InSight Bulletin: A Multidisciplinary Interlink International Research Journal
Peer Reviewed International, Open Access Journal.
ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-2 / February - 2026

Original Article

A study of an AI-Based System for Action Recognition and Human Pose Estimation

Swati Santosh Jamble
ATSS College of Business Studies and Computer Application, Chinchwad, Pune

Manuscript ID:
IBMIIRJ -2026-030243

Submitted: 10 Jan. 2026

Revised: 15 Jan. 2026

Accepted: 18 Feb. 2026

Published: 28 Feb. 2026

ISSN: 3065-7857

Volume-3

Issue-2

Pp. 209-212

February 2026

Correspondence Address:
Swati Santosh Jamble
ATSS College of Business Studies and Computer Application, Chinchwad, Pune
Email: swatiborkar1@gmail.com

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.21357452

DOI Link:
https://doi.org/10.5281/zenodo.21357452

Abstract

The rapid evolution of artificial intelligence has significantly advanced the capabilities of computer vision systems, particularly in the domains of human pose estimation and action recognition. This study examines the design and conceptual development of an AI-driven framework capable of identifying human body postures and recognizing actions from image sequences and video streams in real time. Human pose estimation focuses on detecting anatomically significant joint locations, whereas action recognition involves categorizing observed motion patterns into semantic activity classes. The proposed framework integrates convolutional neural networks for spatial feature extraction, transformer-based architectures for long-range spatiotemporal reasoning, and recurrent neural networks for modeling temporal dependencies. The paper outlines the system architecture, methodological pipeline, dataset selection, preprocessing strategies, and evaluation metrics, while also discussing potential applications in healthcare, intelligent surveillance, and human–computer interaction.

Keywords: *AI-enabled vision systems, deep neural networks, convolutional and transformer architectures, human pose inference, activity and action classification, computer vision applications.*

Introduction

1. Motivation

Recognizing human actions and estimating body poses are essential tasks in computer vision, with applications in areas like intelligent surveillance, medical rehabilitation, sports analysis, and interactive systems. Accurate identification of human activities allows for proactive monitoring in security settings, while precise pose estimation supports biomechanical evaluations and assistive healthcare tools (Carreira & Zisserman, 2017). Despite substantial advancements, these tasks are still difficult due to the complexity of body movements, changes in the environment, and frequent instances of objects blocking the view.

2. Problem Definition

Although recent deep learning models have shown good results, most existing methods have trouble maintaining accuracy in real-world conditions such as different viewpoints, partial visibility of the body, and the need for fast processing. Moreover, the ability to work effectively in various environments and handle different situations remains a challenge (Donahue et al., 2015). Because of this, there is an increasing need for a single, efficient system that can deal with these problems while ensuring high levels of recognition accuracy.

3. Research Objectives

The main goal of this study is to examine an AI-based framework that can perform real-time human pose estimation and action recognition using visual data. The system is designed to operate reliably in difficult situations, such as when parts of the body are hidden, lighting conditions change, or the background is moving.

4. Scope of the Study

This research focuses on combining convolutional neural networks to learn spatial features, pose estimation models like OpenPose and HRNet for identifying key body points, and sequence modeling methods, including recurrent neural networks and transformers, for classifying actions over time. The study also investigates preprocessing techniques for datasets and evaluation strategies to measure how well the system performs.

How to cite this article:
Jamble, S. S. (2026). A study of an AI-Based System for Action Recognition and Human Pose Estimation. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(2), 209–212. https://doi.org/10.5281/zenodo.21357452

Literature Review

The importance of advancements in human action recognition and pose estimation have emerged alongside the development of deep learning and the availability of large-scale annotated datasets. A landmark contribution by Carreira and Zisserman (2017) introduced the Kinetics dataset and the Inflated 3D ConvNet (I3D), enabling effective spatiotemporal feature extraction for large-scale action recognition. Earlier, Simonyan and Zisserman (2014) proposed the Two-Stream CNN architecture, which independently models spatial appearance and temporal motion through RGB frames and optical flow.

To capture temporal dynamics more effectively, Donahue et al. (2015) demonstrated that combining CNNs with long short-term memory (LSTM) networks allows joint modeling of visual appearance and motion sequences. Similarly, Tran et al. (2015) introduced 3D convolutional networks (C3D), which learn spatial and temporal features simultaneously, achieving strong results on benchmark datasets.

Parallel progress has been observed in pose estimation research. Newell et al. (2016) proposed the Stacked Hourglass Network, which employs repeated bottom-up and top-down processing to refine joint predictions. For real-time multi-person scenarios, Cao et al. (2017) developed OpenPose, leveraging Part Affinity Fields to efficiently associate detected joints across individuals. Advances in three-dimensional pose estimation further expanded the field. Sun et al. (2018) utilized the Human3.6M dataset to train deep networks capable of estimating 3D joint positions from monocular images. Kanazawa et al. (2018) introduced graph convolutional networks to encode skeletal structure, improving the transition from 2D detections to 3D pose representations.

Recent studies have explored joint modeling of pose and action. Ke et al. (2017) demonstrated that integrating pose trajectories with appearance features in a CNN–RNN framework enhances action classification accuracy. Transformer-based approaches further advanced this integration by employing self-attention mechanisms to capture long-range spatiotemporal dependencies (Xu et al., 2020; Vaswani et al., 2017). Modern research trends also highlight multimodal fusion and graph-based modeling to improve robustness and interpretability (Liu et al., 2020; Wang et al., 2020).

System Design

1. **Overview**

The proposed framework comprises two interconnected components:

1. Pose Estimation Module: Responsible for detecting anatomical keypoints using CNN-based architectures such as OpenPose and HRNet.
2. Action Recognition Module: Utilizes temporal models to classify human actions based on extracted pose and appearance features.

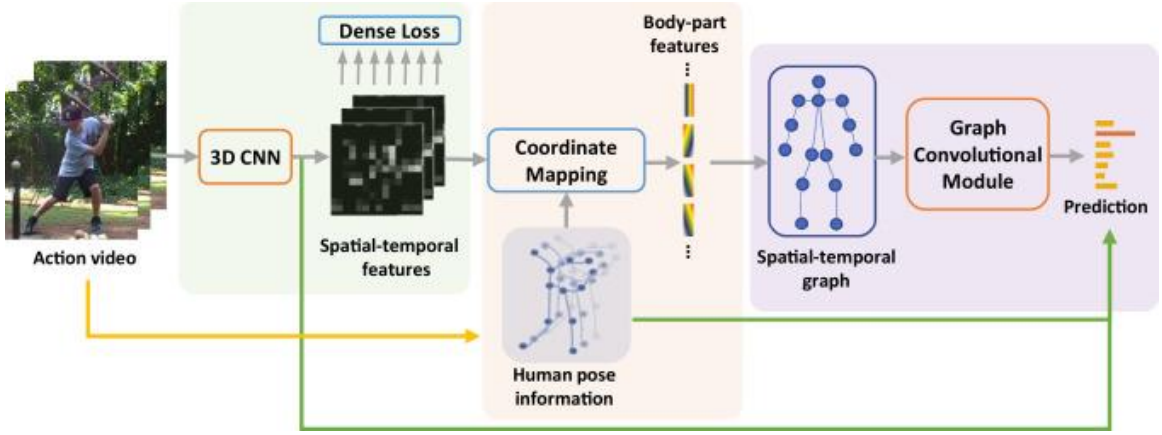

Figure 1: System architecture illustrating the interaction between pose estimation and action recognition modules (adapted from Carreira & Zisserman, 2017).

2. **Pose Estimation Module**

Pose estimation is performed using OpenPose and HRNet, which are designed to maintain high localization accuracy even under occlusions and illumination variations (Cao et al., 2017; Newell et al., 2016). These architectures rely on multi-scale feature representations and high-resolution feature maps to predict joint coordinates.

3. **Feature Extraction**

Convolutional neural networks are employed to extract hierarchical spatial features from video frames. These features encode structural and contextual information necessary for precise keypoint localization and subsequent action recognition.

4. **Action Recognition Module**

Temporal dependencies across successive frames are modeled using LSTM and GRU networks. In addition, transformer architectures are combined to capture long-range temporal relationships more effectively, leveraging self-attention mechanisms for improved categorization understanding (Vaswani et al., 2017).

5. **Fusion Strategies**

Early Fusion: Combines pose and presence features at the demonstration level.

Late Fusion: Incorporates predictions from individual modules at the decision stage.

Datasets and Preprocessing

1. **Datasets**

Commonly used action recognition datasets include UCF101, HMDB51, and Kinetics (Carreira & Zisserman, 2017; Simonyan & Zisserman, 2014). For pose estimation, datasets such as COCO, MPII, and Human3.6M are widely adopted, with the latter supporting 3D pose analysis (Sun et al., 2018).

2. **Preprocessing Techniques**

Preprocessing steps include data augmentation through rotation, scaling, flipping, and noise injection to enhance robustness. Video frames are sampled strategically to reduce computational overhead, while pose keypoints are normalized to ensure consistency across varying resolutions and body proportions.

Methodology

The pose estimation process uses OpenPose and HRNet to identify 2D and 3D joint positions through high-resolution data (Cao et al., 2017; Newell et al., 2016).For action recognition, recurrent and transformer-based models are used to study movement patterns and how posture changes over time (Vaswani et al., 2017). Both early and late fusion methods are explored to combine information from pose and action in an effective way.

Evaluation Strategy

1. Performance Metrics

Pose estimation accuracy is measured using metrics like mean Average Precision (mAP), Percentage of Correct Keypoints (PCK), and joint error in localization (include citation if metric definitions are modified).Action recognition effectiveness is measured using accuracy, precision, recall, and F1-score (Simonyan & Zisserman, 2014).

2. Experimental Setup

The experiments are carried out on standard datasets using environments that take advantage of GPU processing.The data is divided into training, validation, and testing sets to fine-tune and assess the model's performance.

Application Domains

This framework can be used in several areas, such as monitoring health, smart surveillance, analyzing sports performance, and creating gesture-based ways for people to interact with computers (Carreira & Zisserman, 2017; Cao et al., 2017).

Challenges and Future Research Directions

Some ongoing challenges involve dealing with heavy occlusions, ensuring real-time processing without losing accuracy, and working with large and varied datasets. Future work might look into using multiple types of sensors, designing more efficient transformer models, and using synthetic data to improve performance (Vaswani et al., 2017).

Conclusion

This study thoroughly examines an AI-based system that integrates human pose estimation and action recognition. By using convolutional feature extractors, pose estimation networks, recurrent models, and transformer structures, the system shows strong potential for real-time analysis of human activities. Future improvements will concentrate on making the system more efficient, integrating multiple data sources, and enhancing how well it understands sequences over time.

Acknowledgment

The author would like to express sincere gratitude to the faculty and management of ATSS College of Business Studies and Computer Application, Chinchwad, Pune, for their academic guidance and support during the preparation of this research work.
Special thanks are extended to colleagues and peers for their valuable discussions and encouragement, which helped in improving the quality of this study.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6296–6305). IEEE.
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2017). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7291–7299). IEEE.
3. Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2625–2634). IEEE.
4. Kanazawa, A., Sharma, A., Srinivasan, P. P., & Malik, J. (2018). 3D human pose estimation from a single image via graph convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 71–80). IEEE.
5. Ke, X., Li, Q., & Qiao, Y. (2017). Joint action recognition and pose estimation using CNN–RNN networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 3167–3175). IEEE.

6. Liu, Z., Wang, D., & Xu, G. (2020). Multimodal action recognition with RGB and depth images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8), 1934–1946. https://doi.org/10.1109/TPAMI.2019.2904515
7. Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 483–499). Springer.
8. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 568–576).
9. Sun, X., Yang, Y., & Tang, X. (2014). Human3.6M: Large-scale datasets and methods for 3D human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(2), 317–328.
10. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 4489–4497). IEEE.
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS).
12. Wang, W., Li, Y., Liu, Z., & Wu, L. (2020). Graph neural networks for human pose estimation and action recognition. IEEE Transactions on Neural Networks and Learning Systems, 31(10), 3951–3963.
13. Xu, X., Zhang, Z., & Li, X. (2020). Action recognition with pose estimation using transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1721–1730). IEEE.