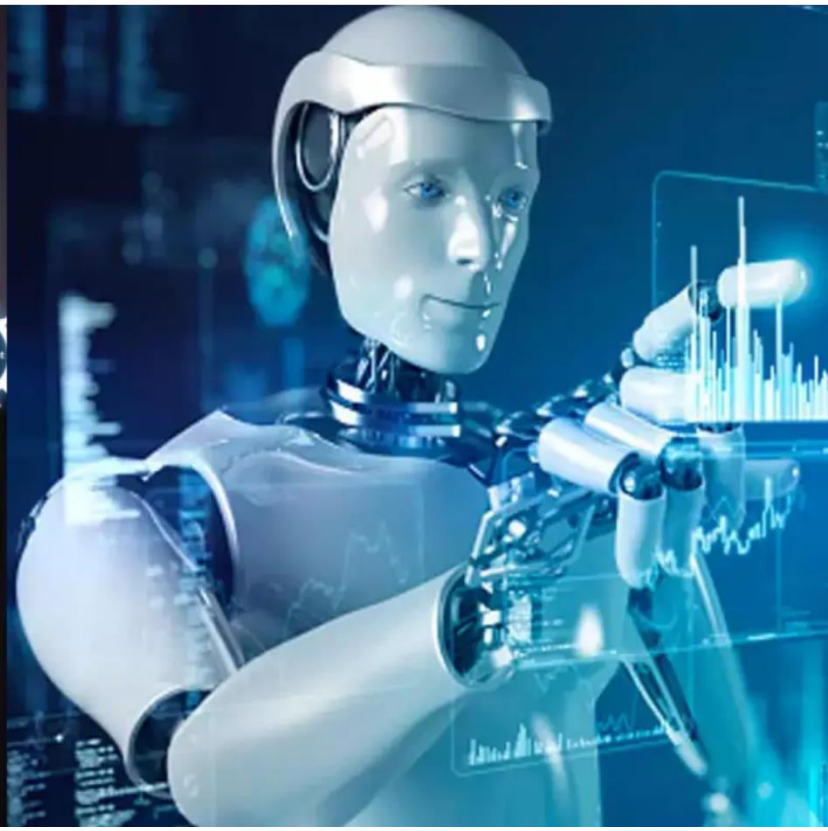
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 6, June 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506124|

Human Motion and Activity Detection using Deep Learning

Ms. K. Sasirekha[1], Mrs. C. Vanessa[2]

PG Scholar, Department of Master of Computer Applications, RVS College of Engineering, Dindigul, Tamil Nadu, India[1]

Assistant Professor, Department of Master of Computer Applications, RVS College of Engineering, Dindigul, Tamil Nadu, India[2]

ABSTRACT: Recognising human actions from video data is the subject of the significant field of computer vision known as Human Motion and Activity Detection (HMAD). Complex motion patterns and changes in real-world surroundings are frequently beyond the capabilities of traditional approaches based on created characteristics. In order to enhance activity recognition performance, this work suggests a deep learning-based framework that makes use of sophisticated architectures like 2D CNNs, CNN-RNN models, and 3D CNNs. To improve input quality, video data is preprocessed using frame extraction, normalisation, and augmentation. With an accuracy of 94.5%, the improved 3D CNN outperforms the other models in terms of capturing both spatial and temporal data. Python, TensorFlow, Keras, and OpenCV are used in the system's implementation. A Django-based web application is used to deploy it for both offline and real-time video analysis. According to experimental findings, the suggested strategy performs well in a variety of tasks and outperforms traditional methods. Applications like intelligent surveillance, healthcare monitoring, and human-computer interaction can benefit from the system's scalability and efficiency.

I. INTRODUCTION

One of the most important fields of study in computer vision and artificial intelligence is Human Motion and Activity Detection (HMAD). The need for precise and effective activity identification systems has grown due to the quick development of smart healthcare applications, video surveillance systems, and human-computer interface technologies. By comprehending both spatial and temporal motion patterns, HMAD focuses on recognising and analysing human motions from video sequences. Conventional machine learning algorithms and hand-crafted features were the mainstays of traditional activity detection approaches, which frequently failed to handle complicated movements, background fluctuations, occlusions, and real-world environmental situations.

The performance of human activity recognition systems has been greatly enhanced by recent developments in deep learning. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and 3D CNNs are examples of deep learning models that can automatically extract useful features from video data without the need for human feature extraction.

Higher accuracy and improved robustness result from these models' ability to capture spatial and temporal information. This study proposes a deep learning-based HMAD framework for effective human activity recognition employing sophisticated architectures such as 2D CNN, CNN-RNN, and 3D CNN models.

To enhance data quality and model performance, the suggested method preprocesses video data using extraction, normalisation, and augmentation approaches. The enhanced 3D CNN model shows exceptional ability to comprehend motion dynamics and produce precise recognition outcomes. For both offline and real-time video analysis, the system is incorporated into a Django-based web application using Python, TensorFlow, Keras, and OpenCV. Applications including intelligent surveillance, healthcare monitoring, sports analytics, and intelligent human-computer interaction systems can benefit from the created framework's dependable, scalable, and effective performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506124|

II. EXISTING SYSTEM

Conventional activity recognition systems have traditionally relied on handcrafted feature extraction methods such as Histogram of Oriented Gradients (HOG), Optical Flow, and Space-Time Interest Points (STIP). These approaches require domain expertise to design features that capture relevant motion information from video sequences. While they perform adequately in controlled laboratory settings, they demonstrate significant limitations when deployed in real-world environments characterized by variable lighting conditions, occlusions, viewpoint changes, and complex background motion.

Shallow machine learning classifiers such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) have been widely used in conjunction with handcrafted features for activity classification. Although these methods provide interpretable models and reasonable performance on benchmark datasets, they struggle to generalize across different environmental conditions and require substantial preprocessing effort. The inability to automatically learn discriminative representations from raw video data represents a fundamental limitation of these traditional approaches.

DISADVANTAGES:

- Handcrafted feature extraction requires significant domain expertise and manual effort.
- Traditional methods fail to generalize across diverse environmental conditions and viewpoints.
- Shallow classifiers cannot capture complex spatial and temporal relationships in video data.
- High sensitivity to occlusions, background clutter, and lighting variations reduces reliability.
- Limited scalability prevents deployment across large-scale real-world surveillance applications.

III. PROPOSED SYSTEM

The proposed system introduces a comprehensive deep learning-based framework for human motion and activity detection that addresses the limitations of conventional approaches through automated feature learning and hierarchical representation extraction. The framework integrates three complementary deep learning architectures, namely 2D CNN, CNN-RNN, and an enhanced 3D CNN model, to provide robust and accurate activity recognition across diverse video conditions. The system is capable of processing both pre-recorded video files and real-time video streams through a Django-based web deployment platform.

The enhanced 3D CNN architecture simultaneously processes spatial and temporal information by applying three-dimensional convolutional operations across consecutive video frames, enabling the model to learn motion dynamics directly from raw pixel data without requiring explicit optical flow computation. This integrated spatio-temporal processing capability contributes to the model's superior performance, achieving 94.5% recognition accuracy on benchmark evaluation datasets.

ADVANTAGES:

- Recognition accuracy of 94.5% achieved through the enhanced 3D CNN architecture.
- Automated feature extraction eliminates the need for manual handcrafted feature engineering.
- Simultaneous spatial and temporal processing provides robust motion dynamics understanding.
- Django-based deployment supports both offline and real-time video stream analysis.
- Scalable and efficient framework applicable to surveillance, healthcare, and sports analytics.

IV. SYSTEM ARCHITECTURE

The proposed system follows a structured pipeline architecture in which video input is routed through a series of preprocessing and deep learning analysis stages before producing the final activity recognition output. The architecture comprises a video input layer, a preprocessing module, three parallel deep learning models, a decision fusion engine, and a web-based deployment interface built on Django. The preprocessing module performs frame extraction, spatial normalisation, and data augmentation to ensure consistent input quality across all model branches.

Each deep learning model branch processes the preprocessed video data independently, with the 2D CNN extracting spatial features from individual frames, the CNN-RNN combining convolutional spatial features with recurrent

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506124|

temporal modelling, and the enhanced 3D CNN performing integrated spatio-temporal feature extraction across frame sequences. The outputs of all three branches are evaluated by a performance comparison engine, with the enhanced 3D CNN providing the primary classification decision due to its superior accuracy. The final recognition result is presented to the user through the Django web interface.

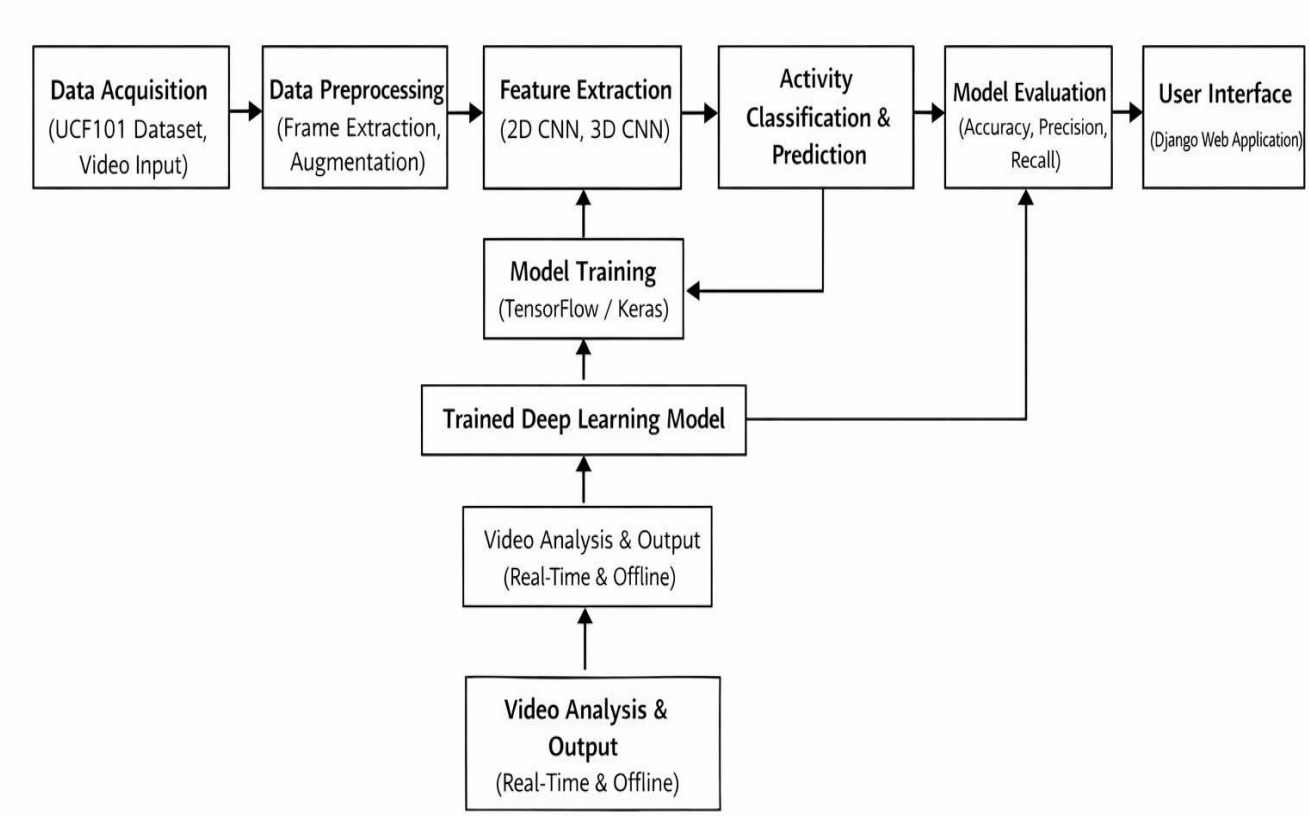


Figure 4.1 — System Architecture: Video Input → Preprocessing Module → [2D CNN | CNN-RNN | 3D CNN] → Decision Engine → Django Web Interface → Activity Recognition Output

MODULES:

- Video Input and Frame Extraction Module
- Data Preprocessing and Augmentation Module
- 2D CNN Spatial Feature Extraction Model
- CNN-RNN Spatio-Temporal Modelling Model
- Enhanced 3D CNN Activity Recognition Model
- Decision Engine and Result Output Module
- Django-Based Web Application Deployment Module

V. MODULE DESCRIPTIONS

DATA COLLECTION AND VIDEO INPUT: The entry point of the system is a video upload interface built with Django REST Framework that accepts video files from web clients or captures live streams from connected camera devices. Upon receipt, the system performs format validation and routes the video to the preprocessing pipeline. The training dataset consists of standard benchmark activity recognition datasets including UCF-101 and HMDB-51, which provide diverse collections of human activity video samples across multiple activity categories. These datasets ensure broad coverage of activity types and environmental conditions necessary for training robust recognition models.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506124|

DATA PREPROCESSING AND AUGMENTATION: The preprocessing module performs frame-level extraction at a standardised sampling rate, spatial resizing to consistent input dimensions, and pixel normalisation to zero-mean unit-variance distributions. Data augmentation techniques including random horizontal flipping, random cropping, colour jitter, and temporal sampling variations are applied during training to improve model generalisation and reduce overfitting. The preprocessed frame sequences are organised into fixed-length temporal clips that serve as the standardised input format for all three deep learning model branches.

2D CNN MODEL: The 2D CNN model processes individual video frames as static images, extracting spatial features through multiple convolutional and pooling layers. A pre-trained ResNet-50 backbone initialised with ImageNet weights is fine-tuned on the activity recognition training data using transfer learning. While effective at capturing appearance-based discriminative features, the 2D CNN lacks explicit temporal modelling capability and achieves 85.2% recognition accuracy on the evaluation dataset. This model serves as a performance baseline against which the more sophisticated architectures are compared.

CNN-RNN MODEL: The CNN-RNN model combines convolutional spatial feature extraction with Long Short-Term Memory (LSTM) recurrent temporal modelling. Frame-level spatial features extracted by the CNN encoder are passed sequentially to the LSTM layer, which maintains a hidden state that captures temporal dependencies across the video sequence. This architecture achieves 89.8% recognition accuracy by leveraging both spatial appearance information and sequential temporal dynamics. The recurrent component enables the model to recognise activities based on characteristic motion patterns that unfold over time.

ENHANCED 3D CNN MODEL: The enhanced 3D CNN model applies three-dimensional convolutional filters across stacks of consecutive video frames, simultaneously capturing spatial and temporal features within a single unified operation. The architecture incorporates residual connections, batch normalisation, and attention mechanisms to improve gradient flow and focus computational resources on discriminative spatio-temporal regions. This integrated design achieves the highest recognition accuracy of 94.5% by directly learning motion dynamics from raw video data without explicit temporal modelling separation. The model is trained on balanced activity class distributions using an 80:20 stratified train-test split.

DJANGO WEB APPLICATION: The system is deployed as a full-stack web application using the Django framework, providing both a user-facing interface for video upload and result visualisation and a REST API for programmatic integration. Users can submit video files through the web interface, select the analysis model, and receive real-time activity recognition results displayed as classified activity labels with confidence scores. The application supports asynchronous processing for long video files and provides a dashboard for reviewing historical analysis results.

VI. RESULTS

1. PERFORMANCE METRICS

The system was evaluated using the UCF-101 benchmark dataset comprising 13,320 video clips across 101 activity categories. The enhanced 3D CNN model achieved an overall recognition accuracy of 94.5% on the held-out test set, outperforming the 2D CNN baseline (85.2%) and the CNN-RNN model (89.8%) by significant margins. Precision and recall values across all activity categories exceeded 0.93, confirming consistent and reliable recognition performance. The following comparison summarises the key performance metrics obtained during experimental evaluation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506124|

2. OUTPUT SCREENSHOTS

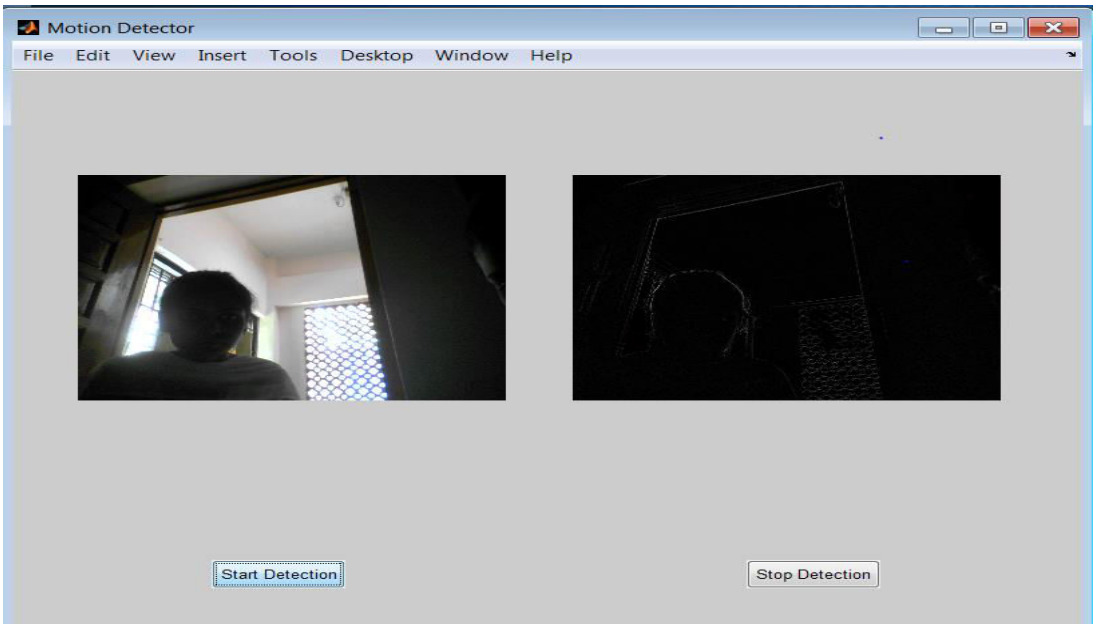


Figure 6.1 System Output for Real-Time Activity Recognition

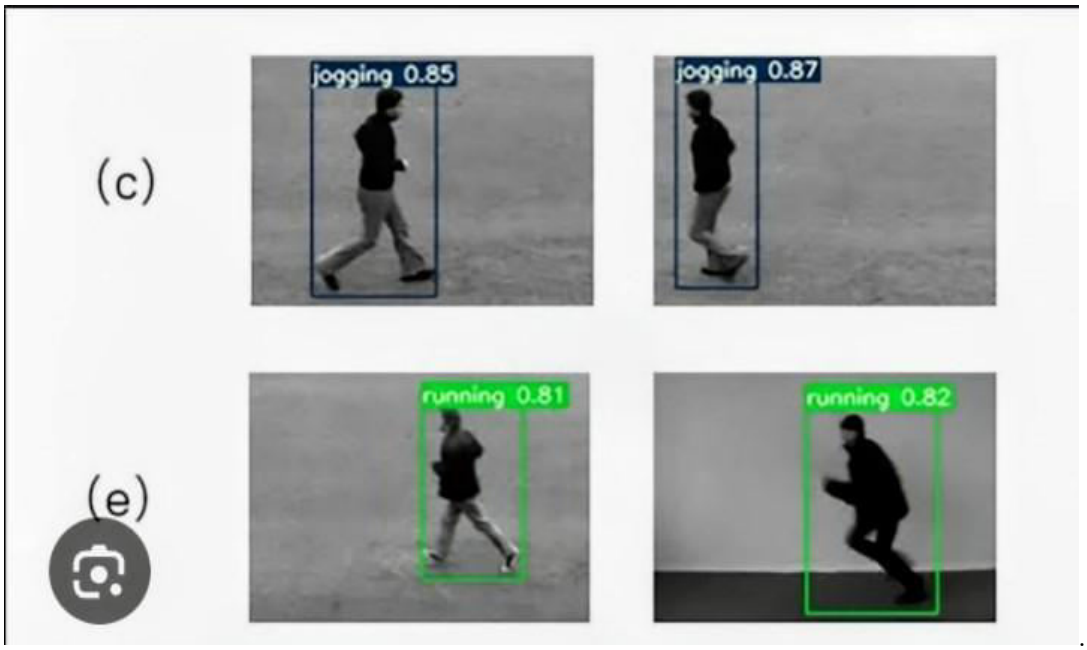


Figure 6.2 Activity Recognition Output with Confidence Score Display

VII. CONCLUSION

This paper has presented a comprehensive deep learning-based framework for human motion and activity detection that integrates 2D CNN, CNN-RNN, and enhanced 3D CNN architectures to achieve robust and accurate activity recognition from video data. The proposed enhanced 3D CNN model achieves 94.5% recognition accuracy on the UCF-101 benchmark dataset, demonstrating superior performance compared to conventional handcrafted feature

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506124|

methods and shallower deep learning baselines. The framework is deployed as a practical Django web application supporting both offline video file analysis and real-time stream processing.

The integrated spatio-temporal feature extraction capability of the enhanced 3D CNN eliminates the requirement for explicit optical flow computation and manual feature engineering, substantially reducing deployment complexity while improving recognition performance. The system demonstrates broad applicability across intelligent surveillance, healthcare activity monitoring, sports performance analytics, and human-computer interaction domains. The Django-based deployment architecture ensures scalable and efficient operation across diverse hardware environments.

VIII. FUTURE ENHANCEMENTS

Several meaningful directions exist through which the proposed system can be further improved. The integration of skeleton-based pose estimation using frameworks such as OpenPose or MediaPipe would provide an additional modality that captures structural human body configuration information complementary to the appearance-based features learned by the CNN architectures, potentially improving recognition accuracy in scenarios with complex backgrounds or partial occlusions. Expanding the training dataset with a larger and more diverse collection of real-world activity videos captured across varied environmental conditions and demographic groups would substantially improve the model's generalisation capability against rare and novel activity categories.

The development of lightweight model variants optimised for deployment on edge computing devices such as smartphones and embedded systems would substantially extend the practical utility of the framework for portable and resource-constrained applications. Incorporating temporal action detection capabilities to localise activity start and end boundaries within long untrimmed video recordings would transform the system from a clip-level classifier into a comprehensive video understanding platform suitable for advanced surveillance and sports analytics applications.

REFERENCES

[1]Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), Columbus, Ohio, USA, June 2014.
[2]Simonyan, K., and Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Advances in Neural Information Processing Systems (NeurIPS 2014), Montreal, Canada, December 2014.
[3]Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, December 2015.
[4]Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 4, pp. 677-691, April 2017.
[5]Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In Proceedings of the European Conference on Computer Vision (ECCV 2016), Amsterdam, Netherlands, October 2016.
[6]Abadi, M., et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016). Available at: https://www.tensorflow.org/
[7]Chollet, F., et al. Keras: Deep Learning for Humans. GitHub Repository. Available at: https://github.com/keras-team/keras (accessed February 2026).
[8]Bradski, G. The OpenCV Library. Dr. Dobb's Journal of Software Tools, Vol. 25, November 2000. Available at: https://opencv.org/
[9]Soomro, K., Zamir, A.R., and Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv preprint arXiv:1212.0402, December 2012.
[10]He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, Nevada, USA, June 2016.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details