Computer Fraud and Security
ISSN (online): 1873-7056

Comparative Evaluation of Deep Learning Architectures for Human Action Recognition

Ravindersingh Rajpal[1]

Meta Platforms, Inc.

rkrajpal@meta.com

Abstract—Human action recognition, or video-based activity classification, remains a challenging task due to complex spatiotemporal dynamics, variations in viewpoint, and motion blur. This study presents a comparative evaluation of three state-of-the-art deep learning architectures for video action recognition: Temporal Segment Networks (TSN), 3D Convolutional Networks (C3D), and Two-Stream Inflated 3D ConvNets (I3D). The models were implemented using the UCF101 dataset, comprising 13,320 videos from 101 action categories. Extensive preprocessing was performed to extract and resize video frames and optical flows. Experimental results demonstrate that C3D achieved the highest test accuracy of 83.2%, followed by I3D with 74.5% and TSN with 38.1%. Despite the computational cost, 3D convolutional architectures provided superior spatiotemporal feature learning. The findings highlight the trade-off between model complexity and training efficiency, and suggest that with improved computational resources, both I3D and TSN could achieve performance closer to current state-of-the-art benchmarks.

Index Terms—Action recognition, video classification, deep learning, convolutional neural networks, temporal modeling, spatiotemporal features.

I. INTRODUCTION

Human action recognition, or video-based action classification, refers to the task of identifying a person's activity or motion from a sequence of video frames. It plays a critical role in numerous applications, including human–computer interaction, video surveillance, sports analytics, and content retrieval. Despite its wide relevance, action recognition remains a challenging problem due to the complex spatiotemporal dynamics of human motion, variations in appearance, background clutter, occlusion, and differences in camera viewpoint and motion [1],[2].

Accurately modeling both spatial and temporal dependencies in video data is essential for robust performance [6]. The spatial component involves understanding visual appearance features within individual frames, while the temporal component captures the motion information that evolves across consecutive frames. Variations in illumination, scale, and camera perspective further complicate the task, making handcrafted feature approaches such as SIFT, HOG, or HOF insufficient for large-scale video understanding [3]. Consequently, deep learning–based architectures have become the dominant approach for this problem [4], [5].

In recent years, several deep neural network architectures have been proposed to model video data effectively. Among them, Temporal Segment Networks (TSN), 3D Convolutional Networks (C3D), and Two-Stream Inflated 3D ConvNets (I3D) have shown remarkable promise in learning discriminative spatiotemporal representations. Each of these models adopts a different strategy to capture temporal structure: TSN focuses on segment-level sampling for long-range temporal reasoning, C3D applies 3D convolutions directly to short video clips to jointly learn motion and appearance, and I3D extends the twostream paradigm by inflating 2D filters into 3D to better exploit pretrained image models.

In this study, we conduct a comparative evaluation of these three representative architectures for video action classification. All experiments are performed on the UCF101 dataset [7], [8], a well-known benchmark comprising 13,320 videos across 101 action categories, collected from YouTube. The dataset provides a diverse range of human activities such as sports, instrument playing, and daily actions, making it a suitable testbed for evaluating generalization capability.

[1] This work was completed prior to joining Meta and is not associated with his current employment at Meta Platforms, Inc.

The remainder of this paper is organized as follows: Section II describes the methodology and implementation details, including data preprocessing and model configurations. Section III presents the experimental results and a detailed comparison of performance across architectures. Section IV concludes the study and highlights directions for future work.

II. METHODOLOGY

Before reviewing various architectures, a significant amount of time was devoted to understanding and extracting the data. We performed data exploration and preprocessing on the dataset, which are discussed as follows.

A. Data Exploration and Preprocessing

In addition to the video files, the UCF101 dataset provides predefined train–test splits. Specifically, the dataset includes three split directories: *trainlist01/testlist01*, *trainlist02/testlist02*, and *trainlist03/testlist03*. These text files define standardized partitions to facilitate consistent benchmarking across different architectures. In this work, we used *trainlist01/testlist01*.

Since the architectures implemented in this study operate on frames and optical/dense flows, which are well-established representations for capturing motion dynamics in videos [5], [9], we considered two possible approaches. The first was to minimize the overhead of repeatedly reading videos by pre-extracting and storing the corresponding frames and flow data on disk. The second was to extract frames and other information on the fly during training. We experimented with both approaches while training one model. Ultimately, storing the frames and optical flows proved to be the optimal solution, albeit requiring substantial preprocessing time for all approximately 11k videos in the selected split before training or testing. During this process, we also resized the frames to 224×224×3, as two of the models required this input size.

Initially, this step was challenging due to limited computational resources—CPU utilization reached 100% on an eightcore machine, and the extraction time was approximately 3–4 minutes per video. To overcome this, we deployed the preprocessing task on Google Cloud Platform using two instances, completing the entire process in roughly 100 hours.

B. Architecture 1: Temporal Segment Networks (TSN)

The Temporal Segment Networks (TSN) architecture was proposed in 2016 by Wang *et al.* [10]. TSN is designed to model long-range temporal dependencies within a video by dividing it into segments. The original authors proposed using three segments, and we followed the same approach. Specifically, the video is divided into early, middle, and late segments, from which one frame and one optical flow are randomly sampled per segment to capture spatial and temporal information, respectively [11]. This results in six convolutional networks—three for spatial and three for temporal processing. The outputs of the spatial networks are averaged to obtain a spatial consensus, and the same is done for the temporal networks. A final consensus between these two outputs produces the overall action prediction for the video [12].

The original TSN used the Batch Normalization Inception network as its base ConvNet. Since this is a deep network, we applied transfer learning to reduce training time, using the *InceptionResNetV2* model provided by Keras for each segment. However, our final results did not meet expectations due to a known issue in Keras related to the Batch Normalization layer during transfer learning. An issue was raised on the official Keras GitHub repository regarding this bug. Consequently, we were required to freeze the Batch Normalization layers (i.e., set them as non-trainable), effectively disabling their functionality, which led to degraded performance. Additionally, the model's high complexity resulted in long training times, and increasing the learning rate caused underfitting [13].

C. Architecture 2: 3D ConvNet (C3D)

The 3D ConvNet (C3D) architecture was introduced in 2014 by Tran *et al.* [14]. This model learns both spatial and short-term temporal features by applying 3D convolutional filters over sequences of video frames. In our implementation, we scaled the frames from 224×224×3 to 112×112×3 to reduce computational cost. The network was implemented in Keras without the inception modules used in TSN. As the model is highly parameter-intensive, we initialized it with pretrained weights from the Sports-1M dataset and fine-tuned it for our action recognition task.

To capture temporal information, we stacked multiple frames in a time-distributed manner before inputting them into the model. The C3D network first learns spatial appearance patterns from initial frames and then captures motion cues from subsequent ones. Due to hardware constraints, we could not stack more than 16 frames per video during training. Once trained, the model was used to predict the class of each video. However, the model's ability to capture long-range temporal

dependencies remained limited, a known limitation of fixed-length 3D convolutional architectures [15], [16], and its computational cost was substantial.

D. Architecture 3: Two-Stream Inflated 3D ConvNet (I3D)

The Inflated 3D ConvNet (I3D) architecture, proposed by Carreira *et al.* [14] in 2017, extends the C3D approach by combining it with the two-stream framework that processes both RGB frames and optical flows. I3D inflates 2D convolutional filters into 3D filters, allowing the use of pretrained 2D weights from ImageNet while modeling temporal features effectively. This design enables the model to leverage spatial and temporal cues jointly.

In our implementation, we trained two separate networks—one for frames and another for flows—each initialized with pretrained weights from the ImageNet and Kinetics datasets. During inference, predictions from the two networks were averaged based on the argmax of the softmax outputs to produce the final classification. The flow stream required stacking optical flow frames into 3D tensors, which made training computationally demanding. Each training epoch for the flow model took approximately 2552 seconds. Due to time constraints, we trained the flow model until it achieved 75% validation accuracy.

III. RESULTS

TABLE I

MODEL ACCURACY ON THE UCF101 DATASET

DNN Architecture	Accuracy
3D ConvNet (C3D)	83.2%
Two-Stream Inflated 3D ConvNet (I3D)	74.5%
Temporal Segment Networks (TSN)	38.1%

The evaluation of all three architectures was conducted on the UCF101 dataset using a 60:20:20 train–validation–test split. The test set accuracy was found to be consistent with the validation performance, indicating that none of the models exhibited significant overfitting or data leakage. Table I summarizes the comparative accuracy of the implemented deep neural network (DNN) architectures.

A. Performance Comparison

Among the evaluated models, the 3D ConvNet (C3D) achieved the highest accuracy, reaching 83.2% on the test set. This superior performance can be attributed to the model's ability to jointly capture short-range spatial and temporal dependencies through 3D convolutions. Despite its computational intensity, C3D demonstrated relatively stable convergence and strong generalization capability once pretrained weights from the Sports-1M dataset were fine-tuned on UCF101.

The Two-Stream Inflated 3D ConvNet (I3D) obtained a test accuracy of 74.5%, with a validation accuracy of 88.4% for the frame-based stream and 75.2% for the flow-based stream. Although I3D theoretically extends the temporal modeling capability of C3D, the computational cost of training two large parallel networks limited the extent of fine-tuning we could perform. Furthermore, the optical flow stream was significantly more resource-intensive, requiring nearly 2550 seconds per epoch on GPU instances hosted on Google Cloud Platform (GCP). As a result, we curtailed training once the validation accuracy plateaued above 75%. The frame stream, however, exhibited faster convergence and smoother learning behavior, suggesting that RGB frame information alone captures substantial temporal cues for many classes within UCF101.

The Temporal Segment Networks (TSN) model achieved a test accuracy of 38.1% after approximately 25 hours of training, covering around 200 epochs. The relatively low accuracy is primarily due to the absence of active batch normalization during transfer learning, as discussed earlier, and the limited computational budget available for longer training. Nevertheless, the observed learning trend (Figs. 1–3) indicates that TSN continued to improve steadily, implying

that additional training time or GPU resources could lead to a notable performance boost. Previous studies have reported substantially higher accuracy for TSN when trained on largescale datasets and high-performance multi-GPU clusters [17], supporting this inference.

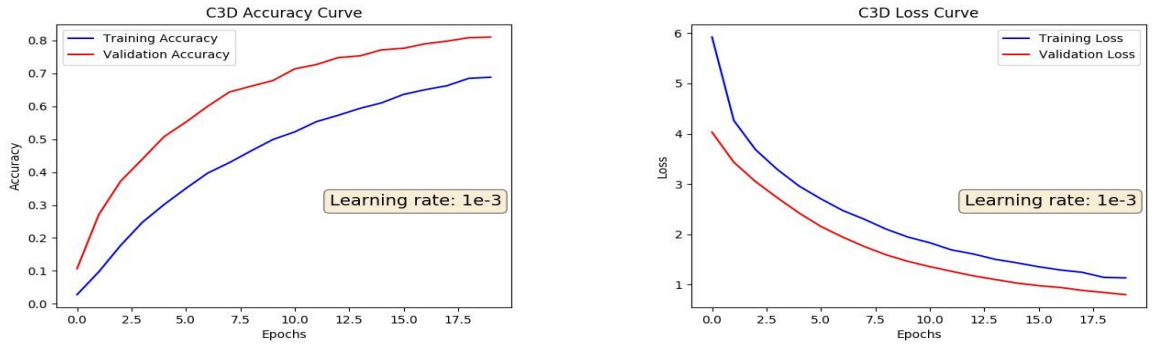
B. Training Efficiency and Observations

All three architectures were trained using GPU-enabled instances on GCP. The preprocessing overhead for frame and optical flow extraction contributed significantly to the total runtime. Training duration per epoch varied greatly between architectures due to differences in model depth and input modalities. C3D offered a balanced trade-off between computational cost and accuracy, whereas I3D, despite being conceptually more powerful, demanded nearly 2–3× longer training cycles due to its dual-stream structure.

Validation curves for the three architectures are shown in Figs. 1–3. The C3D model demonstrated smooth convergence with minimal oscillation, while I3D showed periodic fluctuations likely due to the asynchronous training of its two parallel branches. TSN exhibited slower convergence, which is consistent with the use of frozen normalization layers and relatively shallow gradients during backpropagation.

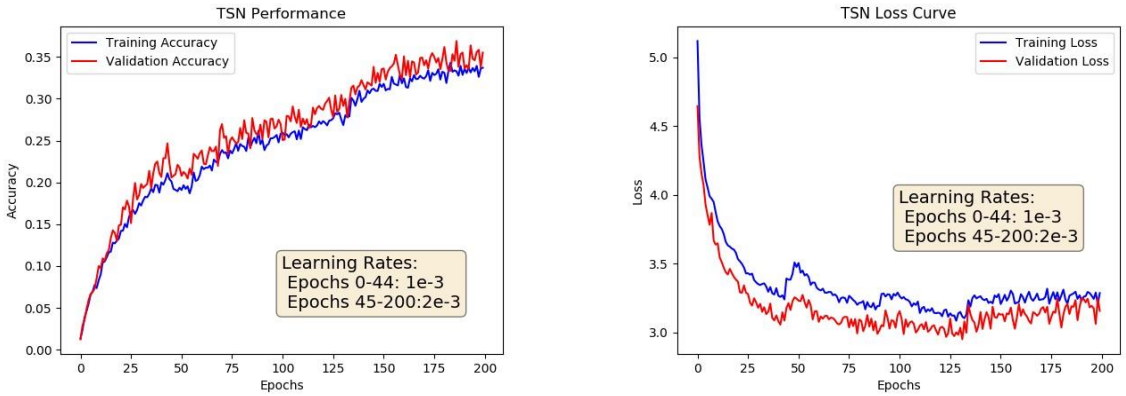
C. Summary of Results

Overall, the experiments confirm that 3D convolution-based architectures outperform 2D segment-based methods when sufficient computational resources and pretrained weights are leveraged. While C3D provided the best empirical accuracy in our setup, the I3D architecture remains more scalable for future extensions involving longer temporal sequences and larger datasets. With improved training configurations and increased GPU availability, both I3D and TSN are expected to achieve performance closer to state-of-the-art benchmarks reported in the literature.



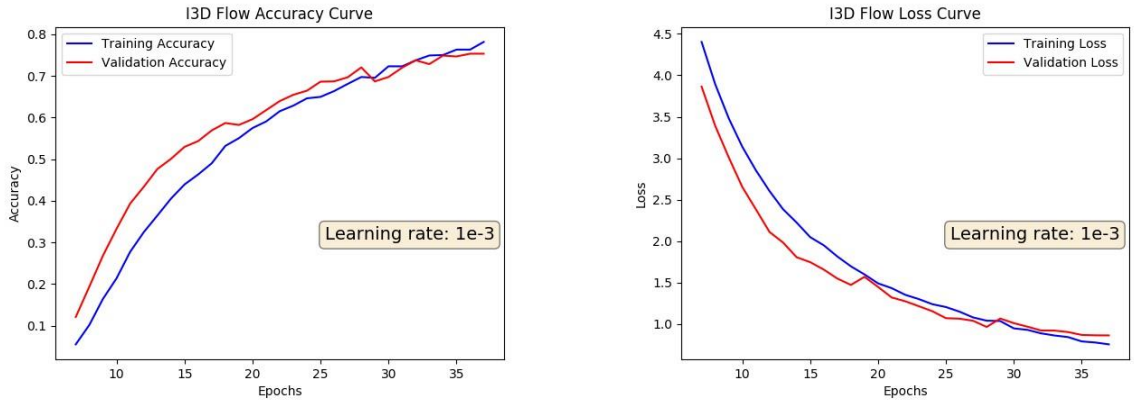
(a) Training and validation accuracy (b) Training and validation loss

Fig. 1. Performance curves for the C3D model. The model demonstrates smooth convergence with stable learning behavior.



(a) Training and validation accuracy (b) Training and validation loss

Fig. 2. Performance curves for the TSN model. Gradual improvement is consistent with limited batch normalization and compute constraints.



(a) Training and validation accuracy (b) Training and validation loss

Fig. 3. Performance curves for the I3D flow network. Periodic fluctuations reflect asynchronous learning in its dual-stream design.

IV. CONCLUSION

This paper presented a comparative analysis of three prominent deep learning architectures—TSN, C3D, and I3D—for video-based human action recognition using the UCF101 dataset. The study demonstrated that 3D convolution-based models outperform 2D segment-based approaches when sufficient computational resources and pretrained weights are available. C3D achieved the best performance, offering a balance between accuracy and computational efficiency, while I3D exhibited strong potential for scalability with larger datasets and extended temporal sequences. TSN, although less accurate in our setup, remains valuable for its lightweight design and suitability for real-time applications.

Future work will focus on extending the dataset to Kinetics600, experimenting with attention-based temporal modules, and optimizing training through distributed GPU frameworks. Additionally, investigating transformer-based video models, such as TimeSformer and ViViT, may further improve long-range temporal reasoning and recognition accuracy.

REFERENCES

1. J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: A review,” *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.
2. R. Poppe, “A survey on vision-based human action recognition,” *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.
3. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
4. A. Karpathy *et al.*, “Large-scale video classification with convolutional neural networks,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732.
5. K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 568–576.
6. J. Donahue *et al.*, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.
7. K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A dataset of 101 human action classes from videos in the wild,” CRCV Technical Report, 2012.
8. H. Kuehne *et al.*, “HMDB: A large video database for human motion recognition,” in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2011, pp. 2556–2563.
9. T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in *Proc. European Conf. on Computer Vision (ECCV)*, 2004, pp. 25–36.

10. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 9, pp. 2220–2232, 2018.
11. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 4489–4497.
12. D. Tran *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.
13. G. Varol *et al.*, "Long-term temporal convolutions for action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510–1517, 2018.
14. J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.
15. P. Goyal *et al.*, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," arXiv:1706.02677, 2017.
16. G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021.
17. A. Arnab *et al.*, "ViViT: A video vision transformer," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2021