Deep Learning Approaches for Human Activity Recognition

Eman Hato[1], Amal Sufiuh Ajrash[2] and Muthana Salih Mahdi[1]
[1]*Department of Computer Science, College of Science, Mustansiriyah University, 10011 Baghdad, Iraq*
[2]*Department of Computer Science, College of Science for Women, University of Baghdad, 10011 Baghdad, Iraq*
emanhato@uomustansiriyah.edu.iq, amalsa_comp@csw.uobaghdad.edu.iq, muthanasalih@uomustansiriyah.edu.iq

Keywords: Confusion Matrix, Human Action Recognition, Human Activity Recognition, Recurrent Neural Networks, Graph Neural Networks.

Abstract: Recent advancements in computer vision have led to increased interest in complex video analysis tasks, particularly in understanding and predicting content. Recognizing human actions and activities is crucial in everyday life, as it allows for the collection of comprehensive and accurate information about human behavior through wearable or stationary devices. Numerous machine learning and deep learning techniques have been investigated in human activity recognition in order to categorize human activities. This paper presents the impact of deep learning architectures for human action and activity recognition. It outlines the structure, performance, and limitations of the deep learning techniques used. These findings are intended to resolve limitations found in previous research, paving the way for enhanced models with greater accuracy across handling varied datasets. The evaluation analysis showed that action recognition achieved an average accuracy of 89.2%, while activity recognition achieved an average accuracy of 93.9%. Furthermore, the combination of ResNets and ConvLSTM algorithms on the SKIG dataset outperformed other methods in accuracy, reaching 99.72%. Discovering and extracting semantic features and the complexity of activities are key challenges facing human activity recognition systems.

1 INTRODUCTION

Today's Human Action/Activity Recognition (HAR) plays an important role in human interaction and interpersonal relationships, providing information about a person's identity, personality, and psychological state. HAR is a critical field in artificial intelligence that involves detecting, classifying, and predicting human activities [1], [2]. Many applications, including video surveillance systems, healthcare, sports, and event analysis, require a HAR system. For example, the HAR systems have impressively improved human safety by helping monitor patients' health and prevent damage to all types of wearable devices [3]. The fundamental goal of HAR systems is to predict human movements based on action data collected from various sensors and acquisition devices. In healthcare applications, HAR systems have impressively improved patient safety by enabling continuous monitoring of individuals' health conditions through both wearable and stationary devices. The Internet of Things (IoT) has proven highly effective in this context, facilitating automated control and security while enabling the recognition of both normal and abnormal human movements. In the transportation sector, HAR technologies enhance safe driving management through real-time analysis of driver behavior. Furthermore, HAR represents a significant advancement in social research, allowing researchers to detect behaviors related to violence and stress in social interactions [4], [5]. Many of these challenges are being addressed through the use of advanced machine learning techniques, leading to significant improvements in activity recognition. As illustrated in Figure 1, machine learning-based HAR consists of three stages: data collection, the machine learning-based HAR process itself, and framework support. Due to the superior performance of deep learning algorithms in this field, traditional methods have seen a significant decrease in usage [6].

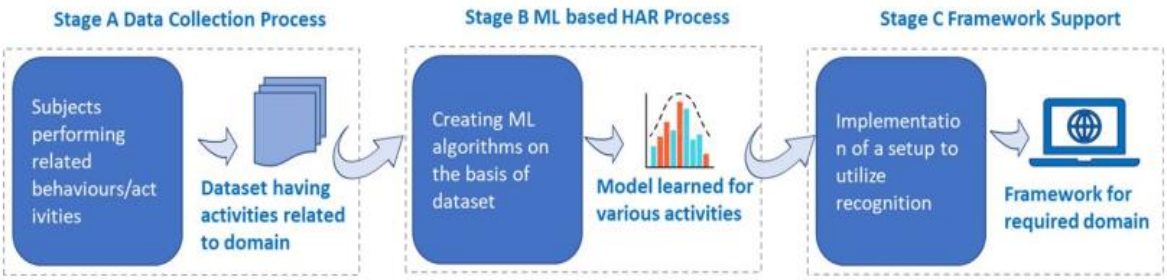

Figure 1: Machine learning processing for HAR.

2 DEEP LEARNING TECHNIQUES

The shortcomings of conventional techniques are addressed by deep learning, which does not require manual feature engineering. It is capable of extracting high-level representations through its deep layers, making it well-suited for complex and dynamic situations [7]. The most commonly used deep learning models are:

2.1 Convolution Neural Networks (CNNs)

Among the most often used deep learning architectures are CNNs. Applications involving image processing frequently employ this method. The convolutional layer, pooling layer, and fully connected layer are the three different types of layers found in CNNs. Each CNN has two training phases, which include the feed-forward stage and the backpropagation stage. GoogLeNet, VGGNet, ResNet, and AlexNet are typical CNN architecture examples. The scientific literature covers a variety of application fields, including computational mechanics, energy, remote sensing, and electronics systems, even though CNNs are mainly utilized for image processing applications [8].

2.2 Recurrent Neural Networks (RNN)

The purpose of RNNs is to identify patterns in speech, handwriting, text, and other applications. Through repeated computations, RNNs' architecture makes use of cyclic connections, which enable them to process input data sequentially. RNNs are really simply regular neural networks that have been modified to include edges that feed data into the subsequent time step rather than just the subsequent layer at the same time step itself. The hidden units store all of the prior input data in a state vector, which is then utilized to calculate the outputs. RNNs represent a relatively newer approach in deep learning. Current applications mentioned in the literature include hydrological modeling, energy forecasting, expert systems, navigation, and economics [9].

2.3 Deep Belief Networks (DBNs)

It is used to learn data using high-dimensional manifolds. DBNs are a multi-layer hybrid neural network that contains directed and undirected connections. There are connections between layers and excluding connections between units within each layer. Deep belief networks contain eagerly trained Restricted Boltzmann Machines (RBMs). In this model, each RBM layer communicates with both the previous and subsequent layers. It consists of a feedforward network along with multiple RBM layers that serve as feature extractors. The hidden layer and the visible layer represent two of these RBM layers. DBNs are among the most reliable methods in deep learning, offering both computational efficiency and high accuracy [10].

2.4 Long Short-Term Memory (LSTM)

The purpose of Long Short-Term Memory networks (LSTMs) is to overcome the limitations of standard RNNs in learning long-range dependencies. LSTMs are really simply a special kind of RNN that has been augmented with a dedicated memory cell and a system of gating units that control the flow of information, rather than just a single, simply connected hidden state. The hidden units are modulated by these gates, which decide what information to store, update, or discard. LSTMs represent a significant advancement in sequence modeling [11].

3 DATASETS USED

Over the years, large datasets for human activity and action recognition have been collected and made available, each of which has been specifically used to study different models and algorithms. Examples of video action and activity datasets are listed in Table 1 [12], [13].

Table 1: Human action and activity datasets.

	Dataset Name	Type
1	Weizmann	Action
2	KTH dataset	Action
3	UT-Interaction dataset	Action
4	BIT-Interaction dataset	Action
5	UCF101 dataset	Action
6	HMDB51 dataset	Action
7	Sports -1M dataset	Action
8	PASCAL VOC 2012 action	Action
9	Stanford40	Action
10	IsoGD,	Action
11	NTU RGB+D120	Action
12	NTU RGB+D	Action
13	Northwestern-UCLA	Action
14	MSR-Action3D	Action
15	UTKinect-Action	Action
16	Florence3D-Action	Action
17	UCI HAR	Activity
18	WISDM	Activity
19	Ordonez	Activity
20	IMU	Activity
21	THUMOS14	Activity
22	ActivityNetv1.2	Activity
23	W-HAR	Activity
24	Kinetics400	Activity
25	Opportunity	Activity
26	ADL	Activity

4 PERFORMANCE METRICS

Various performance metrics were used to evaluate the effectiveness of the proposed classification model. The confusion matrix provides a visual representation of classification performance by showing correctly and incorrectly classified instances for each class, as illustrated in Figure 2 [4].

To assess model performance comprehensively, five widely used evaluation metrics were employed: sensitivity, specificity, precision, recall, and F-score. Sensitivity measures the ability of the model to correctly identify positive instances, while specificity evaluates its capability to correctly recognize negative instances. Precision indicates the reliability of positive predictions, whereas recall reflects the model's effectiveness in detecting all relevant positive cases. The F-score combines precision and recall into a single measure, providing a balanced assessment of classification performance. In addition, overall accuracy was used to measure the proportion of correctly classified samples among all evaluated instances [4], [14], [15].

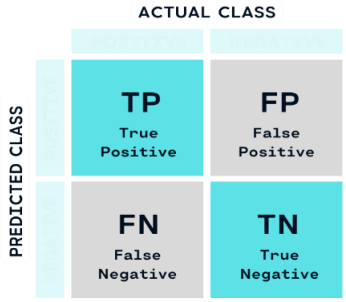

Figure 2: The confusion matrix.

5 LITERATURE REVIEW

Numerous studies have focused on identifying human actions and activities using deep learning techniques. A video-based HAR system aims to automatically detect actions or activities from video information. Various performance metrics were used to determine the effectiveness of the HAR; some of these metrics can be found in. The previous work is presented here, with a brief explanation of each:

A. Elboushaki et al. [16] built a multi-dimensional feature architecture model for human gesture recognition. First, the spatio-temporal features were extracted from video data through 3D Convolutional Residual Networks (ResNets), which were then linked to a ConvLSTM to capture the temporal dependencies between them. Then, the temporal data was encoded into a motion representation based on 2D-ResNets. After that, the spatial and temporal features were integrated for a stable spatiotemporal feature classification as shown in Figure 3. The proposed model was tested on public datasets. The highest classification accuracy was 94.9 based on UCF101 dataset.

S.K. Yadav et al. [17] A method to recognize Yoga asanas using deep learning algorithms was presented. Features were extracted from each frame's keypoints using the CNN layer. Temporal predictions were made using the LSTM. Ten males and five females with a normal RGB were used to test the suggested method on the dataset of six yoga poses

that the authors had developed. 99.04% accuracy on single frames and 99.38% accuracy after polling guesses on 45 video frames were the results.

A method for human activity recognition that incorporates both offline training and online testing was proposed by F. Xiao et al. [18]. Initially, the Archive of Motion Capture as Surface Shapes (AMASS) dataset's features were extracted using a deep convolutional neural network. The parameters were tuned in the fully-connected layers with the Database of Interacting Proteins (DIP) and Inertial Measurement Units (IMU) datasets. In the online testing stage, the DeepConvLSTM architecture was used for classification.

P. Agarwal and M. Alam [19] proposed a model for human activity recognition on edge devices. The proposed model employed a Shallow Recurrent Neural Network (SRNN) combined with a LSTM deep learning algorithm. The WISDM dataset was utilized for training and testing the proposed model. The performance of the proposed model was 95.78 in terms of accuracy and 95.73 in terms of F-score.

For human action recognition, N. Jaouedi et al. [20] introduced a hybrid deep learning approach. The motion tracking was extracted based on the Gaussian Mixture Model (GMM) and the Kalman filter, which was represented as input to the Gated Recurrent Neural Networks (GRNN). The GRNN was used to predict human actions. The GRNN parameters' weights were modified using back propagation over time to minimize error and loss. Figure 4 presents the proposed model.

In their work on human activity recognition, Z. N. Khan and J. Ahmad [21] proposed a model as illustrated in Figure 5, featuring three lightweight CNN heads for feature extraction. They integrated a Squeeze-and-Excitation attention module between two heads to prioritize relevant features by assigning them higher weights. The proposed framework obtained 0.9818 accuracy and 0.9720 F-score based on the WISDM dataset.

N. Rashid et al. [22] presented an adaptive CNN model for human activity recognition to save energy of edge devices. The first step was filtering the data to smooth it. Then, the filtered data was divided using a sliding window of one hundred samples with 70% overlap. After that, each segment was down-sampled to thirty-two samples and the statistical features were extracted. In the proposed design, two convolution blocks each had a corresponding output block. A decision tree, operating on statistical features, was responsible for selecting the appropriate output block during the protection phase. The method's performance according to the F-score was 91.57 based on the Opportunity dataset.

Y. Yang et al. [23] extracted both lower-order and higher-order motion information in order to propose an action recognition method. These characteristics were used to determine the action class using a Graph Convolutional Network (GCN) with channel-wise attention. The suggested strategy was tested using three datasets. On the Northwestern-UCLA dataset, the accuracy was 96.8%.

C. Yang et al. [24] introduced a human action recognition method that combined LSTM with an attention mechanism. To extract key frames from different actions, they employed the K-means clustering algorithm. The LSTM network was optimized using an attention mechanism to improve its accuracy and recognition efficiency for human action sequences. The accuracy achieved was 94.0% according to the dataset acquired with an IMU sensor-based motion capture device.

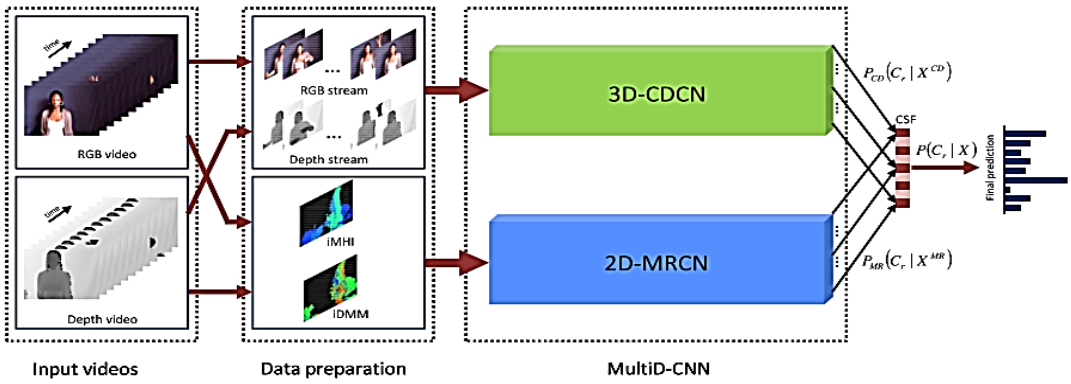

Figure 3: The multi-dimensional feature architecture model.

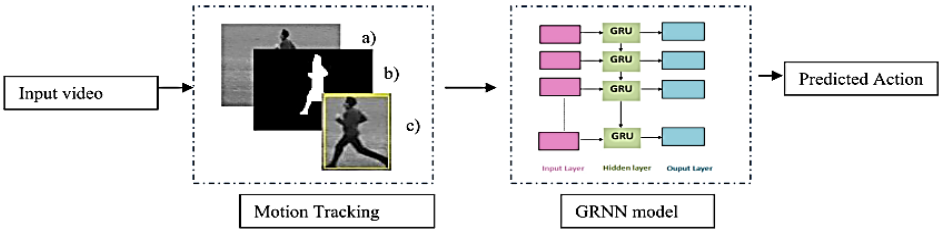

Figure 4: The hybrid deep learning model.

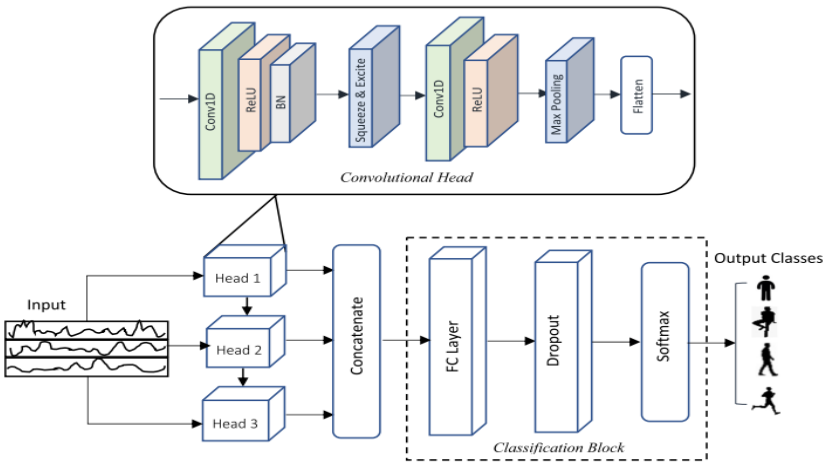

Figure 5: The attention technique and CNN model.

C. Yang et al. [25] and M. Joudaki et al. [26] combined a two-dimensional RBM (2D Conv-RBM) model with an LSTM network for human action recognition. Frames were selected to remove redundancy and retain the most informative frames. The 2D Conv-RBM model extracted spatial features like textures, motion patterns, and edges while minimizing parameters and preserving spatial relationships through weight sharing. In order to effectively identify short- and long-term activity patterns, LSTM processes the features to capture temporal dependencies across frames.

A technique for motion-based human action recognition was presented by E. L. Lozada et al. [27]. Using a pre-trained model on ImageNet for feature extraction, motion estimation through the use of optical flow, person identification and tracking, and action recognition through the use of a two-stream model were the four stages of the suggested approach. RGB data is processed by one stream, and the second stream uses the optical flow's temporal information. The CNN and LSTM models forecasted the actions as seen in Figure 6. The proposed method obtained 94.9% accuracy according to the custom dataset comprises five action classes.

H. Chen et al. [28] proposed a motion recognition method that utilizes keyframe selection based on the human skeleton. The process begins with selecting keyframes through the K-means clustering algorithm. Pose feature vectors are extracted from skeletal joint data for each frame. Turning point frames are then detected based on each body part's momentum flow, after which the vectors are input into a multi-objective optimization model for keyframe selection. Finally, the resulting keyframes are used as input for a Support Vector Machine (SVM) classifier to recognize motion.

R. D. Rani and C.J. Prabhakar [29] proposed recognizing human actions in still images by first using YOLOv8 for human part detection. The detected parts were extracted, split into uniform patches, and each patch was flattened and converted to a vector through linear embedding. Positional embeddings and a class token were added to maintain patch order, and these tokens were subsequently passed to a vision encoder for action classification.

In their work on human action recognition, D. Lamani et al. [30] employed an SVM-directed classifier. They used an improved version of a lightweight directed acyclic graph-based 2D CNN to extract motion features from videos, which were then

passed to the SVM for recognition. Experimental results on the UCF101 dataset demonstrated that this method achieved 92.3% accuracy.

Table 2 provides a brief comparison of the human action and activity recognition methods described in the previous section.

4 RESULTS DISCUSSION AND ANALYSIS

Extensive studies have been conducted on human activity detection, with mixed results. In general, deep neural networks are a popular and effective method for identifying human activity. Detecting human activity in videos using deep learning has demonstrated high accuracy in recognizing human activity. In general, distinguishing between actions is easier and more accurate than distinguishing between different activities, for several reasons. These include: human actions are less complex, and their data is widely available, facilitating the extraction of features that reflect the finest details, which in turn improves the efficiency and accuracy of the recognition. Nevertheless, Previous studies have demonstrated high accuracy in human activity recognition, as reported in Table 3.

One of the primary goals of human activity detection is to classify and distinguish activities based on their type. Previous studies have shown that various deep neural network methods can classify activities with a high degree of accuracy. Although many studies have focused on tracking specific activities such as running, climbing stairs, walking, sleeping, and waking up, their limited scalability restricts their use outside of laboratories or smart homes. By expanding their activity-tracking capabilities, these systems can achieve better performance, intelligence, and responsiveness for the user. Furthermore, the majority of HAR systems can execute HAR in situations involving just one individual. However, many scenarios in the real world call for multi-person HAR, including in living rooms, kitchens, and retail establishments, or when several people are performing atomic actions like embraces and handshakes.

Despite the importance of contextual information, no human activity recognition datasets or systems have been found that take this information into account.

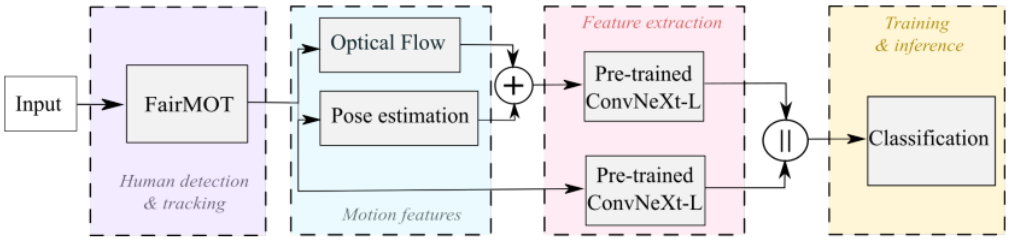

Figure 6: The CNN and LSTM model.

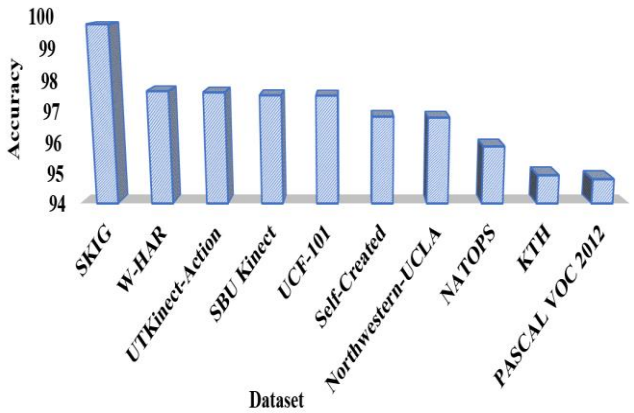

Figure 7: Average accuracy by dataset.

Table 2: Literature review abstraction.

Ref	Type	Technique	Limitation	Domain	Dataset	Acc.
[16]	Action	Multiple CNN	High computational complexity	Analyze videos	HMDB51 UCF101 THUMOS14	71.0 94.9 80.1
[17]	Activity	CNN, LSTM	The low scalability makes such research less usable.	Recognize Yoga asanas	Self-Created	99.38
[18]	Activity	DeepConv, LSTM	High computational	Sensor-based human activity recognition	IMU	91.15
[19]	Activity	RNN, LSTM	Simple daily activities,	Recognize individual's daily activity	WISDM	95.78
[20]	Action	GMM, GRNN	High computational complexity	Human activity analysis	KTH UCF Sport UCF101	96.3 89.01 89.3
[21]	Activity	CNN, Attention Technique	There is a weakness in distinguishing between certain activities.	Human-computer interaction	WISDM UCI HAR	98.1 95.3
[22]	Activity	CNN	Difficulty distinguishing the complex activities.	Health monitoring	Opportunity W-HAR	91.57 97.64
[23]	Action	GCN, Attention Technique	High complexity and costly calculations.	Skeleton-based action recognition	NTU RGB+D120	91 96.9
[24]	Action	LSTM, Attention Technique	Limited action is recognized.	Human motion recognition	Self-Created	94
[25]	Action	2D-CNN, 3D-CNN	Unable to identify activities that occur outside of their specific time context.	Analyze human action	HMDB-51 UCF-101	77.7 97.5
[26]	Action	2D-CNN, LSTM	Several parameters need to be adjusted.	Real-time applications and sports analytics	KTH, UCF Sports HMDB51.	97.3 94.8 81.5
[27]	Action	CNN, LSTM, Optical Flow	A limited number of human actions are identified.	Video analysis	Self-Created	94.9
[28]	Action	K-means, SVM	Action recognition depends on the accuracy of the multi-objective optimization model	Human behavior understanding	MSR-Action3D, UTKinect-Action, Florence3D-Action	94.6 97.6 94.2
[29]	Action	YOLOv8	The model sometimes has difficulty managing occlusions or overlapping objects in intricate scenes.	Detecting human actions in still images	Stanford40 PASCAL VOC 2012 action	97.4 94.8
[30]	Action	2D CNN, SVM	High accuracy requires effective extraction of spatial.	Surveillance video analysis	UCF101 HMDB51 KTH	92.3 84.1 98.92

Table 3: Action vs activity recognition performance.

Type	Average Accuracy	Min Accuracy	Max Accuracy
Action	89.2 %	70.1 %	99.72 %
Activity	93.9 %	78.3 %	100%

Table 4: Performance of the deep learning technique.

Technique	Best Dataset	Accuracy
ResNets+ConvLSTM	SKIG	99.72
GCN +Attention	NTU RGB+D	96.9
2D-CNN + 3D-CNN	UCF-101	97.5
CNN + LSTM	Self-Created	94.9
YOLOv8	Stanford40	97.4

It is important to note that there are many publicly available datasets pertaining to the detection of human activities utilizing different sensor techniques. These datasets are not standardized, though, so not all of them offer the same degree of information regarding test subjects, test settings, and suitable data quantities. According to Table 2, the UCF101 and HMDB51 datasets are the most commonly used in the field of action recognition. However, the SKIG dataset achieved the highest accuracy (99.72%). Figure 7 illustrates the accuracy levels achieved by the different datasets.

The difference between Elboushaki et al. [16] and Yosry et al. [25] in reported accuracy (94.9% vs. 77.7% on HMDB-51) is primarily due to a combination of factors: dataset difficulty (UCF-101 is generally considered less challenging than HMDB-51) and model complexity (Elboushaki's multi-dimensional feature architecture vs. Yosry's fusion methods).

The most popular techniques are the combination of CNNs and LSTMs, which is widely used. Additionally, attention mechanisms are gaining popularity in recent research across various base architectures. Table 4 illustrates the accuracy distribution based on technique type and the databases used. Numerous studies have been carried out in the security domain to detect and stop criminal activity and suspicious conduct in public areas. There is increasing hope that more precise techniques for detecting human activity may become available in the future due to recent developments in deep learning and the utilization of large databases.

5 CONCLUSIONS

The main objective of motion or activity identification algorithms is similar; they differ in their ability to identify specific movements and employ different methods for data collection. Multiple actions at once, like daily routines, are the main empha-sis of motion recognition. Motion recognition, on the other hand, concentrates on particular, predetermined human movements. To get good and acceptable results in this field, deep learning algorithms have produced several efficient ways for recognizing a variety of human actions. Despite extensive research efforts, the performance of these systems remains poor. The detection and extraction of semantic features, knowledge of input data, and the complexity of activities are major challenges facing human activity recognition systems. Some observations drawn from the test results presented in the review papers presented in this survey are:

1) Basic activities like running, walking and sitting are relatively easy to recognize, while complex activities like performing exercises that involve multiple routines are much more difficult to recognize;
2) Current human activity recognition techniques struggle with HAR in complex, multi-person real-world scenarios, despite their superior performance in atomic and basic activities in single-person scenarios;
3) Most HAR systems also operate based on a single person being engaged in a single activity at any given time. However, in real-life scenarios, individuals often perform multiple tasks and engage in simultaneous activities, such as walking while drinking coffee, snacking, or drinking while reading.

Future work should focus on developing systems that can model human postures and interactions with surrounding objects and scenes to gain a more comprehensive and accurate understanding of human activity.

REFERENCES

[1] R. Gómez-Ramos, J. Duque-Domingo, E. Zalama and J. Gómez-García-Bermejo, “An unsupervised method to recognise human activity at home using non-intrusive sensors,” Electronics, vol. 12, no. 23, pp. 1-24, Nov. 2023, [Online]. Available: https://doi.org/10.3390/electronics12234772.

[2] F. Demrozi, G. Pravadelli, A. Bihorac and P. Rashidi, “Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey,” IEEE Access, vol. 8, pp. 210816-210841, 2020.

[3] N. Sedaghati, S. Ardebili and A. Ghaffari, "Application of human activity/action recognition: a review," Multimed. Tools Appl., Jun. 2025, [Online]. Available: https://doi.org/10.1007/s11042-024-20576-2.

[4] D. R. Beddiar, B. Nini, M. Sabokrou and A. Hadid, "Vision-based human activity recognition: A survey," Multimed. Tools Appl., vol. 79, pp. 30509-30555, 2020, [Online]. Available: https://doi.org/10.1007/s11042-020-09004-3.

[5] H. Khalaf and M. Riyadh, "Human activity recognition using inertial sensors in a smartphone: technical background (review)," Al-Nahrain J. Sci., vol. 27, no. 1, pp. 108-120, Mar. 2024, [Online]. Available: https://doi.org/10.22401/ANJS.27.1.10.

[6] N. S. Khan and M. S. Ghani, "A survey of deep learning based models for human activity recognition," Wirel. Pers. Commun., vol. 121, pp. 3099-3133, 2021, [Online]. Available: https://doi.org/10.1007/s11277-021-08525-w.

[7] H. Yin, R. O. Sinnott and G. T. Jayaputera, "A survey of video-based human action recognition in team sports," Artif. Intell. Rev., vol. 57, pp. 293-348, Sep. 2024, [Online]. Available: https://doi.org/10.1007/s10462-024-10934-9.

[8] V. Sharma, M. Gupta, A. Kumar and D. Mishra, "Video processing using deep learning techniques: a systematic literature review," IEEE Access, vol. 9, pp. 139489-139504, Oct. 2021, [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3118541.

[9] F. Shafizadegan, A. R. Naghsh-Nilchi and E. Shabaninia, "Multimodal vision-based human action recognition using deep learning: a review," Artif. Intell. Rev., vol. 57, Art. no. 178, Jun. 2024, [Online]. Available: https://doi.org/10.1007/s10462-024-10730-5.

[10] J. Wang, Y. Chen, S. Hao, X. Peng and L. Hu, "Deep learning for sensor-based activity recognition: a survey," Pattern Recognit. Lett., vol. 119, pp. 3-11, Mar. 2019, doi: 10.1016/j.patrec.2018.02.010.

[11] X. Chai, B. G. Lee, C. Hu, M. Pike, D. Chieng, R. Wu and W.-Y. Chung, "IoT-FAR: a multi-sensor fusion approach for IoT-based firefighting activity recognition," Inf. Fusion, vol. 113, p. 102650, Aug. 2025, [Online]. Available: https://doi.org/10.1016/j.inffus.2024.102650.

[12] T. F. Naik Bukht, H. Rahman, M. Shaheen, A. Algarni, N. A. Almujally and A. Jalal, "A review of video-based human activity recognition: theory, methods and applications," Multimed. Tools Appl., Oct. 2024, [Online]. Available: https://doi.org/10.1007/s11042-024-19711-w.

[13] X. Cheng, L. Zhang, Y. Tang, Y. Liu, H. Wu and J. He, "Real-time human activity recognition using conditionally parametrized convolutions on mobile and wearable devices," IEEE Sens. J., vol. 21, no. 15, pp. 16661-16673, Aug. 2021.

[14] E. Hato, Z. S. Abduljabbar and Z. J. Ahmed, "Comparative analysis for bag of features (BoF) performance," Iraqi J. Sci., vol. 65, no. 8, pp. 4606-4622, Aug. 2024, [Online]. Available: https://doi.org/10.24996/ijs.2024.65.8.38.

[15] S. M. Al-Selwi, M. F. Hassan, S. J. Abdulkadir, A. Muneer, E. H. Sumiea, A. Alqushaibi and M. G. Ragab, "RNN-LSTM: from applications to modeling techniques and beyond-systematic review," J. King Saud Univ. Comput. Inf. Sci., vol. 36, p. 102068, 2024, [Online]. Available: https://doi.org/10.1016/j.jksuci.2024.102068.

[16] Elboushaki, R. Hannane, K. Afdel and L. Koutti, "MultiD-CNN: a multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences," Expert Syst. Appl., vol. 139, p. 112829, Mar. 2020.

[17] S. K. Yadav, A. Singh, A. Gupta and J. L. Raheja, "Real-time yoga recognition using deep learning," Neural Comput. Appl., vol. 33, no. 16, pp. 13243-13256, 2021.

[18] F. Xiao, L. Pei, L. Chu, D. Zou, W. Yu, Y. Zhu and T. Li, "A deep learning method for complex human activity recognition using virtual wearable sensors," in Proc. Int. Conf. Hum.-Centric Comput. (HCC), 2020, pp. 455-467.

[19] P. Agarwal and M. Alam, "A lightweight deep learning model for human activity recognition on edge devices," Procedia Comput. Sci., vol. 167, pp. 2364-2373, 2020.

[20] N. Jaouedi, N. Boujnah and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," J. King Saud Univ. Comput. Inf. Sci., vol. 32, no. 4, pp. 447-453, 2020.

[21] Z. N. Khan and J. Ahmad, "Attention-induced multi-head convolutional neural network for human activity recognition," Appl. Soft Comput., vol. 110, p. 107671, Oct. 2021.

[22] N. Rashid, B. U. Demirel and M. A. Al Faruque, "AHAR: adaptive CNN for energy-efficient human activity recognition in low-power edge devices," IEEE Internet Things J., vol. 9, no. 15, pp. 13000-13012, Aug. 2022.

[23] Y. Yang, H. Chen, Z. Liu, Y. Ly, B. Zhang, S. Wu, Z. Wang and K. Ren, "Action recognition with multi-stream motion modeling and mutual information maximization," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2023, pp. 1658-1664.

[24] C. Yang, F. Mei, T. Zang, J. Tu, N. Jiang and L. Liu, "Human action recognition using key-frame attention-based LSTM networks," Electronics, vol. 12, no. 12, p. 2622, Jun. 2023.

[25] S. Yosry, L. Elrefaei, R. ElKamaar and R. R. Ziedan, "Various frameworks for integrating image and video streams for spatiotemporal information learning employing 2D-3D residual networks for human action recognition," Discov. Appl. Sci., vol. 6, no. 3, p. 141, Mar. 2024.

[26] M. Joudaki, M. Imani and H. R. Arabnia, "A new efficient hybrid technique for human action recognition using 2D Conv-RBM and LSTM with optimized frame selection," Technologies, vol. 13, no. 2, p. 53, Feb. 2025.

[27] E. López-Lozada, H. Sossa, E. Rubio-Espino and J. Y. Montiel-Pérez, "Action recognition in videos through a transfer-learning-based technique," Mathematics, vol. 12, no. 20, p. 3245, Oct. 2024.

[28] H. Chen, Y. Pan and C. Wang, “An optimization method of human skeleton keyframes selection for action recognition,” Complex Intell. Syst., vol. 10, no. 3, pp. 4659-4673, Jun. 2024.

[29] D. R. Rani and J. P. C., “Vision transformer-based model for human action recognition in still images,” J. Comput. Anal. Appl., vol. 33, no. 8, pp. 522-531, 2024.

[30] D. Lamani, P. Kumar, A. Bhagyalakshmi, J. M. Shanthi, L. P. Maguluri, M. Arif, C. Dhanamjayulu, S. Kumar and B. Khan, “SVM directed machine learning classifier for human action recognition network,” Sci. Rep., vol. 15, no. 1, p. 672, Jan. 2025.