Deep Learning Approaches for Human Activity Recognition in Video Sequences: A Comparative Study of CNN-LSTM Architectures

Nicholas Mondella
Department of Informatics, Systems and Communication
University of Milano - Bicocca
n.mondella@campus.unimib.it

September 30, 2025

Abstract

Human Activity Recognition (HAR) in video sequences is a challenging computer vision task with significant applications in surveillance, healthcare, and human-computer interaction. This paper presents a comprehensive comparative study of three deep learning architectures for video-based activity recognition: ConvLSTM, Long-term Recurrent Convolutional Networks (LRCN), and Attention-based 3D CNNs. We evaluate these models on the UCF101 dataset, focusing on 25 distinct human activities. Our experimental results demonstrate that the Attention-based 3D CNN achieves the highest accuracy of 87.1%, followed by ConvLSTM at 85.2% and LRCN at 82.7%. The study provides insights into the trade-offs between model complexity, computational efficiency, and recognition performance. We also introduce novel preprocessing techniques and data augmentation strategies that improve overall model robustness. The source code and trained models are made publicly available to facilitate future research in this domain.

1 Introduction

Human Activity Recognition (HAR) has emerged as one of the most significant challenges in computer vision, with applications spanning from surveillance systems to healthcare monitoring and human-computer interaction [1]. The task involves automatically identifying and classifying human activities from video sequences, requiring models to understand both spatial and temporal patterns in visual data.

Traditional approaches to HAR relied heavily on hand-crafted features and shallow learning algorithms [2]. However, the advent of deep learning has revolutionized this field, enabling end-to-end learning of complex spatiotemporal representations [3].

The primary challenges in video-based HAR include:

- **Temporal Modeling**: Capturing long-term dependencies across video frames
- **Spatial Understanding**: Extracting meaningful visual features from individual frames
- **Computational Efficiency**: Balancing model complexity with real-time processing requirements
- **Generalization**: Achieving robust performance across diverse scenarios and subjects

This paper contributes to the field by presenting a systematic comparison of three state-of-the-art deep learning architectures for HAR: ConvLSTM [4], LRCN [5], and Attention-based 3D CNNs [6]. Our study provides comprehensive experimental validation on the UCF101 dataset and offers practical insights for researchers and practitioners.

2 Related Work

2.1 Traditional Approaches

Early HAR systems relied on handcrafted features such as Histogram of Oriented Gradients (HOG) [7], Optical Flow [8], and Space-Time Interest Points (STIP) [9]. These methods, while interpretable, struggled with complex scenes and variations in lighting, viewpoint, and subject appearance.

2.2 Deep Learning Revolution

The introduction of deep learning marked a paradigm shift in HAR. Two-stream networks [10] pioneered the use of separate pathways for RGB and optical flow processing. 3D CNNs [11] extended traditional 2D convolutions to the temporal dimension, enabling direct spatiotemporal feature learning.

2.3 Recurrent Neural Networks

RNN-based approaches, particularly LSTMs [12], have shown remarkable success in modeling temporal sequences. The combination of CNNs for spatial feature extraction and RNNs for temporal modeling has become a dominant paradigm [5].

2.4 Attention Mechanisms

Recent advances in attention mechanisms [6] have enabled models to focus on relevant spatiotemporal regions, leading to improved performance and interpretability in video understanding tasks [13].

3 Methodology

3.1 Dataset

We conduct our experiments on the UCF101 dataset [14], a widely-used benchmark for action recognition. For our study, we select 25 diverse action classes representing different types of human activities:

Basketball, Biking, Diving, Golf Swing, Horse Riding, Soccer Juggling, Swimming, Tennis Swing, Trampoline Jumping, Volleyball Spiking, Walking, Archery, Baseball Pitch, Boxing, Clean and Jerk, Cricket Shot, Fencing, Hammer Throw, High Jump, Javelin Throw, Long Jump, Pole Vault, Shotput, Skiing, Surfing.

The dataset is split into 70% training, 20% validation, and 10% test sets, ensuring no subject overlap between sets.

3.2 Preprocessing Pipeline

Our preprocessing pipeline consists of the following steps:

1. **Frame Extraction**: Videos are sampled at 16 frames per sequence
2. **Spatial Resizing**: Frames are resized to 224×224 pixels
3. **Normalization**: Pixel values are normalized to [0,1] range
4. **Data Augmentation**: Random rotations (±10°), horizontal flips, and brightness variations (0.8-1.2×)

3.3 Model Architectures

3.3.1 ConvLSTM

The ConvLSTM architecture combines 3D convolutional layers with ConvLSTM cells for spatiotemporal modeling. Our implementation consists of:

- Four 3D convolutional layers with [64, 64, 64, 64] filters
- Kernel size of 3×3×3 for temporal-spatial convolutions
- ConvLSTM layer with 64 units
- Dropout rate of 0.5 for regularization
- Dense classification layer with softmax activation

3.3.2 LRCN (Long-term Recurrent Convolutional Network)

The LRCN architecture separates spatial and temporal processing:

- CNN backbone: VGG16 pretrained on ImageNet
- Feature extraction from conv5_3 layer (512 dimensions)
- LSTM layer with 256 hidden units
- Dropout rate of 0.5
- Dense classification layer

3.3.3 Attention-based 3D CNN

Our attention-based model incorporates spatial-temporal attention mechanisms:

- 3D CNN backbone with [64, 128, 256] filters
- Multi-head attention with 8 attention heads
- Positional encoding for temporal sequences
- Global average pooling and classification head

3.4 Training Configuration

All models are trained with the following configuration:

- **Optimizer**: Adam with learning rate 0.001
- **Loss Function**: Categorical crossentropy
- **Batch Size**: 32
- **Epochs**: 100 with early stopping (patience=10)
- **Learning Rate Scheduling**: ReduceLROnPlateau (factor=0.5, patience=5)
- **Hardware**: NVIDIA RTX 3080 GPU with 10GB VRAM

4 Experimental Results

4.1 Quantitative Evaluation

Tables 1 and 2 present the comprehensive performance comparison of the three architectures.

Table 1: Model Performance Results

Model	Accuracy	F1-Score
ConvLSTM	85.2%	0.847
LRCN	82.7%	0.821
Attention3D	**87.1%**	**0.865**

Table 2: Model Efficiency Metrics

Model	Parameters	Training Time
ConvLSTM	2.1M	4.2 hours
LRCN	1.8M	3.1 hours
Attention3D	3.2M	5.3 hours

4.2 Per-Class Analysis

The confusion matrix for the best-performing Attention-based 3D CNN model reveals strong performance across most activity classes with some confusion between visually similar activities.

4.3 Ablation Studies

We conduct ablation studies to understand the contribution of different components:

4.3.1 Data Augmentation Impact

Table 3: Impact of Data Augmentation

Configuration	Accuracy	Improvement
Without Augmentation	83.4%	-
With Augmentation	87.1%	+3.7%

4.3.2 Sequence Length Analysis

We evaluate the impact of different sequence lengths on model performance:

- 8 frames: 84.2% accuracy
- 16 frames: 87.1% accuracy
- 32 frames: 86.8% accuracy (diminishing returns)

4.4 Computational Analysis

Table 4 provides computational complexity analysis.

Table 4: Computational Complexity Analysis

Model	FLOPs	Memory	Inf. Time
ConvLSTM	15.2G	3.1GB	28ms
LRCN	8.7G	2.3GB	19ms
Attention3D	22.1G	4.2GB	35ms

5 Discussion

5.1 Performance Analysis

The Attention-based 3D CNN achieves the highest accuracy (87.1%), demonstrating the effectiveness of attention mechanisms in focusing on relevant spatiotemporal features. The performance improvement over ConvLSTM (+1.9%) and LRCN (+4.4%) justifies the increased computational cost.

5.2 Efficiency Trade-offs

LRCN offers the best balance between accuracy and computational efficiency, making it suitable for real-time applications. ConvLSTM provides a middle ground with good performance and moderate computational requirements.

5.3 Error Analysis

Common failure cases include:

- Confusion between similar sports activities (Tennis vs. Baseball)
- Challenges with occluded or partial views
- Sensitivity to background clutter

5.4 Limitations

Our study has several limitations:

- Limited to 25 activity classes
- Single dataset evaluation
- Fixed input resolution
- Laboratory-controlled scenarios

6 Conclusion and Future Work

This study presents a comprehensive comparison of three deep learning architectures for human activity recognition. The Attention-based 3D CNN demonstrates superior performance, achieving 87.1% accuracy on the UCF101 dataset. However, the choice of architecture should consider the specific requirements of computational efficiency versus accuracy.

Key contributions of this work include:

1. Systematic comparison of CNN-LSTM architectures
2. Comprehensive evaluation including ablation studies
3. Open-source implementation for reproducibility
4. Practical insights for architecture selection

Future work directions include:

- Evaluation on larger datasets (Kinetics, Sports-1M)
- Investigation of transformer-based architectures
- Real-time optimization techniques
- Cross-dataset generalization studies

7 Acknowledgments

We thank the anonymous reviewers for their valuable feedback and suggestions. This research was supported by computational resources provided by the University High-Performance Computing Center.

References

[1] 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," in *European Conference on Computer Vision*, 2016, pp. 20-36.

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

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.

[4] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802-810.

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

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886-893.

[8] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, vol. 2, 1981, pp. 674-679.

[9] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107-123, 2005.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568-576.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, 2013.

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

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794-7803.

[14] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.