Towards Efficient Construction Monitoring: An Empirical Study on Action Recognition Models

Sudheer Kumar Nanduri[1] and Venkata Santosh Kumar Delhi[1]

[1] Indian Institute of Technology Bombay, Mumbai, India
n.sudheer@iitb.ac.in, venkatad@iitb.ac.in

Abstract –

Monitoring fatigue is challenging under computer-vision-based action recognition due to the changes in motion patterns caused by fatigue. Particularly in the construction scenario, the motion patterns are unique per trade and longer than daily life actions, causing challenging scenarios. This paper aims to understand the patterns that can guide the selection of optimal clip durations for aggregating motion features specific to each task. We compare the performance of three action recognition models (I3D, MViT, and VideoMAE) on different construction tasks (excavation, masonry, plastering, etc.) at varying clip lengths. We evaluate the models based on frame-wise accuracy, sequence predictability error, and normalized evaluation duration. Our results show that the transformer-based models outperform the convolutional neural network-based models. The model trained directly over videos performs better than those trained on images. Also, the clip duration affects the model performance differently depending on the task type. Neither the 3s context window from the Atomic Visual Actions (AVA) dataset nor the 10s context window from the Kinetics-400 dataset is suitable for construction tasks. Instead, we suggest a variable clip duration between 5s and 7s, which is preferable depending on the tasks and model architecture. Our work provides insights for developing a dynamic and context-aware duration selection system for action recognition in construction.

Keywords –

Action Recognition; Construction Activities; Clip Durations

1 Introduction

Worker fatigue is a much-studied problem in construction, considering the adverse effects on productivity, safety, and health. Prior attempts to automate fatigue detection utilized computer vision (CV) or on-body sensors for collecting necessary data. Sensors are limited by the contextual information they can collect. For example, an IMU sensor can collect the motion of a specific body part to which it is connected. Computer vision is a better fit for field application because it can collect information from the worker and surroundings simultaneously. In CV, current action recognition approaches analyze patterns in features aggregated from a set of frames. For measuring the duration of a specific action, existing works run action recognition on fixed-length input clips in sequence and append the results.

In prior works for developing fatigue monitoring, work-rest status [1] is set manually for biomechanical evaluation of joint movements. In a pragmatic approach, manual identification and biomechanical evaluation reduce their applicability to real-time monitoring. Utilizing the changes in the movement patterns is a better approach for automating part of these tasks.

While conventional understanding associates muscular fatigue with a decline in performance, the literature suggests that performance is maintained with changes in movement patterns under fatigue. Depending on the variable selected, movement variability may increase or decrease under fatigue [2] [3]. Muscle groups behave differently under fatigue [4]. Fatigue diminishes the force-producing capacity and the ability for smooth and controlled action. This aspect can be utilized for fatigue monitoring and skipping biomechanical analysis. Humans cannot detect movement variability due to cognitive limitations, so computer vision is the best fit. However, breaking down the action into small clips will not be sufficient for fatigue monitoring as the models will lose the context and motion patterns they can use.

One solution is to adopt a dynamic context-aware approach in selecting the clip durations for aggregating motion features specific to each task. The dynamic selection will improve the detection performance while reducing the resource usage for recognition models. Context awareness will also be helpful for safety monitoring and improve the interpretability of action recognition models. In developing a dynamic and context-aware selection system, this work focuses on the first step of understanding the patterns that can guide the selection of clip durations.

2 Literature Review

Though empirical data collection is preferred, two significant works have already provided such data in construction. The first video-based dataset on construction activities [5] is available with 11 action classes. Bricklaying and plastering are manually identified as 9.8 s of mean clip length, with a variance of 3.6s for bricklaying and 5.2 seconds for plastering. An average clip length of 6.8s is observed in the case of all 11 activities, with a variance of 2.7 s. In a later dataset [6], the average activity lengths for bricklaying with sub-activities ranged between 27.33 and 33.30 frames. Considering a 25 FPS video or 30 FPS video, which are the standard practices, these range to only one second of video. The plastering application has 50 frames, which comes close to 1.5 to 2 seconds of video. This difference in clip duration is a significant concern. The subsequent models developed using the datasets are expected to learn correctly from the motion patterns available within the dataset. This assumption makes it difficult to transfer the model trained on the first dataset to the second dataset. Thus, it is preferable to study the performance of different models on standard datasets to identify suitable selection patterns. Developing a new dataset needs to consider this aspect to reduce potential bias.

Recent approaches to action recognition utilize three seconds as a fixed clip length following the Atomic Visual Actions (AVA) dataset [7] standards. The AVA dataset focuses on the action recognition of a single person in a frame with a context window of one and a half seconds before and after the frame. The smaller context window enables fine-scale annotation and improves the action boundary precision. The dataset contains 430 15-minute video clips, 1.58 million class labels, and 80 classes. The dataset is built from movies without actual construction-related actions.

Another important dataset is the Kinetics-400 [8], which has 400 action classes. It has 306,245 clips sourced from YouTube videos, mostly from amateur videographers. Thus, it also provides variety in how the action is performed, along with the clothing, pose, and other parameters. Each clip lasts around ten seconds, providing a context window of 5 seconds before and after the keyframe. Some of the action classes in the dataset, like ‘laying bricks’, ‘plastering’, ‘welding’, and ‘bending metal’, can be utilized for evaluating construction action videos. For this reason, in the current work, we utilized the models with pretraining using the Kinetics-400 dataset.

CV-based action recognition models utilize four major algorithmic approaches - Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformers, and Graph Neural Networks (GNN). Transformer-based models provide the most accurate results and can be considered the latest upgrade over RNNs. GNNs need human joint key points to create graph nodes for analysis. However, they have yet to be proven to perform better than the Transformers, which utilize frame features similar to CNNs.

Two-stream inflated 3D ConvNet (I3D) is a CNN-based model proposed along with the Kinetics-400 dataset [9]. It performs well and is considered a baseline for models and datasets developed afterward. Multiscale Vision Transformers (MViT v2) is a Transformer-based image classification model [10] extended for video classification. Video Masked Autoencoders with dual masking (VideoMAE v2) are also transformer-based models but are built with a specific focus on video data for all types of tasks on videos [11]. Both transformer models utilize 3D convolutions to convert the video frames into patches for training and testing. After conversion, both models use attention networks to identify patterns from the data.

3 Methodology

3.1 Data Collection

The study focuses on evaluating CV models on standard construction processes. Processes selected include excavation, scaffolding, formwork, reinforcement, concreting, masonry, and plastering. The videos are carefully chosen from YouTube, with actions relevant to the processes. The selection of the videos considered video clarity, showing critical activities in the operations without cropping and focusing on the workers doing the actions. Activities within the processes are identified concerning construction literature while ensuring a representative dataset capturing realistic scenarios. A total of 14 videos, ranging from 00:15 to 26:50 minutes (average 08:11 min), are utilized for the current study. Most videos have a 30 FPS frame rate and 1280x720 frame dimensions. Other valuable details are presented in Table 1 below, and the example frames for each task category are presented in Appendix 1 for reference.

The videos are then annotated for the ground truth labels. However, the ground truth labels are a subset of action classes from the Kinetics-400 dataset. This subset comprises actions that can be observed in construction sites. For example, Kinetics-400 does not have any formwork-related class label. In the video of formwork, the annotated labels include classes like ‘moving furniture’, which has the keyword ‘moving’ relevant to the action context.

However, several frames only fall under some of the subsets of the action labels. These frames are marked with a new class named ‘Background’. Frames that show transitions, empty land, and other problems are marked in this class.

Table 1 Frame Count & Task Category of Dataset Used

File Name	Task Category	FPS	Total Frames
CONC_2	Concreting	30	30425
EXCV_2	Excavation	29.94	450
EXCV_6	Excavation	30	3709
EXCV_7	Excavation	30	16382
FMWK_2	Formwork	29.97	17069
FMWK_4	Formwork	30	12651
MASO_2	Masonry	25	4065
MASO_3	Masonry	30	48304
MASO_4	Masonry	29.97	24581
PLAS_1	Plastering	30	18151
PLAS_2	Plastering	30	4350
RFMT_2	Reinforcement	29.97	8364
RFMT_3	Reinforcement	29.97	2131
SCFL_1	Scaffolding	30	15010

3.2 Model Selection

The three action recognition models mentioned in the literature section (i3D, MViT v2, VideoMAE v2) are utilized for comparative analysis. All three models are utilized from the same open-source toolbox, MMAction2, based on PyTorch, to standardize the comparison. Transformer-based models typically have more parameter count and can perform better in general. However, a comparison with the CNN model will help us evaluate the strengths and weaknesses when applying the models to construction scenarios. The I3D model is trained directly on the Kinetics-400 dataset. The MViT model is primarily an image detection model, and hence, it is pre-trained on the ImageNet dataset before training on the Kinetics-400 dataset. The VideoMAE v2 model architecture makes it difficult to train with small datasets. Hence, it is trained with larger hybrid datasets, and the classification head is trained for the Kinetics-400 dataset. Other model details are presented in Table 2 below. Thus, comparing the performance of models also helps us compare the Kinetics-400 dataset with the construction scenario.

Table 2 Model Details

Model	Sampling protocol	FLOPs	Params
I3D	10 clips x 3 crop	43.5G	28.0M
MViT	5 clips x 1 crop	225G	51.2M
MAE	5 clips x 3 crops	180G	87M

A high-performance computer with 2xIntel-Xeon G-6348 CPU and 4x64 GB RAM is used. The available GPUs are not utilized as the work focuses only on model evaluation, and no training is involved.

3.3 Evaluation Parameters

The two parameters mentioned in Table 2 – floating point operations (FLOPs) and parameters (Params)- indicate model performance. Lower FLOPs and Higher Params are the best combinations for CV models. However, a few other considerations also come into play while evaluating the model throughput. The two most important considerations are the model architecture and input variations.

The current work evaluates three aspects - the models' performance on different construction tasks at varying clip lengths. Models and Tasks are detailed in the earlier subsections. Clip length is the final variable discussed in the present subsection.

When a video is chunked into multiple clips, two other parameters that can be useful are the gap duration between two subsequent clips and the overlap duration of the first clip over the second clip. In general, a gap between clips increases the speed at the cost of accuracy, and overlap increases the accuracy by providing additional context at the cost of reduced speed. However, the actual performance might differ due to the model and input variations.

For the current work, clip lengths of 1s, 3s, 5s, 7s, 9s are utilized to cover various temporal scales. Overlaps and gaps are not mixed; when the overlap is present, a gap is not considered, and vice versa. Overlaps chosen are 0s, 2s, and 4s, provided they are always less than the clip duration. For a 1s clip, overlap cannot be 2s as it is the same as a 3s clip length and takes in more features than expected. The gaps chosen are 0s, 1s, 10s. In general practice, gaps are provided such that the frame rate is only 1 Hz, that is, a gap of 1s. However, larger gaps can be considered for the action of longer durations, typically observable in construction sites. A 10s gap is chosen to verify whether a large gap will be helpful. With the given conditions, 22 combinations are formed for durations.

In prediction, the gap durations will have no outputs. This approach improves model performance by reusing the last frame results for all frames within the gap duration. Overlap durations only provide the context for current frame prediction, so there is no effect on the outputs for each frame.

3.4 Evaluation Criteria

Evaluation is based on the accuracy of models without any fine-tuning or transfer learning to avoid any biases from additional training. Doing so will also help maintain classification consistency, even when the specific class labels are absent in the pre-trained dataset. Hence, the dataset prepared is utilized for model output

evaluation.

Three evaluations are made on the model's overall predictions and task-wise predictions - Frame-wise evaluation, Sequence-wise evaluation, and Time duration evaluation. Frame-wise evaluation captures the model's qualitative performance per frame. Sequence-wise evaluation captures the model's sensitivity to the motion pattern changes surrounding the frames. Finally, the time duration evaluation captures the model's quantitative performance per frame. By comparing the results of these three criteria, analysis is carried out to compare construction tasks and extract useful patterns.

Frame-wise evaluation matches the current frame prediction with manual annotations. Accuracy is the ratio of total positive to total positive and negative classifications for a given action. The models are bound to provide noisy predictions for frames annotated as 'Background'. So, the frame prediction is skipped in evaluation for accuracy, but the duration is considered for measuring model prediction performance. Although a multi-class confusion matrix can be utilized, we use the simple metric given the choice of models and a small dataset.

The methodology followed for the sequence-wise predictions is as follows. The previous frame (A) and current frame (B) predictions together form the consecutive clip for evaluation. The confusion matrix is built based on whether the sequence is correctly predicted, as shown in the table below.

Table 3 Confusion Matrix for Sequence-Wise Evaluation

Predicted Sequence	True Sequence	
	A-B	A-A
A-B	TP	FP
A-A	FN	TN

The prediction should correctly capture the change in actions for good sensitivity. Hence, the change in action class is marked as positive, and no change is marked as negative. If consecutive clips have different actions, but the same classifications are provided, a 'False Negative' is considered, and when a different classification is provided, a 'True Positive' is considered. Suppose consecutive clips have the same actions, but a different classification is provided for the second frame. In that case, a 'False Positive' is considered, and a 'True Negative' is considered if the same classification is provided.

In general, accuracy and precision metrics are evaluated from the confusion matrix. An issue with these usual metrics occurs when comparing the sequential predictions. The change in the action may be detected at a very different frame than annotations. This can occur due to changes in the frames, which are invisible to the human eye or missed easily during annotations. Hence, a different metric is developed for the specific case using the same elements of the confusion matrix.

$$PC = (TP + FP)/(TN + FN) \quad (1)$$

$$AC = (TP + FN)/(TN + FP) \quad (2)$$

$$SPE = (PC \,/\, AC) - 1 \quad (3)$$

Where,

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative
PC = Predicted Changes
AC = Actual Changes
SPE = Sequence Predictability Error

In cases where the actual changes can be zero, that is, no changes in the actions in the video, the denominator will be considered as 1 to overcome the division by zero error. Also, the predicted changes can be far more than actual changes. The sequence predictability error must be close to zero.

Finally, the time duration evaluation compares the durations for inference of all the models over the clip duration combinations specified earlier. The time taken per inference is captured and averaged for each model, video, and task. As we compare video clip size variations, considering the number of evaluations made within each variation will provide a better metric for model performance evaluation. The average duration per video is divided by the total number of evaluations made within the video to evaluate the performance.

$$T_{NE} = 100 * T_{AE} \,/\, N_E \quad (4)$$

Where,

T_{NE} = Normalized Evaluation Duration, in seconds
T_{AE} = Average Evaluation Duration, in seconds
N_E = Total number of evaluations

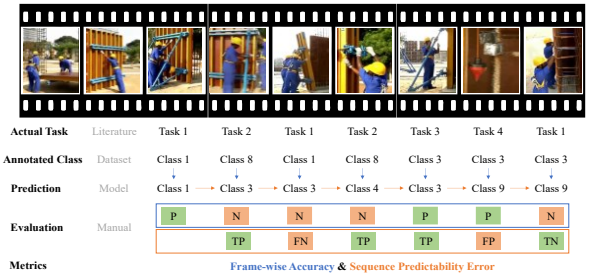


Figure 1 Annotation and Evaluation Methodology

The annotation and evaluation methodology is depicted in 'Figure 1' above, except the duration evaluation.

4 Results

4.1 Frame-wise Evaluation

The overall model performances are unexpectedly low. The best-performing model is MAE with 56.42%, followed by MViT model with 49.27%, and I3D model with 43.13%. The clip overlap and gap durations show slight improvements but do not form any meaningful patterns in frame-wise accuracy. Models offer good performance for tasks with action classes available in the Kinetics-400 dataset. Across models, the average accuracy stays in a similar ratio, as shown in 'Figure 2' below.

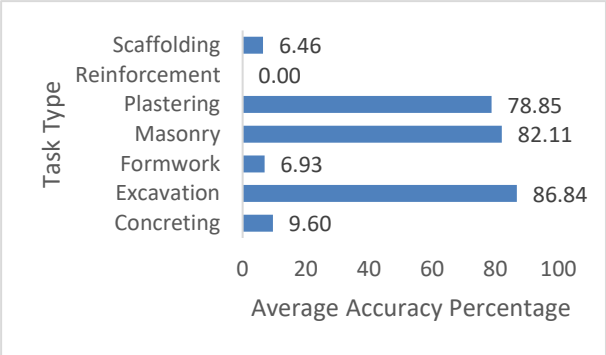
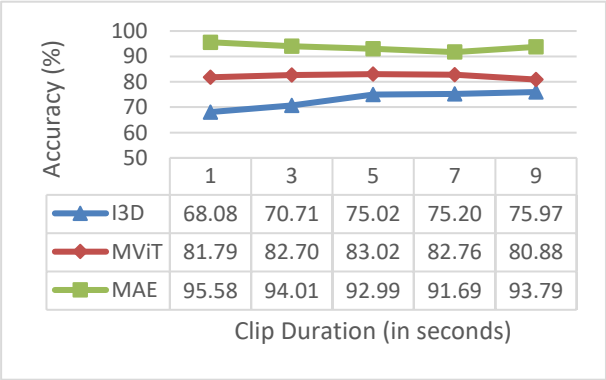


Figure 2 Average Accuracy Percentage Per Task Type

The need for more accuracy indicates disagreement between manual annotations and model predictions for action classes unavailable in the pre-trained dataset. Utilizing the results from the top three tasks – Excavation, Masonry, and Plastering, the frame-wise accuracies are reported in 'Figure 3' below.



	1	3	5	7	9
I3D	68.08	70.71	75.02	75.20	75.97
MViT	81.79	82.70	83.02	82.76	80.88
MAE	95.58	94.01	92.99	91.69	93.79

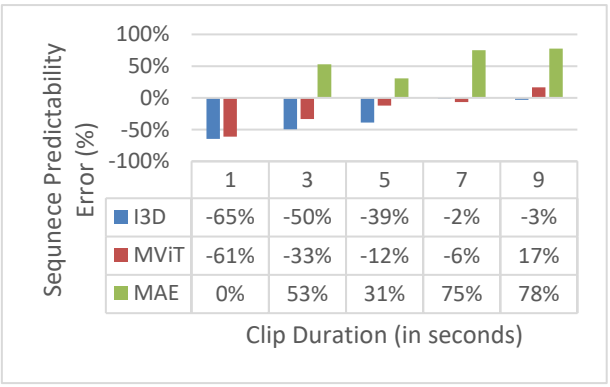
Figure 3 Model-wise Average Accuracy Percentage Per Clip Duration

The I3D model shows a considerable increase in performance with an increase in the input clip duration. After 5-second clip length, a plateauing of I3D performance and a decrement of MViT performance can be observed. The decrement in the MAE model might be related to the fact that it is trained on datasets beyond the K-400 dataset, owing to its architectural needs.

4.2 Sequence-wise Evaluation

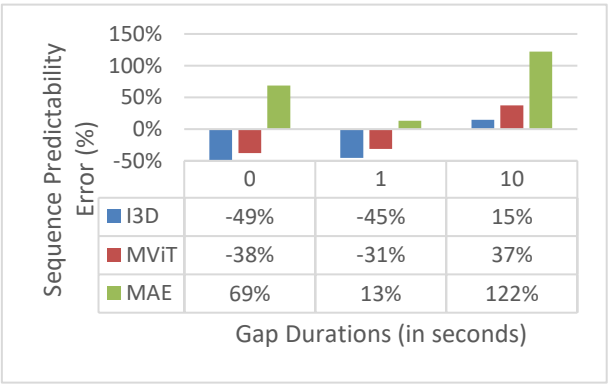
Models might capture more action transitions than annotated ones because they can see more details than humans. However, the duration, overlap, and gap combinations also affect the predictive capabilities due to the sampling strategies for testing. A 10-clip x 3-crop strategy takes ten clips from the given video, crops three different zones within each clip, and utilizes the information for prediction. Thus, a longer video duration, an overlap, and a gap between clips will all provide different features.

The Sequence Predictability Error (SPE) of models for different durations is presented from 'Figure 4' to 'Figure 6' below. The negative predictability shows that the models predict less than the actual, and positive values indicate that the models predict more. Being closer to zero is preferred, as the models are expected to perform best in correctly identifying sequences.



	1	3	5	7	9
I3D	-65%	-50%	-39%	-2%	-3%
MViT	-61%	-33%	-12%	-6%	17%
MAE	0%	53%	31%	75%	78%

Figure 4 Duration-wise Sequence Predictability Error of Models



	0	1	10
I3D	-49%	-45%	15%
MViT	-38%	-31%	37%
MAE	69%	13%	122%

Figure 5 Gap Duration-wise Sequence Predictability Error of Models

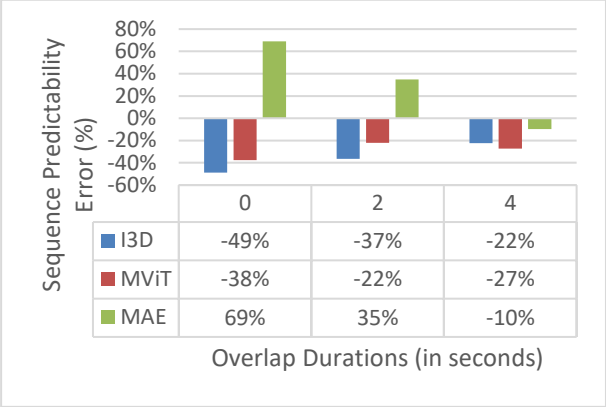


Figure 6 Overlap Duration-wise Sequence Predictability Error of Models

As the duration increases, the predictability of the models – I3D and MViT- improves. However, the MAE model over-predicts the number of changes in the video. A slight gap in the durations improves the predictability, but too much will throw the models far away. Also, a slight overlap improves the predictability.

4.3 Time Duration Evaluation

The models are compared with different clip lengths and durations and presented in 'Figure 7' below.

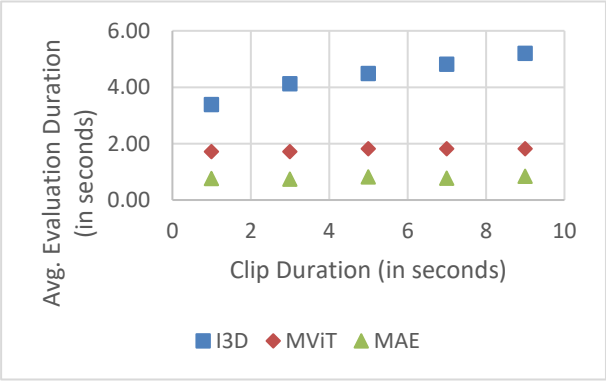


Figure 7 Average Evaluation Duration Per Model at Different Clip Lengths

The I3D, a CNN-based model, shows a linear increase in the evaluation duration with increasing clip lengths. Combined with the need for more accuracy beyond 5 seconds, using large clip lengths for CNN-based models is not valuable. The Transformer-based models do not show any linear increase and are stable across the clip lengths.

Although evaluation duration increases with time, that is the case when there is no overlap. Additional overlaps of 2 and 4 seconds did not show the same incremental behavior in performance time. One exception is the 9-second limit for the video clips. The I3D model took longer in any duration-overlap combination while evaluating the clips of length 9 seconds. But beyond the 9 seconds, the evaluation duration reduces. There is no clear explanation for this behavior.

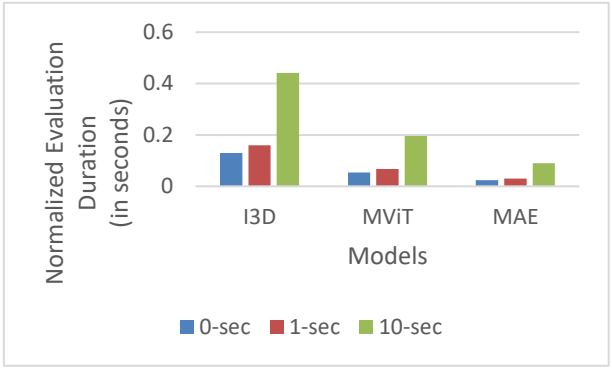


Figure 8 Gap Duration-wise Normalized Evaluation Durations of Models

From 'Figure 8' above, having a large gap between clips increased the normalized evaluation duration across all models. Technically, the models consider each clip a separate video and only predict the action within the clip context. But, the behavior here suggests that the models utilize the previous videos as context.

4.4 Task-wise Evaluation

Breaking down the model performance task-wise, 'Figure 9' below shows the different performance of models for the tasks.

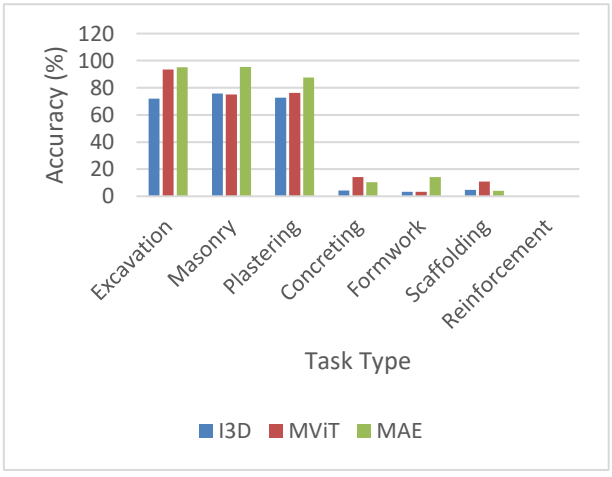


Figure 9 Average Accuracy Percentage per Task Type for Models

The need for correct labels for other tasks limits our

ability to evaluate model performance systematically. However, MViT performs well for concreting and scaffolding, while MAE performs well for formwork.

Task-wise evaluations are focused only on the best-performing tasks. The models show different behavior under different clip durations as presented in Table 4.

Table 4 Average Accuracy of Models for Best Performing Task Types under Different Clip Lengths

Task Type	Model	Clip Duration (in seconds)				
		1	3	5	7	9
Excavation	I3D	66	68	74	72	76
	MViT	92	92	94	95	93
	MAE	100	95	94	94	95
Masonry	I3D	77	74	76	75	77
	MViT	77	77	76	75	71
	MAE	95	96	96	95	96
Plastering	I3D	58	71	75	80	74
	MViT	73	76	77	76	78
	MAE	90	89	87	84	89

Table 5 Average Accuracy of Models for Best Performing Task Types for Overlap-Gap Combinations

Task Type	Model	Overlap - Gap Combinations				
		0-0	0-1	0-10	2-0	4-0
Excavation	I3D	72	70	78	73	64
	MViT	93	93	92	95	94
	MAE	95	96	94	95	97
Masonry	I3D	76	76	75	77	77
	MViT	74	74	79	74	73
	MAE	95	96	95	96	96
Plastering	I3D	67	75	69	75	81
	MViT	76	76	76	77	78
	MAE	87	88	89	86	88

The rounded-off average accuracies for overlap-gap duration combinations is presented in Table 5 above. For the I3D model, increasing the overlap duration increases the model performance for plastering tasks but decreases accuracy for the excavation task. A 2-sec overlap in the I3D model improves the performance when the clip durations are below 9 seconds.

Comparing the task-wise sequence predictability of the models, it was observed that the plastering task has too many change predictions than actual. The results are presented in 'Figure 10' below. Across the tasks, none of the models can predict the number of changeovers sufficiently. Although performance for formwork tasks seems promising, the high background percentage might also lead to this erroneous evaluation.

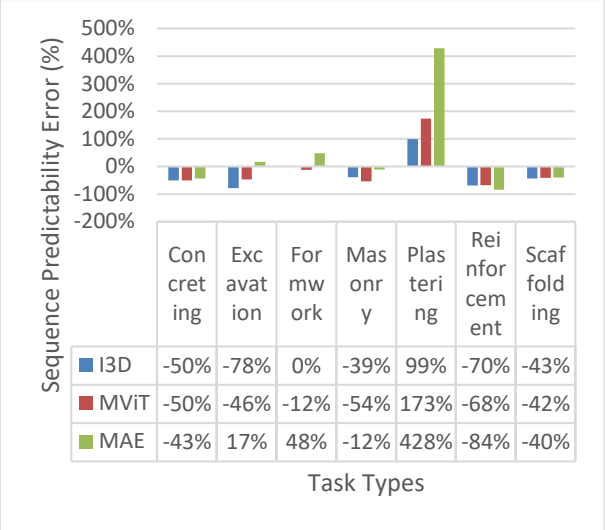


Figure 10 Task-Type wise Sequence Predictability Error of Models

Separating the task-wise performance, each model shows a different pattern for the duration. Results from the comparison are presented in 'Figure 11' below. The excavation task takes most of the time for evaluation, followed by reinforcement and plastering.

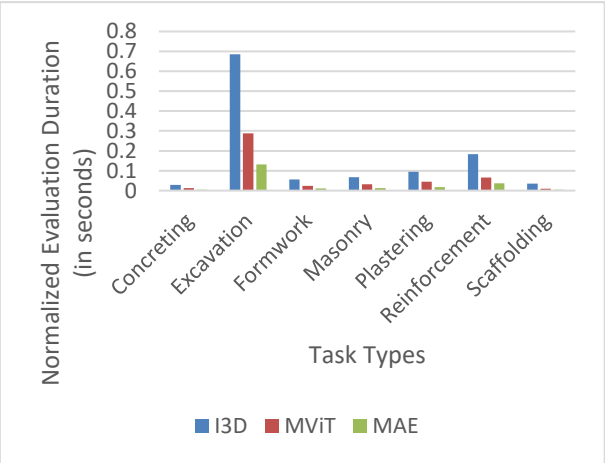


Figure 11 Normalized Evaluation Duration per Task Type for Models

5 Discussion

Summarizing the results across models, duration combinations, and task types, a few patterns are found useful. Frame-wise accuracy suggests that the Transformer models are the best-performing models. The

MAE model, which is trained on videos directly, performs better than MViT, which is trained on images and extended to videos. For I3D, a CNN-based model, the clip duration affects the performance by providing a larger context window. However, a 5-second limit occurs across models trained on images. The context window of 5 seconds seems sufficient for our use cases.

For the I3D model, the excavation task with a context window of 5 seconds and the plastering task with a context window of 7 seconds perform best. However, the exact durations are the worst performing for the MAE model. The MAE model is best performing for clips of short durations.

For excavation and masonry, providing a gap between clips increased the accuracy, but extra overlap increased the accuracy for plastering. This suggests that the motion patterns within the former two actions are highly repeated, whereas, for the latter, a considerable difference occurs.

In the sequence predictability, higher clip durations increase the predictability of the I3D and MViT models but reduce the MAE model. A slight gap or overlap in the durations improves the predictability.

For most tasks, all the models predict less than actual sequence changes. Only for the plastering task, the models predict more than the actual. This might be a one-off case and needs further investigation. The excavation and reinforcement tasks have higher evaluation durations than the average model performances.

In the concreting activity videos, the actions are the concreting of a floor slab and a road. The most repetitive actions are dumping the concrete and evening (smoothing) the surface. The annotations given are ‘unloading the truck’ and ‘sweeping the floor’, considering the closeness of these labels to the actions. However, most model detections classified the surface evening as ‘digging’. The motion patterns relevant to these two classes need to be differentiated. Similarly, most formwork tasks are annotated under ‘moving furniture’ since they involve moving and fixing the components. The best-performing model, MAE, detects some of these actions as ‘building shed’ and ‘bending metal’.

There is no pre-training involved in the study, and the annotations are mapped to nearest action class of the Kinetics-400 dataset. Consequently, the accuracy results in task-wise results are not useful for concreting, formwork, scaffolding and reinforcement works. Since the focus of the current work is on temporal precision rather than frame level accuracies, some useful interpretations can be derived from the task-wise results of these actions also. The sequential predictability error identifies how well the model can detect the action switching from one to the next. Even in the mapped action classes, the actual action is irrelevant, and only the change of action is important. From the task-level results, a negative SPE is seen in most cases suggesting that the models predict less switches than that can be detected by the manual annotators. Also, the normalized evaluation duration results hint at a possible correlation with the video clip lengths. The smallest video clip of excavation has only 450 frames and the results of excavation point to a very high duration for evaluation. However, this is contrary to the general expectation that a smaller clip can be evaluated faster. There is no relation found between the accuracy and duration because the accuracies are very less for the reinforcement work yet the action class took higher duration for evaluation. For the action classes existing in the dataset and are directly related to construction activities like masonry, plastering and excavation, results across all evaluations are useful.

These observations lead to the development of a classification system for tasks and models. Excavation tasks can be predicted better with 5-second context windows but need too much time for prediction. Plastering tasks can be predicted best with 7-second context windows, but the number of switches detected can be far higher than actual. Masonry tasks can be predicted with 5-second context windows without any drawback. Overall, additional overlap or gap increases the prediction performance on speed and accuracy. However, models trained on video datasets directly perform the best. If the sequence changes are not a particular concern, then using the video data-trained models with the least clip durations will improve the model performance dramatically. However, if sequence predictability is the primary concern, the transformer-based model trained on the image dataset will work best with a clip duration between 5 and 7 seconds.

Overall, this work suggests that neither the 3-sec context window from the AVA dataset nor the 10-sec context window from the Kinetics-400 dataset is suitable for construction tasks. Instead, it is a variable that needs to be carefully evaluated and considered for better performance.

The study offers useful insights for the construction organizations adopting automated visual surveillance for applications like fatigue and safety monitoring. One of the important components of such systems is the worker action recognition model. Action recognition is a resource-intensive task, requiring the model to classify a fixed set of frames from the input videos into relevant actions. Changing the fixed set of frames to a variable set according to the task is a valuable optimization. Consider an analogy: Suppose you watch a live stream of your favorite action-oriented sport. Usually, there is a minimum of 30-second delay in live streaming the sport, depending on the technology. Now, imagine a computer watching the live match at the location, informing you whether there is any useful action in progress. If it can

analyze the match faster and more accurately, your cost and time will also be saved. In the game, two people can be doing different actions simultaneously, one running towards the opponents and another standing in some corner for quite some time. We too, focus on the running person rather than the standing person. By enabling the computer to focus on the running person, we can save the cost and time of operating the computer while increasing accuracy. Our proposed approach is one such method to achieve this. By remembering how humans move for different tasks, the computers can speed up their performance, by using smaller clip lengths for important actions and vice versa. Using technical terminology, the proposed method improves the speed of action recognition by aligning the clip durations with task-specific motion patterns. This strategic optimization reduces the data analysis costs while increasing the speed of evaluations for the organizations. Also, some action recognition models depend on upstream models like object detection. When these upstream models cause erroneous detections or missed predictions, our methodology can improve the action recognition as the systems can still make correct predictions due to the varying clip lengths according to the task.

6 Limitations and Future Work

Compared to other related works on action recognition for construction activities, this study stands out in its comprehensive evaluation of different models and clip durations. While most previous studies have focused on a single model or a fixed clip duration, this study provides a more nuanced understanding of how the model performance varies with the clip duration and task type.

However, like all studies, this one also has its limitations. The study is based on a limited number of construction tasks and a specific dataset on which the selected models are pre-trained. The generalizability of the findings to other tasks or datasets remains to be tested. The reliance on YouTube videos for data collection may only partially capture the complexity and diversity of construction activities in real-world scenarios. Future studies could benefit from using a dataset collected directly from construction sites. The study does not discuss potential challenges such as sensitivity to noise, occlusion, camera motion, or lighting conditions, which are common in construction sites. It is expected that the videos collected from YouTube will include these issues in the dataset.

The need for correct labels for all the tasks limits the applicability of the pre-trained models directly. Yet, the work proceeds to utilize the model, and as a result, the respective estimates for tasks like reinforcement could have yielded more useful clip duration estimates. However, these observations proved helpful in studying the model performance for these tasks. The higher evaluation time for reinforcement tasks and higher accuracy of the MViT model for concreting and scaffolding compared to the MAE model are useful predictions for further work.

Lastly, the sequence predictability error metric measures the total number of predicted sequence changes. It does not capture the sensitivity of correctly identifying the change precisely at the annotated frame. This is an approximation, considering that the models might see the frames in videos differently than humans and capture more details for their analysis. Future studies can also modify the metric and measure the duration between annotated and predicted change for more sensitivity.

There are several promising directions for extending this work. One possibility is to incorporate more construction tasks into the study to validate further and refine the findings. Another possibility is to create a dataset of the same set of actions with varying execution time and use the dataset for evaluation following the methodology discussed in the current work. If the performance of models depends on execution time also, then the context-awareness of models can add another dimension of variance. For example, a dataset of masonry work is created with action lengths between 3 and 9 seconds. The context-aware systems can identify the mean and variance from this dataset. In a future application, the system can consider this variance while varying the clip lengths. Another direction is to develop a new dataset with more accurate and diverse annotations, which could help to improve the model's performance and robustness. Fine-tuning or transferring the models to the construction domain could also be explored to exploit the domain-specific knowledge and data. Finally, other datasets (like AVA), modalities (using skeletal frame, optical flow), architectures for action recognition, and tasks (recognition, segmentation, localization) could be investigated. Evaluation for different mechanization levels (manual, tools, equipment, machinery) can be done for more detailed analysis. Additionally, the application of these models to other tasks, such as safety analysis or productivity assessment, could be explored.

7 Conclusion

This paper compares three action recognition models for construction activities: I3D, MViT, and VideoMAE. The models are evaluated on a YouTube video dataset covering seven standard construction processes. The paper analyzes the effect of clip length, overlap, and gap on the model performance, using frame-wise accuracy, sequence predictability error, and normalized evaluation duration as the criteria. The results show that the transformer-based models outperform the CNN-based

model in accuracy but have different sensitivity to the clip duration and motion patterns. The paper also suggests that the optimal clip length for construction action recognition is between 5 and 7 seconds, depending on the task and the model. The paper contributes to understanding the strengths and weaknesses of different action recognition models for construction scenarios. It provides insights for developing a dynamic and context-aware selection system for clip durations.

References

[1] Y. Yu, H. Li, X. Yang, L. Kong, X. Luo, and A. Y. L. Wong, "An automatic and non-invasive physical fatigue assessment method for construction workers," *Autom. Constr.*, vol. 103, pp. 1–12, Jul. 2019, doi: 10.1016/j.autcon.2019.02.020.

[2] N. Cortes, J. Onate, and S. Morrison, "Differential effects of fatigue on movement variability," *Gait Posture*, vol. 39, no. 3, pp. 888–893, Mar. 2014, doi: 10.1016/J.GAITPOST.2013.11.020.

[3] K. L. Mudie, A. Gupta, S. Green, and P. J. Clothier, "Adaptation of lower limb movement patterns when maintaining performance in muscle fatigue," *Hum. Mov. Sci.*, vol. 48, pp. 28–36, Aug. 2016, doi: 10.1016/J.HUMOV.2016.04.003.

[4] J. C. Cowley and D. H. Gates, "Proximal and distal muscle fatigue differentially affect movement coordination," *PLoS One*, vol. 12, no. 2, p. e0172835, Feb. 2017, doi: 10.1371/JOURNAL.PONE.0172835.

[5] J. Yang, Z. Shi, and Z. Wu, "Vision-based action recognition of construction workers using dense trajectories," *Adv. Eng. Informatics*, vol. 30, no. 3, pp. 327–336, Aug. 2016, doi: 10.1016/j.aei.2016.04.009.

[6] D. Roberts, W. Torres Calderon, S. Tang, and M. Golparvar-Fard, "Vision-Based Construction Worker Activity Analysis Informed by Body Posture," *J. Comput. Civ. Eng.*, vol. 34, no. 4, p. 04020017, Jul. 2020, doi: 10.1061/(ASCE)CP.1943-5487.0000898.

[7] C. Gu *et al.*, "AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions," May 2017, Accessed: Dec. 15, 2023. [Online]. Available: http://arxiv.org/abs/1705.08421.

[8] W. Kay *et al.*, "The Kinetics Human Action Video Dataset," May 2017, Accessed: Dec. 15, 2023. [Online]. Available: http://arxiv.org/abs/1705.06950.

[9] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 4724–4733, doi: 10.1109/CVPR.2017.502.

[10] Y. Li *et al.*, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, vol. 2022-June, pp. 4794–4804, doi: 10.1109/CVPR52688.2022.00476.

[11] L. Wang *et al.*, "VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking," Mar. 2023, Accessed: Dec. 16, 2023. [Online]. Available: http://arxiv.org/abs/2303.16727.

Appendix 1: Sample Task Frames