6. Chen, X., Wang, L., & Huang, Z. (2023). Context-aware AI for adaptive augmented reality interfaces. IEEE Transactions on Visualization and Computer Graphics, 29(5), 2451–2465.
7. Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64. https://doi.org/10.1518/001872095779049543
8. Grand View Research. (2025). Augmented reality market size, share & trends analysis report, 2025–2030. Grand View Research.
9. Rudinska, O., & Kolesnikov, O. (2025). Menedzhment tekhnolohichnykh innovatsii u kvest-kimnatakh [Management of technological innovations in escape rooms]. Stalyi Rozvytok Ekonomiky, 5(56), 595–602. https://doi.org/10.32782/2308-1988/2025-56-82

DOI 10.70286/ISU-01.04.2026.007

DEEP LEARNING ARCHITECTURES FOR GESTURE RECOGNITION

Bielobrov Artur Oleksandrovich
Master's student in Computer Science
Hodovychenko Mykola Anatoliiovych
Ph.D., Associate Professor
Department of Information Systems
Odessa Polytechnic National University, Ukraine

The growing demand for natural and intuitive human-computer interaction (HCI) has stimulated substantial research interest in automated gesture recognition systems capable of interpreting human body movements from video sequences. Applications ranging from virtual and augmented reality (VR/AR) interfaces to contactless control systems, security monitoring, and medical rehabilitation increasingly rely on the ability to accurately classify complex spatio-temporal motion patterns without manual feature engineering. Deep learning has emerged as the dominant paradigm for this task, offering architectures that automatically learn hierarchical spatial and temporal representations from raw video data. However, the diversity of available deep learning architectures – including 3D Convolutional Neural Networks (3D CNN), hybrid CNN+LSTM models, and Transformer-based approaches – presents a non-trivial selection problem, as each architecture class exhibits distinct trade-offs between classification accuracy, computational cost, inference latency, and scalability. This paper presents a comparative analysis of these three architecture families for gesture recognition for intelligent human motion identification using the Jester video dataset [1].

Three-Dimensional Convolutional Neural Networks (3D CNN) represent a foundational approach to spatio-temporal video analysis. Unlike conventional 2D CNNs that process individual frames independently, 3D CNNs employ convolutional kernels of size (t × h × w) – where t denotes the temporal dimension – thereby

computing features that jointly encode spatial structure and inter-frame motion within a single operation. This architectural principle enables 3D CNNs to capture spatio-temporal patterns naturally, without requiring a separate temporal modelling stage. Notable variants include C3D [2], which applies 3D kernels across all layers; I3D (Inflated 3D ConvNets) [3], a hybrid approach that inflates pre-trained 2D kernels into 3D to reduce parameter count while preserving spatial knowledge; and 3D Temporal Dilation Convolution, which uses dilated kernels to expand the temporal receptive field without increasing the number of parameters. On the Jester dataset [1] – a benchmark comprising approximately 148,000 short gesture video clips across 25 classes — 3D CNN architectures achieve an accuracy of approximately 94% with a relatively compact parameter footprint ranging from 3.7 million to 50 million, and inference latency of 30-100 milliseconds. The primary strength of 3D CNN lies in its balanced trade-off between accuracy, implementation simplicity, and computational cost, making it a robust baseline for gesture recognition deployments. Its principal limitation is reduced scalability to very large datasets compared to attention-based architectures.

The CNN+LSTM hybrid architecture decouples spatial and temporal processing into two distinct stages. In the first stage, a 2D or 3D CNN backbone, such as ResNet, MobileNetV2, or EfficientNet, extracts spatial feature vectors from individual video frames or short frame groups, producing a sequence of compact representations. In the second stage, a Long Short-Term Memory (LSTM) network [4] processes this feature sequence to model temporal dependencies, leveraging its gating mechanisms (forget gate, input gate, output gate) to capture both short-term dynamics and long-range temporal relationships that are critical for gestures unfolding over extended time intervals. This two-stage design offers significant flexibility: the CNN backbone can be selected based on computational constraints (lightweight MobileNetV2 for edge devices versus deeper ResNet for server-side processing), while the LSTM's capacity can be adjusted independently through the number of hidden units and layers. Experimental comparisons reported in the internship study indicate that CNN+LSTM configurations achieve approximately 92% accuracy on the Jester dataset, with parameter counts ranging from 10 million to 100 million and latency of 50-150 milliseconds. While this accuracy is slightly lower than that of pure 3D CNN, the CNN+LSTM approach is particularly effective for dynamic gestures with complex temporal structure, where explicit sequential modelling of frame-level features yields superior discrimination of fine-grained motion patterns.

Transformer-based architectures represent the most recent and architecturally distinct paradigm for video understanding, replacing local convolutional operations and sequential recurrence with global self-attention mechanisms. The Vision Transformer (ViT) treats images as sequences of patches and applies standard Transformer encoding to learn inter-patch relationships, while its video extensions – TimeSformer, ViViT, LS-VIT, and RViT – adapt this principle to spatio-temporal data. TimeSformer [5] factorises attention into separate spatial and temporal dimensions, applying self-attention within each frame and across time independently, which reduces computational complexity compared to full 3D attention. ViViT [6] partitions video into spatial and temporal tokens processed by separate Transformer encoders and has demonstrated state-of-the-art results

on benchmarks such as UCF101 and Kinetics-400. LS-VIT (Long and Short-term Temporal Difference Vision Transformer) explicitly models both long-range and short-range motion through dedicated LMIM and SMIF modules, achieving higher efficiency than TimeSformer at lower FLOPS. RViT (Recurrent Vision Transformer) introduces a recurrent mechanism into the ViT framework for handling variable-length video sequences, achieving 92.31% accuracy on the Jester dataset – one of the highest reported results among Transformer models. The key advantages of Transformer architectures include a global receptive field (in contrast to the local nature of convolutions), parallelised processing (unlike sequential RNN computation), strong scalability on large datasets, and effective transfer learning to new tasks. Their primary disadvantage is substantially higher computational requirements: parameter counts typically exceed 80-150 million, and inference latency ranges from 70 to 300 milliseconds, making them less suitable for real-time deployment on resource-constrained devices.

Table 1. Performance comparison of deep learning architectures for gesture recognition on the Jester dataset.

Architecture	Accuracy (Jester)	Computational Complexity	Parameters	Latency (real-time)	Adaptability
3D CNN	~94%[1]	High (3D operations)	3.7M–50M	30-100 ms	Medium
CNN + LSTM	~92% [1]	High	10M–100M	50-150 ms	Medium
2D CNN (per-frame)	~80% [1]	Low	5M–25M	10-30 ms	Low
TimeSformer	~88% [1]	Very high	100M+	100-300 ms	High
LS-VIT	~90% [1]	Medium	80M–150M	70-300 ms	High
RViT	~92.31% [1]	High	85M–120M	80-250 ms	High
Two-Stream (CNN+OF)	~93% [1]	Very high (dual)	50M–150M	50-200 ms	Medium

The comparative analysis reveals that architecture selection for gesture recognition is fundamentally governed by the trade-off between accuracy, latency, and computational budget. For standard deployment scenarios where a balance between classification quality and resource consumption is required, 3D CNN emerges as the most robust choice: it achieves the highest single-architecture accuracy (~94%) on Jester while maintaining moderate latency (30-100 ms) and a manageable parameter count that permits execution on mid-range hardware. The availability of optimised variants such as I3D and 3D-TDC further enables tuning for specific constraints. CNN+LSTM configurations are preferable when temporal dependency modelling is critical – for instance, in applications involving gestures that unfold across multiple distinct phases – and when the flexibility to interchange CNN backbones is valued. Their moderate accuracy (~92%) and the ability to pair lightweight backbones (e.g., MobileNetV2) with compact LSTM heads make them a viable compromise for mobile and embedded deployments. Transformer architectures, despite higher resource demands, are the most promising direction for scenarios with sufficient computational capacity: they exhibit superior adaptability through transfer learning, scale effectively

to larger and more diverse datasets, and model long-range dependencies through global attention. RViT's accuracy of 92.31% demonstrates that Transformer-based models are already competitive with established CNN-based approaches on gesture benchmarks, and continued architectural innovations (such as factorised attention in LS-VIT) are progressively reducing their computational overhead.

Based on the research conducted within the analysis of deep learning architectures for gesture recognition on the Jester dataset, the following conclusions can be stated. First, 3D CNN architectures provide the most favourable accuracy-to-cost ratio for gesture recognition, achieving approximately 94% accuracy on the Jester dataset [1] with 3.7-50 million parameters and latency within the 30-100 ms range [2][3], which satisfies real-time constraints for most HCI applications. Second, CNN+LSTM hybrids achieve approximately 92% accuracy [1] while offering architectural flexibility through interchangeable CNN backbones and adjustable LSTM [4] capacity; this makes them suitable for scenarios where explicit temporal modelling is prioritised and deployment platforms vary in computational resources. Third, Transformer-based architectures (TimeSformer [5], ViViT [6], LS-VIT, RViT) represent the most scalable and adaptable paradigm, with RViT reaching 92.31% accuracy on Jester [1]; however, their parameter counts (85M-150M+) and latency (70-300 ms) currently restrict their deployment to environments with adequate GPU resources.

Finally, practical selection criteria should prioritise 3D CNN as the baseline architecture for constrained-resource deployments, CNN+LSTM configurations for applications demanding explicit sequential modelling of complex gesture dynamics, and Transformer-based models for research-oriented or high-capacity production environments where scalability, transfer learning, and long-range temporal reasoning are paramount.

References

1. Materzynska, J., Berger, G., Bax, I., & Memisevic, R. (2019). The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).
2. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 4489–4497.
3. Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6299–6308.
4. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.
5. Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? Proceedings of the International Conference on Machine Learning (ICML), 38, 813–824.
6. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 6836–6846.