Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ23495

Analysis of Human Action Recognition with Deep Learning

Mo Zhou

College of Computer Science, Nankai University, Tianjin, China
2211349@mail.nankai.edu.cn

Abstract: Human Action Recognition (HAR) is a critical research area in computer vision, with notable applications in security surveillance, health monitoring, and human-computer interaction. This study explores the latest advancements in deep learning techniques applied to HAR over the past decade. This study also presents a comprehensive framework that incorporates various advanced methodologies, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) enhanced with Long Short-Term Memory (LSTM) and attention mechanisms, Graph Convolutional Networks (GCN), and Transformer architectures. The paper discusses the architecture and principles of these models, analyzing their advantages and disadvantages in various tasks. Moreover, the study identifies current challenges and proposes future directions for enhancing the performance of HAR systems. Key areas for future research include integrating sensor data with visual information to improve recognition accuracy, reducing computational complexity, and increasing real-time applicability in practical systems. This research provides an in-depth overview of the current status of HAR technology and sets the stage for further advancements and applications, serving as a foundation for future research in the field.

Keywords: Human Action Recognition (HAR), Deep Learning, CNN, GCN.

1. Introduction

Human action recognition (HAR) aims to identify and classify human behaviors automatically from images and videos, an important research direction in computer vision. HAR has wide application value in security and surveillance, human-computer interaction, and health monitoring [1, 2]. For instance, the automatic abnormal behavior recognition technology has been greatly improved in terms of security in recent years [3], which replaces the limited surveillance of human eyes and prevents and stops dangerous behavior events [4]. The interaction between computers and humans has always been a challenging problem. To effectively interact with humans, computers must interpret body movements, such as gestures, and respond accordingly using sensor-based systems. This capability enhances both daily life and industrial processes by enabling more convenient services and advancing virtual reality applications [5]. Furthermore, HAR is crucial in healthcare, particularly in real-time monitoring of patients' health conditions. It is especially valuable for providing personalized assistance to the elderly within smart home environments, making it a prominent research focus in recent years [6]. The HAR field has developed rapidly and has achieved many significant breakthroughs in recent years. However, handling complex dynamic scenes, improving the model's generalization ability, and ensuring real-time performance are still challenges facing this field.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ23495

Therefore, in-depth research on HAR technology has important academic significance and application value.

With the growth of deep learning and its popularity in computer vision, researchers have begun to apply it to the HAR field. Deep learning technology is a revolutionary change that overcomes the shortcomings of manual feature design in traditional methods. It can automatically learn feature representation from large amounts of data [7]. This paper will present previous work in human action recognition and the development of related technologies and methods in the past decade. Furthermore, the paper will also discuss the current research focus and the latest technical methods. The main recent research work can be divided into vision-based and sensor-based methods (more turning to mobile device-based HAR) [8]. There are many models for vision-based HAR, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM), Graph Convolutional Network (GCN), and Transformer. Simonyan and Zisserman utilized two CNNs that process optical flow information and static Red, Green, and Blue (RGB) images in parallel [9]. LSTM was employed by Donahue et al. to generate a single action label based on RGB features extracted by a 2D CNN. Tran et al. first used the 3D CNN architecture for video classification [10]. The 3D convolution kernel can simultaneously process spatial and temporal information, thereby achieving better motion features and context information [11]. Since Vaswani et al. proposed that the Transformer model relies on the self-attention mechanism, more and more work has begun on the Transformer [12]. For instance, to address the computational complexity and memory cost issues, Fan et al. transformed 3D video frames into 2D super images as the input of the Transformer network, which became the Super Image Action Recognition (SIFAR) network [13]. Sensor-based HAR uses sensor nodes (worn by users) such as accelerometers and gyroscopes to capture and process sensor data for activity recognition. In recent years, this technology has adopted devices such as Radio Frequency Identification (RFID) systems and Wireless Fidelity (WiFi) devices to sense human activities through the radio signals they generate [8]. To enhance portability and reliability, it has now developed into mobile devices.

This paper aims to sort out the relevant concepts, background, and latest progress of HAR. It analyzes HAR's core technologies, including sensor-based and vision-based methods in deep learning. In addition, the paper deeply explores its principles and advantages and discusses the advantages and disadvantages of key technologies and future development prospects. This paper puts forward the shortcomings of current methods and looks forward to future research directions.

2. Methodology

2.1. Dataset description

The current popular datasets in the HAR field can also be divided into vision-based and sensor-based. The field of vision technology is now highly mature in HAR. The dataset contains 400 human action categories, each with at least 400 video clips. The categories cover various interactive scenes such as playing musical instruments, shaking hands, hugging, and so on [14]. It is the benchmark dataset for training large-scale video understanding models like SlowFast and Temporal Segment Networks (TSN). Compared with visual 2D, 3D Sensor-based data is the primary method, which is technically complex but highly accurate and more suitable for professional situations. Sensor-based human motion data collection mainly requires wearable devices or motion capture in specific environments. The Human3.6M dataset is a popular 3D human pose data from Carnegie Mellon University, collected by 480 VGA cameras, 30+ High Definition (HD) cameras, and 10 Kinect sensors [15]. This dataset comprises 3.6 million three-dimensional human poses associated with 17 distinct action scenarios, including discussion, eating, exercise, and greeting. Currently, it is extensively utilized in 3D pose estimation and motion modeling.

DOI: 10.54254/2755-2721/2025.TJ23495

2.2. Proposed approach

The primary objective of this paper is to review the recent advancements in the field of HAR, analyze key technologies, evaluate their performance, and explore potential future directions. The paper will address the following tasks: First, it will introduce the concepts, background, and application scenarios of HAR. Second, it will summarize the significant work in vision-based and sensor-based HAR over the past decade. Third, it will analyze and discuss mainstream datasets and methods used in HAR, including CNN, RNN, GCN, and Transformer-based approaches. Finally, it will compare the performance of these different methods and discuss their future development prospects. The pipeline of the entire article is shown in Figure 1.

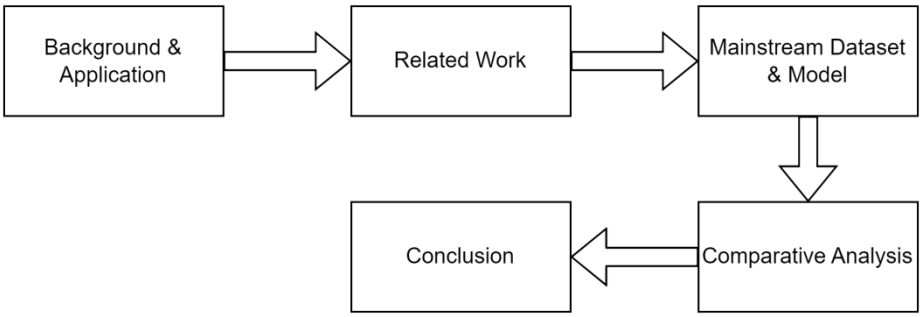

Figure 1: The pipeline of the study (picture credit: original)

2.2.1. CNN-based HAR

CNN is one of the most famous and representative deep learning models in the field of HAR. It can effectively capture the spatial hierarchical structure information in the data because it automatically learns features from raw data through convolution operations.

A CNN consists of three key components: convolutional, pooling, and fully connected layers. The processing flow and structural representation are illustrated in Figure 2, and the following is a detailed explanation. The input image first enters the convolution layer. Then the convolution kernel will slide on different data positions and perform a weighted summation in each convolution kernel size area. Define a^l is the feature map after l convolutional layers, W^l is the convolution kernel of the l-th layer, b means the bias vector, and σ is the activation function, which is usually Rectified Linear Unit (ReLU). The output of the convolutional layer of the lth layer is:

$$a^l = \sigma(a^{l-1} * W^l + b^l) \tag{1}$$

The next component is the pooling layer, whose primary role is to downsample the feature map produced by the convolution operation. This downsampling effectively decreases the data dimensionality and simplifies subsequent computations. In practice, maximum and average pooling are the most frequently employed techniques. They select the highest or the mean value within a specified region. The process can be shown as:

$$a^l = pool(a^{l-1}) \tag{2}$$

Thirdly, the whole connection layer. Following a series of convolution and pooling operations, the output from the preceding layer is converted into a one-dimensional vector. This vector is then subjected to a weighted transformation using a weight matrix, combined with a bias term, and processed through an activation function to generate the final classification result. The final output can be expressed as:

DOI: 10.54254/2755-2721/2025.TJ23495

$$a^L = soft\,max(\,a^{L-1} * W^L + b^L) \tag{3}$$

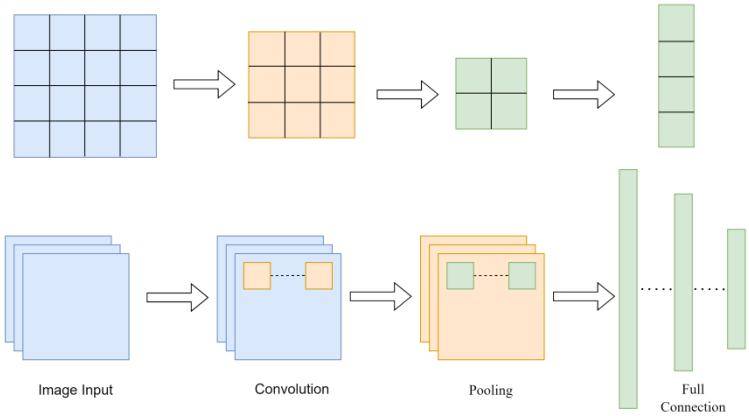

Figure 2: CNN structure (picture credit: original)

2.2.2. RNN-based HAR

RNN is very effective in processing sequence data. RNN can capture the time series information of input data and is suitable for processing data with strong time correlation. The HAR field mainly processes sensor data and video frame sequences.

As shown in Figure 3, RNN is composed of multiple recurrent units, where each unit updates its hidden state at time step t by utilizing both the current input and the secret state from the preceding time step. Given an input sequence $X = \{x_1, x_2, \ldots, x_n\}$, at the t-th time step, RNN first updates the hidden state:

$$h_t = \sigma(W_x * x_t + W_h * h_{t-1} + b) \tag{4}$$

The current input is denoted as x_t, while the previous hidden state is represented by h_{t-1}. The trainable parameters, namely W_x, W_h, b, are employed in the network, and the activation function σ (such as tanh or ReLU) is applied to introduce nonlinearity. Then calculate the output result through:

$$y_t = soft\,max(\,W_y * h_t + c) \tag{5}$$

where c is the trainable parameter, the SoftMax function converts the output into a probability distribution.

When processing long sequences, standard RNN often suffers from issues like gradient vanishing or explosion; to address these challenges, researchers have developed enhanced variants, most notably, the LSTM model, which mitigates these problems by incorporating specialized gating mechanisms.

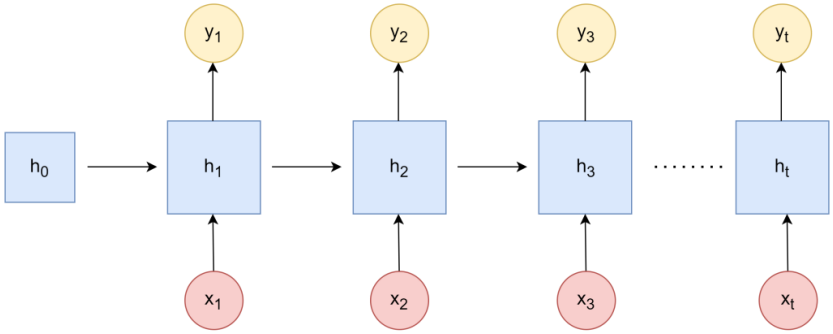

Figure 3: RNN structure (picture credit: original)

DOI: 10.54254/2755-2721/2025.TJ23495

2.2.3. GCN-based HAR

GCN specifically captures the spatial topological relationship and structural information between nodes and is suitable for neural network models that process graph-structured data. GCN learns high-level feature representations of nodes through multiple layers of iteration and understands complex graph structure information. Each layer updates the features of the nodes.

As shown in Figure 4, in GCN, each layer updates the features of the node and finally obtains the output result (label) after weighted summation, linear transformation, and nonlinear activation. The graph's adjacency matrix is A, the node feature matrix is X, the weight matrix is W, and the activation function is σ. An expression representing the computations executed in the lth layer is given by

$$h^l = \sigma(A * h^{l-1} * W^l) \tag{6}$$

GCN is suitable for skeleton-based HAR. The human skeleton can be regarded as a graph structure. Because it is insensitive to changes in lighting, background, and perspective, it can effectively capture the spatial structural information between human joints in complex environments.

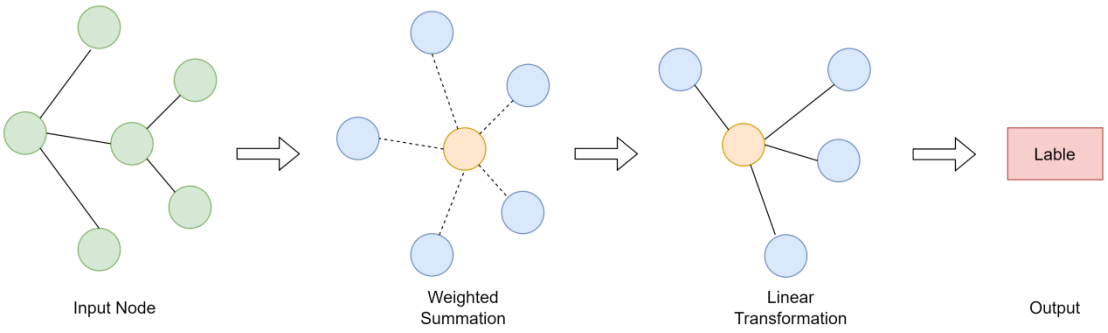

Figure 4: GCN structure (picture credit: original)

2.2.4. Transformer-based HAR

In recent years, the Transformer model has emerged as a powerful tool in the HAR domain, demonstrating substantial advancements. Unlike traditional RNN and CNN, the Transformer leverages a self-attention mechanism that enables it to capture global dependencies throughout the entire sequence.

Transformer itself cannot process sequence order. First, it is necessary to introduce position encoding to integrate position information into the input sequence and retain the order information. Then map the input to the Query(Q), Key(K), and Value(V) vectors, and then calculate the dot product of the query and key, and divide the dot product result by the square root of the dimension, that is. Finally, the SoftMax function is applied to it to get the attention weight. Therefore, use the attention weight to weight the value vector and get the

$$output \quad Attention(Q,K,V) = Soft\,max(\frac{QK^T}{\sqrt{d^k}})V \tag{7}$$

At the same time, Transformer can calculate multiple self-attention mechanisms in parallel to enhance the model's expressiveness. Finally, the feedforward network performs a nonlinear transformation on the output of self-attention. The whole process is shown in Figure 5.

Transformer captures global dependencies and can improve the ability to recognize complex actions. Also, the architecture supports parallel computing, which enhances the training and reasoning efficiency of the model.

DOI: 10.54254/2755-2721/2025.TJ23495

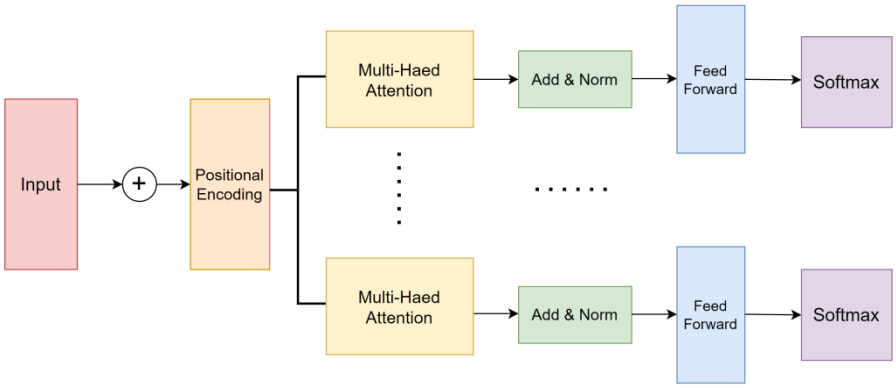

Figure 5: Transformer structure (picture credit: original)

3. Discussion

Owing to its inherent capability to extract feature representations from raw data via convolutional operations automatically, CNN has established itself as one of the quintessential models in deep learning. The accuracy of traditional CNN methods is low. Now, most studies convert bone sequences into pseudo images and use 3D CNN to extract spatiotemporal features. On fixed-structure bone data, 3D CNN can usually achieve high recognition accuracy. However, converting bone sequences into pseudo images may lead to losing the original skeleton's topological relationship information. The need for manual expertise in configuring the input structure and preprocessing pipeline can restrict the model's generalization capabilities.

RNN excel at modeling time series data by capturing long-term dependencies and are particularly well-suited for processing continuous sequences of actions. RNN takes each time step's sensor data or video frame features as input, and gradually updates the hidden state through a recursive structure. The emergence of LSTM alleviates the problem of RNN gradient disappearance or explosion. The LSTM model has significantly improved through the optimized attention mechanism in some sensor-based HAR tasks. However, sequence lengths that are too long will also face challenges in fitting and computing performance.

GCN is suitable for directly processing bone data and regards the human skeleton as a graph structure. Compared with other models, it can make full use of the topological structure of the skeleton and avoid information loss during pseudo-image conversion. The model simultaneously captures spatial topological relationships and temporal dynamics through graph convolution operations. At present, the ST-GCN model can usually show high robustness and recognition accuracy in complex environments. However, the structural design of the GCN model is relatively complex, with more parameters, and has high requirements for the accuracy of bone detection and posture estimation.

Transformer is currently the most popular model in the HAR field. Long-range dependencies are effectively captured by the self-attention mechanism, which remains unaffected by the length of the input sequence. Compared with RNN, Transformer can process all inputs in parallel, significantly improving training and reasoning efficiency. However, Transformer usually requires large-scale data sets for training to fully utilize the advantages of Transformer. In addition, it consumes a lot of computing resources: the model has many parameters and high computational complexity.

Multimodal fusion is an essential trend in the development of HAR. For example, combining sensor data with visual data (such as RGB video and skeleton data) or model fusion can complement each other's shortcomings, thereby improving recognition accuracy and the robustness of the model in complex scenarios. Secondly, given the problem that some current complex models (such as Transformer and deep GCN) have high computational overhead, in the future, technologies such as

DOI: 10.54254/2755-2721/2025.TJ23495

model compression, quantization and pruning can be used to achieve model lightweight and real-time performance improvement, making it more suitable for edge computing and real-time applications. In the future, HAR is not yet mature in human-computer interaction, such as personalized health management and telemedicine, but it also has broad development potential. How to improve real-time performance and accuracy is the direction of future research.

4. Conclusion

This research provides an overview of HAR, highlighting its crucial applications in security surveillance, human-computer interaction, and health monitoring. Additionally, it reviews the advancements of deep learning in HAR over the past decade. It systematically analyzes the architectures and underlying mechanisms of essential models, such as CNN, RNN, GCN, and Transformer. The paper also provides a detailed analysis of the advantages and disadvantages of these models, followed by a discussion on future research directions. Future work will focus on integrating sensor data with visual data to enhance recognition performance further, reduce computational complexity, and improve the real-time applicability of HAR systems.

References

[1] *Turaga, P., Chellappa, R., Subrahmanian, V.S., et al. (2008). Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video technology, 18(11), 1473-1488.*
[2] *Wang, Jindong, et al. (2019). Deep learning for sensor-based activity recognition: A survey. Pattern recognition letters, 119, 3-11.*
[3] *Mahdi, M.S., Mohammed, A.J., Jafer, M.M. (2021). Unusual activity detection in surveillance video scene. Journal of Al-Qadisiyah for computer science and mathematics, 13(3), 92–98.*
[4] *Li, T., Liu, J., Zhang, W., et al. (2021). Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 16266-16275.*
[5] *Girdhar, P., Johri, P., Virmani, D. (2021). Vision based human activity recognition: a comprehensive review of methods & techniques. Turkish Journal of Computer and Mathematics Education, 12(10), 7383-7394.*
[6] *Karim, M., Khalid, S., Aleryani, A., et al. (2024). Human action recognition systems: A review of the trends and state-of-the-art. IEEE Access.*
[7] *Jiang, Z., Liu, Y. (2024). A Review of Human Action Recognition Based On Deep Learning. International Conference on Intelligent Informatics and Biomedical Sciences, 9, 78-83.*
[8] *Yin, Y., Xie, L., Jiang, Z., et al. (2024). A systematic review of human activity recognition based on mobile devices: overview, progress and trends. IEEE Communications Surveys & Tutorials, 26(2), 890-929.*
[9] *Wang, Z., Lu, H., Jin, J., et al. (2022). Human action recognition based on improved two-stream convolution network. Applied Sciences, 12(12), 5784.*
[10] *Li, Z. (2024). A Study on Video Classification Based on Deep Learning. International Conference on Interactive Intelligent Systems and Techniques, 81-89.*
[11] *Tran, D., Bourdev, L., Fergus, R., et al. (2015). Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE international conference on computer vision, 4489-4497.*
[12] *Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.*
[13] *Fan, Q., Panda, R. (2021). Can an image classifier suffice for action recognition? arXiv preprint, 14104.*
[14] *Kay, W., Carreira, J., Simonyan, K., et al. (2017). The kinetics human action video dataset. arXiv preprint, 6950.*
[15] *Mehta, D., Rhodin, H., Casas, D., et al. (2017). Monocular 3d human pose estimation in the wild using improved cnn supervision. International conference on 3D vision, 506-516.*