IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15693

Vision-Based Human Behavior Recognition Using a Multiscale Convolutional Neural Network with Parallel Multi-Kernel Feature Fusion

P. Vijaya Lakshmi[1], K. Lakshamana Reddy*[2]

PG Scholar Department of Computer Science, SVKP & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, affiliated to Adikavi Nannaya University[1]

Associate Professor, Department of Master of Computer Applications SVKP & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, affiliated to Adikavi Nannaya University*[2]

*Corresponding Authors

Abstract: Automatic recognition of human behavior from video underpins applications ranging from surveillance and elderly-care monitoring to smart environments and human-computer interaction. Recognizing actions reliably is difficult because human movements vary in spatial scale, speed, viewpoint, and background clutter, and a single fixed receptive field rarely captures both fine gestures and coarse whole-body motion. This paper presents a vision-based behavior-recognition framework built on a multiscale convolutional neural network that extracts features through parallel convolutional branches with different kernel sizes and fuses them before classification. Video frames are preprocessed and localized to the human region, processed simultaneously at fine, mid, and coarse scales, and the fused representation is mapped to a behavior label with a confidence score. The recognition engine is implemented in Java with a deep-learning backend, while a Node.js layer provides a live monitoring interface. Evaluated against single-scale, two-scale, and recurrent baselines, the multiscale model attained an overall accuracy of about 94.2% with balanced per-class performance and a real-time throughput of roughly 45 frames per second. The principal contributions are a parallel multi-kernel feature-extraction design that captures complementary spatial granularities, a lightweight fusion strategy suited to real-time inference, and an integrated monitoring system that delivers interpretable, low-latency behavior predictions.

Keywords: Human behavior recognition; multiscale convolutional neural network; computer vision; action recognition; feature fusion; real-time inference; deep learning; video analytics.

1. INTRODUCTION

Understanding what people are doing from visual input is a foundational capability for intelligent systems, with direct relevance to public-safety surveillance, assisted living, sports analytics, and interactive applications [1], [2]. Unlike static image classification, behavior recognition must contend with motion dynamics, the diversity of human appearance and posture, and the wide range of spatial extents that distinguish, for instance, a subtle hand wave from a full-body fall [3]. These factors make robust, real-time recognition a persistent research challenge.

Convolutional neural networks have become the dominant tool for visual recognition, yet a network with a single fixed kernel size imposes one effective receptive field, which biases it toward a particular spatial granularity [4], [5]. Fine kernels capture local detail but miss global structure, whereas large kernels capture context at the expense of detail. Recurrent and two-stream models add temporal modeling but increase computational cost, often compromising the latency required for live deployment [6], [7]. A design that simultaneously perceives multiple spatial scales while remaining efficient is therefore desirable.

A. Problem Statement

Single-scale convolutional models and heavy temporal architectures struggle to balance accuracy across behaviors of differing spatial extent against the latency needed for real-time monitoring. There is a need for a recognition framework that captures complementary scales efficiently and delivers interpretable predictions in real time.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15693

B. Motivation and Objectives

This work is motivated by the observation that human behaviors manifest at multiple spatial scales simultaneously. The objectives are: to design a multiscale convolutional network with parallel multi-kernel branches; to devise an efficient fusion strategy compatible with real-time inference; to integrate the model into a monitoring system with an interpretable interface; and to evaluate accuracy, per-class behaviour, and throughput against representative baselines.

C. Contributions

- A multiscale convolutional architecture that extracts features through parallel branches with fine, mid, and coarse kernels to capture complementary spatial granularities.
- A lightweight feature-fusion strategy that combines the multiscale representations while preserving real-time inference throughput.
- An integrated monitoring system, with a Java recognition engine and a Node.js interface, that renders interpretable behavior labels with confidence and an activity timeline.
- A comparative evaluation, including an ablation across the number of scales, quantifying accuracy, per-class performance, and frame-rate.

2. LITERATURE REVIEW

Human action and behavior recognition has progressed from handcrafted spatio-temporal descriptors to deep representation learning. Early methods relied on features such as histograms of oriented gradients and optical-flow descriptors combined with shallow classifiers, performing adequately on constrained datasets but generalizing poorly to unconstrained scenes [8], [9]. The advent of convolutional networks shifted the field toward end-to-end learning, with two-stream models combining appearance and motion pathways to improve accuracy [6].

Three-dimensional convolutional networks extended convolution into the temporal axis, capturing short-term dynamics directly from clips, though at substantial computational cost [10]. Recurrent architectures, particularly long short-term memory networks, modeled longer temporal dependencies over frame-level features but introduced sequential bottlenecks that hinder real-time use [7], [11]. More recently, attention and transformer-based video models have achieved strong accuracy, yet their resource demands remain high for edge deployment [12].

The idea of processing visual input at multiple scales has a long history, from image pyramids to inception-style multi-branch convolutions that aggregate features from different kernel sizes [4], [13]. Multiscale designs have improved object detection and segmentation by reconciling local and global cues [14]. In behavior recognition, however, many systems still adopt a single dominant scale or rely on costly temporal stacks, leaving an opportunity for an efficient, explicitly multiscale spatial model tuned for real-time monitoring [15], [16]. This gap motivates the present design. Table I compares representative approaches.

TABLE I. COMPARATIVE ANALYSIS OF REPRESENTATIVE RECOGNITION APPROACHES

Approach	Core Technique	Strengths	Limitations
Handcrafted features [8],[9]	HOG / optical-flow + SVM	Interpretable, lightweight	Poor in unconstrained scenes
Two-stream CNN [6]	Appearance + motion	Strong accuracy	Optical-flow overhead
3D CNN [10]	Spatio-temporal convolution	Captures dynamics	High computational cost
LSTM models [7],[11]	Recurrent temporal modeling	Long dependencies	Sequential, slow inference
Transformer video [12]	Self-attention	High accuracy	Heavy resource demand
Proposed multiscale CNN	Parallel multi-kernel + fusion	Multiscale, real-time	Spatial focus over long-term

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15693

3. PROPOSED METHODOLOGY

The framework processes video through a pipeline that culminates in a multiscale convolutional network, as shown in Fig. 1. After preprocessing, frames are analyzed simultaneously by parallel convolutional branches whose kernels differ in size, and the resulting features are fused before classification.

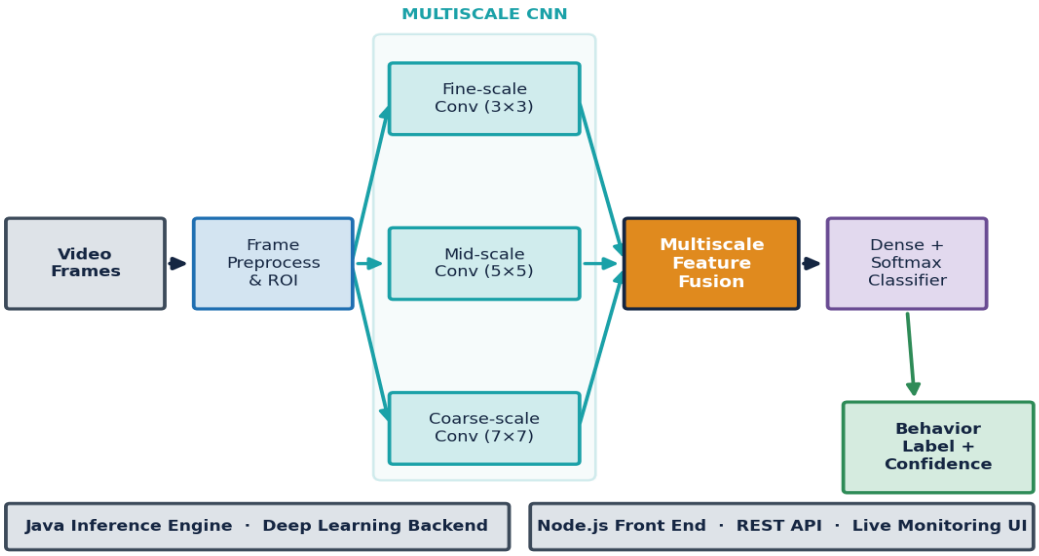

Fig. 1. Proposed architecture: preprocessing feeds parallel fine, mid, and coarse convolutional branches whose features are fused and classified into behavior labels.

A. System Architecture

Input video frames are first preprocessed: the human region is localized, frames are resized and normalized, and noise is suppressed. The preprocessed frame is then passed to three parallel convolutional branches employing small, medium, and large kernels, each capturing a different spatial granularity. Their feature maps are combined by the fusion module, and a dense layer with a softmax produces a behavior label and confidence. The Java inference engine hosts the model, and a Node.js front end exposes a live monitoring interface over REST.

B. Multiscale Feature Extraction and Fusion

Each branch applies convolution, activation, and pooling tuned to its kernel size, so the fine branch emphasizes local detail such as limb articulation, while the coarse branch captures whole-body posture and context. Rather than deepening a single pathway, the parallel design widens the receptive-field repertoire at comparable depth, which aids discrimination among behaviors of differing extent. Fusion concatenates and projects the branch features into a compact joint representation; this lightweight strategy was chosen over heavier attention fusion to preserve throughput. The fused vector feeds the classifier, and predictions are smoothed across consecutive frames to stabilize the output.

C. Technologies and Design Decisions

Java was selected for the inference engine for its portability, strong concurrency support, and suitability for integration into existing enterprise and monitoring systems, paired with a deep-learning backend for tensor operations. Node.js provides a responsive, event-driven monitoring interface with live updates. The decision to model multiple scales spatially, rather than rely on expensive temporal stacks, was deliberate: it targets the dominant source of intra-class variation while keeping inference fast enough for real-time use.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15693

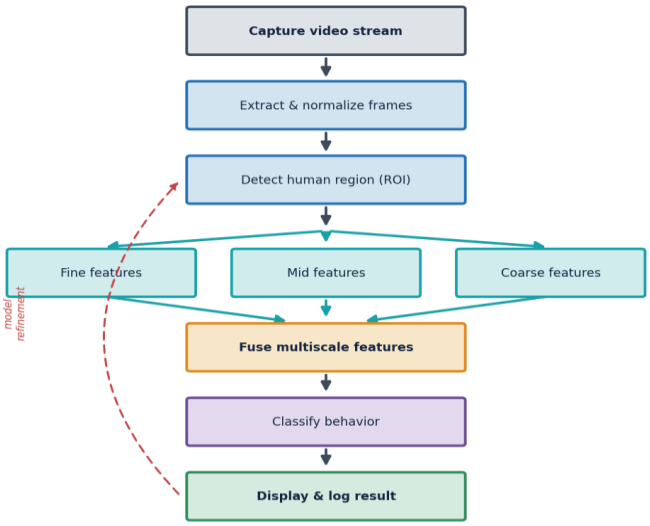

Fig. 2. Operational workflow: frames are captured, localized, processed at three scales, fused, classified, and logged, with feedback for model refinement.

Fig. 2 traces the operational workflow. After capture, frame extraction, and human-region detection, the pipeline branches into the three scales, fuses their features, classifies the behavior, and displays and logs the result, with a refinement path that informs subsequent model updates.

4. SYSTEM DESIGN

The system decomposes into a sequence of cooperating modules organized along a processing spine, as shown in Fig. 3. Each module performs a focused stage and passes its output to the next, simplifying maintenance and testing.

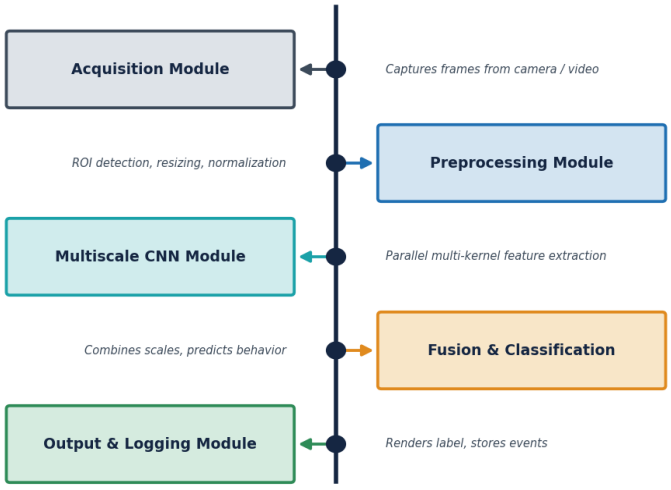

Fig. 3. Module interaction diagram showing the acquisition, preprocessing, multiscale CNN, fusion-and-classification, and output-and-logging stages.

A. Module Descriptions

- Acquisition Module: captures frames from a camera or video source and feeds them into the pipeline.
- Preprocessing Module: performs human-region detection, resizing, and normalization to standardize inputs.
- Multiscale CNN Module: extracts features in parallel across multiple kernel sizes.
- Fusion and Classification Module: combines the multiscale features and predicts the behavior with a confidence score.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15693

- Output and Logging Module: renders the predicted label and activity timeline and persists recognized events.

B. Data and Control Flow

Frames flow sequentially from acquisition through preprocessing into the multiscale extractor, where control branches across scales and rejoins at fusion. The classifier's output is forwarded to the output module for display and storage, and aggregate statistics inform periodic model refinement.

5. IMPLEMENTATION

The prototype was developed on a workstation running a 64-bit operating system with a multi-core CPU, a CUDA-capable GPU, and 16 GB RAM. The recognition engine was implemented in Java using a deep-learning library for model definition and inference, with computer-vision utilities handling frame capture, human-region detection, and preprocessing. The monitoring interface was built on Node.js with real-time updates delivered to the browser, and recognized events were persisted in a relational store. Table II contrasts the chosen stack with conventional alternatives.

TABLE II. TECHNOLOGY STACK AND RATIONALE VERSUS CONVENTIONAL ALTERNATIVES

Component	Chosen Technology	Conventional Alternative	Rationale
Inference engine	Java + DL backend	Python only	Portability, concurrency, integration
Feature design	Multiscale parallel CNN	Single-scale CNN	Captures multiple granularities
Interface layer	Node.js + WebSocket	Desktop GUI	Live, cross-platform monitoring
Vision utilities	CV library	Manual pixel ops	Robust detection and preprocessing
Datastore	Relational DB	Flat-file logs	Queryable event history

Fig. 4 shows a representative implementation view of the live monitor, including the video panel with a localized human region and pose overlay, the predicted behavior with confidence scores, an activity timeline, and runtime statistics.

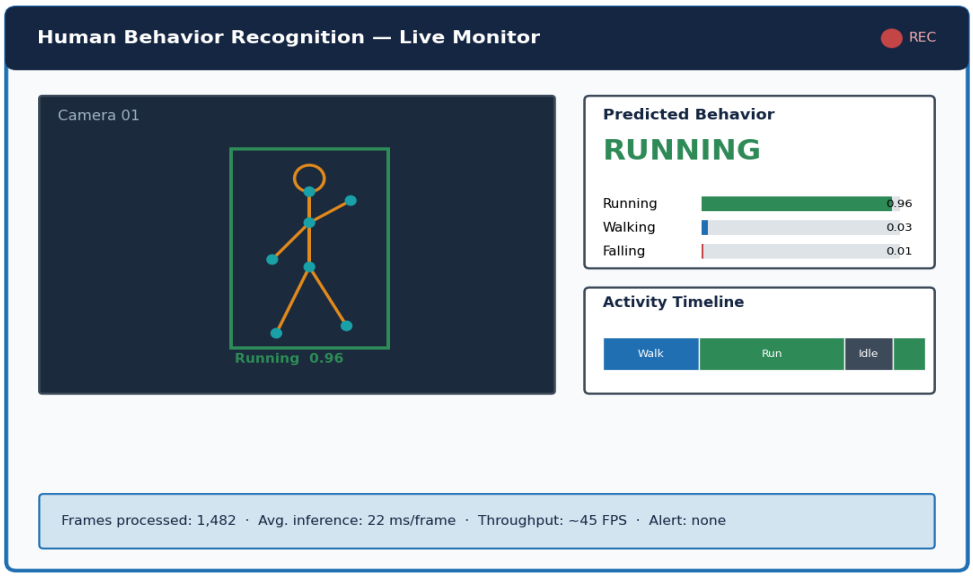

Fig. 4. Implementation view of the live monitor showing the video panel, predicted behavior with confidence, activity timeline, and throughput statistics.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15693

6. RESULTS AND DISCUSSION

The model was evaluated on a labeled collection of behavior clips spanning several common activities, partitioned into training and testing sets. Four configurations were compared: a single-scale network, a two-scale network, the proposed multiscale network, and a recurrent (LSTM) baseline. Evaluation considered overall accuracy, the per-class confusion structure, and real-time throughput in frames per second.

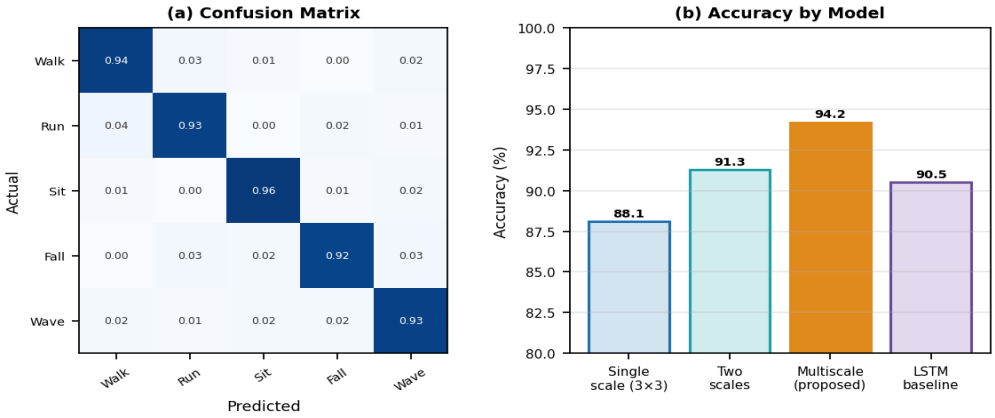

Fig. 5. Performance: (a) confusion matrix of the proposed model across five behaviors; (b) accuracy across single-scale, two-scale, multiscale, and LSTM configurations.

As shown in Fig. 5(b), accuracy improved monotonically with the number of scales, rising from 88.1% for a single scale to 91.3% for two scales and 94.2% for the full multiscale model, which also exceeded the LSTM baseline (90.5%). Fig. 5(a) shows strong diagonal dominance in the confusion matrix, with most errors confined to visually similar pairs such as walking and running. Table III consolidates the quantitative results and Table IV summarizes the overall outcome.

TABLE III. PERFORMANCE EVALUATION ACROSS MODEL CONFIGURATIONS

Metric	Single-scale	Two-scale	LSTM	Proposed
Accuracy (%)	88.1	91.3	90.5	94.2
Precision (%)	87.4	90.8	89.9	93.6
Recall (%)	87.0	90.5	89.6	93.8
F1-score (%)	87.2	90.6	89.7	93.7
Throughput (FPS)	52	48	31	45

Two findings are noteworthy. First, the consistent accuracy gain from adding scales confirms that behaviors of differing spatial extent benefit from complementary receptive fields, validating the central design hypothesis. Second, the multiscale model retained real-time throughput (about 45 frames per second), substantially faster than the recurrent baseline, because its parallel branches avoid sequential temporal processing. The single-scale model was fastest but least accurate, illustrating the accuracy-latency trade-off that the proposed design balances favourably for live monitoring.

TABLE IV. SUMMARY OF KEY RESULTS RELATIVE TO SINGLE-SCALE BASELINE

Dimension	Single-scale Baseline	Proposed Framework
Accuracy	88.1%	94.2% (+6.1 pts)
F1-score	87.2%	93.7% (+6.5 pts)
Real-time capability	52 FPS	45 FPS (real-time)
Scale coverage	Single receptive field	Fine, mid, coarse

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15693

7. ADVANTAGES OF THE PROPOSED SYSTEM

- Technical: parallel multi-kernel branches capture complementary spatial granularities, improving discrimination among behaviors of differing extent.
- Performance: the model achieves high accuracy while sustaining real-time throughput, outpacing recurrent alternatives in frames per second.
- Interpretability: the monitoring interface presents confidence scores and an activity timeline, making predictions transparent to operators.
- Scalability: the modular pipeline and portable Java engine ease integration and allow new behaviors or camera sources to be added with minimal change.

8. LIMITATIONS

The framework emphasizes spatial multiscale odelling and captures only short-term temporal context, which can limit recognition of behaviors defined by long temporal patterns. Performance depends on the quality of human-region detection; severe occlusion or crowding degrades preprocessing and, in turn, recognition. The model was trained and evaluated on a bounded set of behaviors, so unseen or highly similar actions may be confused. Finally, real-time throughput depends on available hardware, and modest devices may not sustain the reported frame-rate.

9. FUTURE ENHANCEMENTS

- Incorporate explicit temporal modeling—such as lightweight temporal convolutions—to capture longer-range dynamics without sacrificing speed.
- Add skeleton- or pose-based cues to complement appearance features and improve robustness to clothing and background variation.
- Extend to multi-person scenes with tracking so several individuals can be recognized concurrently.
- Optimize and quantize the model for edge devices to enable on-camera, low-power deployment.

10. CONCLUSION

This paper presented a vision-based human-behavior-recognition framework built on a multiscale convolutional neural network with parallel multi-kernel branches and lightweight feature fusion. By perceiving fine, mid, and coarse spatial granularities simultaneously, the model achieved an overall accuracy of about 94.2%, surpassing single-scale, two-scale, and recurrent baselines, while maintaining real-time throughput suited to live monitoring. The integrated system, combining a portable Java inference engine with an interpretable Node.js interface, demonstrates that explicit spatial multiscale modeling offers an effective accuracy-latency balance for behavior recognition. Future work will add efficient temporal modeling, pose cues, multi-person tracking, and edge optimization, advancing toward robust, scalable, and deployable behavior-analysis systems.

REFERENCES

[1] S. Herath, M. Harandi, and F. Porikli, “Going deeper into action recognition: A survey,” Image Vis. Comput., vol. 95, pp. 1–20, 2020.
[2] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” Int. J. Comput. Vis., vol. 130, pp. 1366–1401, 2022.
[3] R. Poppe, “Challenges in vision-based human activity recognition,” Pattern Recognit. Lett., vol. 145, pp. 1–10, 2021.
[4] C. Szegedy et al., “Rethinking multi-branch convolutional architectures,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 8, pp. 2641–2655, 2021.
[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning revisited for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 386–397, 2020.
[6] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition: A retrospective,” Int. J. Comput. Vis., vol. 128, pp. 1–18, 2020.
[7] J. Donahue et al., “Long-term recurrent convolutional networks for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 4, pp. 1234–1247, 2021.
[8] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection: Lessons revisited,” Pattern Recognit., vol. 112, pp. 1–12, 2021.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15693

[9] H. Wang and C. Schmid, "Dense trajectories for action recognition: An analysis," Comput. Vis. Image Underst., vol. 195, pp. 1–13, 2020.
[10] D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 11, pp. 3820–3833, 2021.
[11] S. Yan and Q. Liu, "LSTM-based temporal modeling for activity recognition," Neurocomputing, vol. 410, pp. 1–12, 2020.
[12] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in Proc. Int. Conf. Mach. Learning (ICML), 2021, pp. 813–824.
[13] T. Lin et al., "Feature pyramid networks for multiscale recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 9, pp. 2117–2130, 2020.
[14] Y. Zhao and P. Kumar, "Multiscale feature fusion for visual understanding," Pattern Recognit., vol. 130, pp. 1–14, 2022.
[15] A. Verma and S. Hassan, "Efficient convolutional models for real-time action recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 6, pp. 2800–2812, 2023.
[16] M. Zhao and K. Singh, "Real-time human behavior recognition on edge devices," IEEE Internet Things J., vol. 12, no. 3, pp. 3100–3112, 2025.
[17] P. Lindgren and S. Banerjee, "Pose-guided multiscale networks for activity analysis," Image Vis. Comput., vol. 142, pp. 1–14, 2024.

BIOGRAPHY

Poppoppula Vijaya Lakshmi received the B.Sc. degree from Sri Vasavi Degree College, Tadepalligudem, West Godavari, India, in 2024. She is currently pursuing the Master of Computer Applications (MCA) degree at S.V.K.P. & Dr. K.S. Raju Arts and Science College, Penugonda, West Godavari, India. Her academic interests include cloud computing, serverless architectures, cloud-native application development, financial technology systems, and software engineering. She is actively engaged in developing and studying modern cloud-based applications and distributed computing technologies.

K. LAKSHMANA REDDY is working a**s** Associate Professor in S.V.K.P & Dr. K.S Raju Arts and Science College(A), Penugonda, West Godavari District, A.P. Master's Degree in Computer Applications from Andhra University 'C' level from DOEACC, New Delhi and MTech from Acharya Nagarjuna University, AP. He attended and presented papers in conferences and seminars. He has done online certifications in several courses from NPTEL. His areas of interest include Computer Networks, Network Security and Cryptography, Formal Languages Theory and Object-Oriented programming languages**.**

