An Adaptive Hybrid Deep Learning Approach for Human Action Recognition

Shreyas Pagare[1*],
Rakesh Kumar[2], Sanjeev Kumar Gupta[3]
[1,2,3] *Rabindranath Tagore University Bhopal (M.P), India*
[1*] *shreyas_au211443@aisectuniversity.ac.in*
[2]*rakeshmittan@gmail.com*
[3]drskg1973@gmail.com

Abstract: Human Activity Recognition (HAR) technology, which is focused on identifying and analyzing human activities, has gained significant interest in recent years. Traditional approaches have employed manually designed features to identify human activities, leading to limited feature extraction. Neural network detectors are increasingly used in personal and portable devices to detect and recognize human actions. Nevertheless, unimodal methods rely on a solitary sensing modality and employ machine learning techniques to identify human activities. A deep learning-based Human Activity Recognition (HAR) model called Adaptive Hybrid Deep Attentive Network (AHDAN) will be created to address these abstract concepts. This model will combine a 3D Convolutional Neural Network (1DCNN) with gated Recurrent Units (GRU) to enhance the recognition process. Additionally, the parameters of the network will be optimized to improve the recognition process further. Through comprehensive experimental assessments on the UCF101 benchmark dataset, we have established that our proposed method surpasses existing state-of-the-art techniques in action recognition. These findings underscore the capability of our approach to enhance future research in video action recognition. This study presents a novel method for identifying actions in video content. The technique combines attention-based mechanisms with a long short-term memory network and an improved, optimized 3D Convolutional Neural Network to achieve effective action recognition.

Keywords: Attention Mechanism, Long Short-Term Memory (LSTM), 3D Convolutional Neural Network (CNN), and Human Action Recognition (HAR).

1 Introduction

Human activity recognition (HAR) constitutes a critical component in contemporary society due to its capacity to derive novel insights regarding human actions from unprocessed data. Given the increasing prevalence of interpersonal communication applications, the study of HAR innovation has emerged as a significant field of research both domestically and internationally. People can classify various forms of urban transportation and acquire the necessary knowledge for efficient functioning by extracting information from common objects, thus establishing the basis for future applications.

Activity recognition (HAR) is a crucial component of modern society as it can extract

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_26

new insights about human actions from unprocessed data. Due to the increasing popularity of interpersonal communication applications, the study of HAR innovation has become a significant field of research both domestically and globally. People can classify various forms of urban transportation and acquire the necessary knowledge for efficient functioning by extracting information from common objects, thus establishing the groundwork for future applications. [1]

When it comes to personal conversations and interactions, watching people is key [2]. Information regarding a particular person, their attitude, and mental factors makes it difficult to retrieve. The capacity to understand the actions of other people is one of the most important topics studied in artificial intelligence research.[2], [3].

Several consumer goods relied on activity recognition as a core component. To drastically alter the gaming experience, game systems such as Gesture recognition and full-body tracking technologies are essential components of gaming systems like the Nintendo Wii and Microsoft Kinect. [4]. These systems were originally developed for the entertainment industry, but they have since found new uses in areas such as rehabilitation and personal fitness training. They have also been major drivers in activity recognition research.[5] Lastly, there are athletic products that include motion sensors and offer feedback to athletes at all levels, like the Philips Direct Life and Nike+ running shoes [6].

These examples show how important it is for businesses and universities to acknowledge human actions. Despite substantial advancements in activity recognition system prototyping and deployment and inferring activities from on-body encoders, developing HAR systems that meet application and user needs is still a challenging issue. This holds regardless of whether HAR methods that were effective for a particular recognition task are transferred to an entirely new area of study.

Machine Learning

To glean useful insights from respondents and use them for internal restructuring, a Machine Learning algorithm compiles a set of instructions along with empirical methods. A Machine Learning model revolves around this central theme. In machine learning, a model is fundamental. For this purpose, a Machine Learning Algorithm is utilized[2]. In order to get the right answer, a method takes into account all the assumptions that a model can make with the data you give it. The first step in a Machine Learning process is to feed the machine a mountain of data. From there, the algorithm is trained to find patterns and make predictions based on this mountain of data. The next step is to use these observations to create training data that solves the problems methodically[5]. Reproducing the typical look of papers in the area is the target. As strictly as possible, we ask that you adhere to these rules.

Deep Learning

The subfield of machine learning known as "Deep Learning" attempts to model reality according to a hierarchical structure of recursive principles, wherein each dimension is linked to smaller units, and numerous more abstract representations are derived from less abstract ones. It learned to mirror reality as a hierarchical recursive system of ideas, which allowed it to attain outstanding performance and scalability[7]. Millions of images, some of which show actual people or animals. In essence, this type of neuron changes based on the cells in the images it receives. Data sets that resemble specific human characteristics, such as a hand, head, and ears, can be identified and labeled by it, providing visual confirmation of a person's presence[8].

2 Literature Review

Human activity recognition (HAR) is a technique that can identify motions or gestures made by the body and use them to define or predict the steps of an action or behavior [9]. Academic and industrial institutions are interested in conducting further research and development due to its numerous potential uses, such as in the Army care system, in athletic rehabilitation following injury or disability, and in medical trial abnormality therapy. Smartphones, smartwatches, and other portable digital devices generate a plethora of chronic data in the form of recording devices, photo streams, and geographic timbers, among other things. A HAR strategy based on deep learning is going to be very popular, particularly in national healthcare, where personalization based on activity recognition is expected to be highly effective [10]. A power spectrum visual is constructed using an accelerometer pulse and subsequently implanted into a convent by the researchers. Although it increases the initial cost of training the network, the spectrogram synthesis phase essentially replaces the feature extraction step [11]. Execute a one-dimensional inversion on all adequate data utilizing a pure accelerometer sensor as the input to a suitable system. This method may result in the loss of spatial linkages among various sensor components. They focus on publicly available datasets, primarily sourced from embedded sensors (such as smartphones) or monitoring devices. Researchers have utilized identical data and employed analogous methodologies, articulating dynamic impulses within a particular cable through two-dimensional inversion [12]. This study categorizes prominent tasks using a labeled database of fundamental motions to apply convolutional neural networks (CNNs) for job classification, utilizing an inter-multilayer neural network that incorporates linear acceleration and rotational mobility signals.

The data indicators collected from each respondent are employed to train various learning models [7] to develop a monitoring system utilizing six inertial measurement devices. The sorting assignment participants undertake is highly individualized. Following the network analysis, the researchers employed the random forest (RF) classifier to categorize events and developed functionality based on a selection of performance parameters that passed the statistical method. Ultimately, 84.6% of the time was accurate. The study [13] elaborated on a system for activity recognition utilizing haptic feedback inertial sensors and its application in healthcare diagnostics. The integration of Reprieve with consecutive advance drifting scans (RCADS) facilitated feature selection. By the conclusion of the day, k-nearest neighbor (KNN) and Naive Bayes methods have been utilized for choreography comparison and classification. In tasks related to the detection of individuals' daily movements, machine learning algorithms may heavily depend on heuristics, which are subjective techniques for feature extraction. In many instances, the human knowledge base constitutes a significant impediment [8]. A resolution to this issue has been achieved utilizing deep learning methodologies. These techniques can autonomously extract essential features from telemetry data, even in the preprocessing phase, while substantially diminishing the original descriptors and augmenting conceptual patterns. The effectiveness of deep neural networks in classification, computational linguistics, speech synthesis, and other fields has prompted a relatively novel area of research focused on scaling these networks to the capacity of human wearable sensors. Zens et al. [9] proposed utilizing a CNN comprising three convolutional layers and one fully connected layer to represent human movements by converting three-axis sensor readings into a pictorial format. To gain a comprehensive understanding of the field, researchers have examined supplementary materials[10], [11].

The key contributions of our proposed method are:

Reducing the influence of extraneous or noisy data improves recognition accuracy and enables the model to handle complex situations, such as environments with multiple objects or actors, or occlusion. Moreover, we illustrate that attention mechanisms combined with 3D convolutional neural networks and long short-term memory networks can proficiently tackle the subjectivity and variability of action annotations in datasets.

The model exhibits resilience in handling complex scenarios, including those with multiple objects or participants, as well as instances of occlusion, due to its ability to enhance recognition accuracy by mitigating the influence of irrelevant or noisy data. Furthermore, our research demonstrates that the amalgamation of 3D convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, in conjunction with an attention mechanism, significantly reduces the subjectivity and variability commonly encountered in action annotations within datasets.[14].

3 Research Methodology

A system that can identify human actions is built using Deep Learning in this study. A convolutional neural network and a long short-term memory network are the building blocks of the proposed system, which can detect human activity in videos. There are two separate structures and methods that we must employ in TensorFlow. Lastly, we will modify the effective system to estimate videos on YouTube.

Image Classification

The process of image classification involves analyzing a picture to determine its "class" using a computer. (Alternatively, it could truly be a 'class.') 'Vehicle,' 'animal,' and similar terms are examples of classes. To obtain a class prediction, feed an image into a filter. This filter could be a learned deep neural network (CONN-ML) or a traditional classifier.

Video Classification

A key component of video classification is tagging the pixels that make up the majority of the image. A good YouTube clip prophet does a lot too much simply give accurate panel tags, it mimics the actual clip by making use of the tags and elements from successive frames [15], [16]. In one shot, a clip may depict a shrub, but the main title (such as "hiking") could refer to something completely different. The objective specifies the required identifier precision for piece and clip characterization. Typical tasks involve adding a global tag or tags to a clip and adding a description or tags to each panel within the clip.

Convolutional Neural Network

A multi-layer perceptron known as a convolutional neural network (CNN) or ConvNet has become indispensable in visual data operations, particularly in vision estimation and forecasting. It starts with a small set of parameters because it operates with seeds, which are filters, superimposed on the likeness and one generates a salience map, which represents how visible a specific constituent is at a specific location in the image (Fig. 1). But as we go deeper into the intranet, the same number of nodes grows and the size of the graphs shrinks without losing important data using accumulating procedures [17].

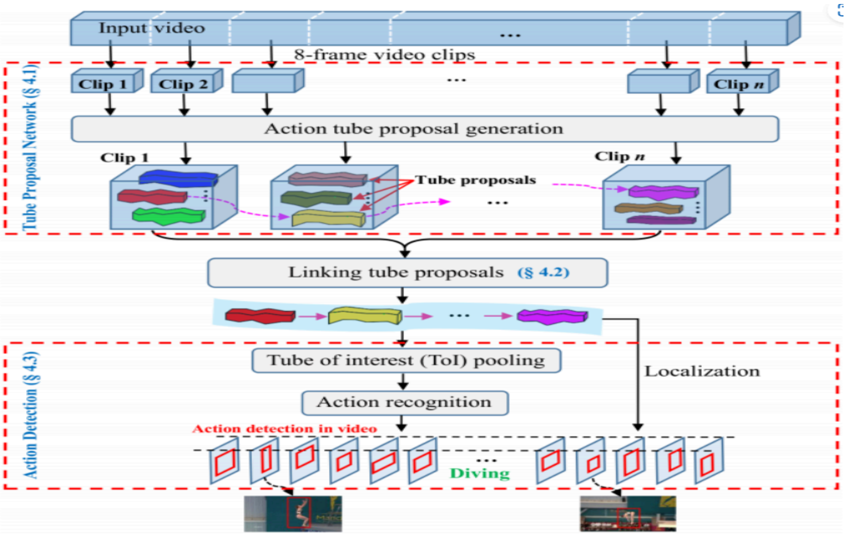

Fig. 1. Convolutional Neural Network architecture used for action detection in video sequences.

Long Short-Term Memory (LSTM)

An LSTM system was developed to process sequential data by effectively synthesizing information from various sources in the correct order to generate output. Although RNN is a subtype of recurrent neural network, it has demonstrated inefficacy in resolving the long-term dependency challenge known as the Vanishing Gradient problem in input sequences. LSTMs were created to mitigate the problem of vanishing gradients, enabling LSTM units to preserve contextual information from lengthy input data sequences [18]. Time series forecasting, speech synthesis, dictionary lookup, and orchestral composition are examples of sequential data challenges that can be more efficiently resolved through the application of an enhanced LSTM. At present, let us exclude all other variables and concentrate exclusively on the manner in which LSTMs can improve our action recognition models[19], [20].

Recurrent Neural Network (RNN)

By recognizing patterns, recurrent neural networks are able to predict what's likely to happen next in the data. Deep learning and the development of algorithms that mimic the actions of neurons in the brain have both made use of RNNs. They're great for situations where understanding is key to anticipating a reaction. In contrast to those other neural networks, they analyze a set of memories that impacts the accuracy of their results by using responses[21]. There are natural cycles and solid documentation that will last. A popular word to describe this occurrence is recall.

Big-O vs. Little-Snap-Ned

A recurrent neural network that incorporates both regular and special subunits is the long short-term memory network. Long short-term memory (LSTM) units have a "memory block" that can hold data for a long time. They can learn dependence over the long term through this memory cell. Their long-term memory retention is the key differentiator between RNN and LSTM (shown in Fig. 2). Since LSTM can retain data in memory for a longer duration, it surpasses RNN in this case[22].

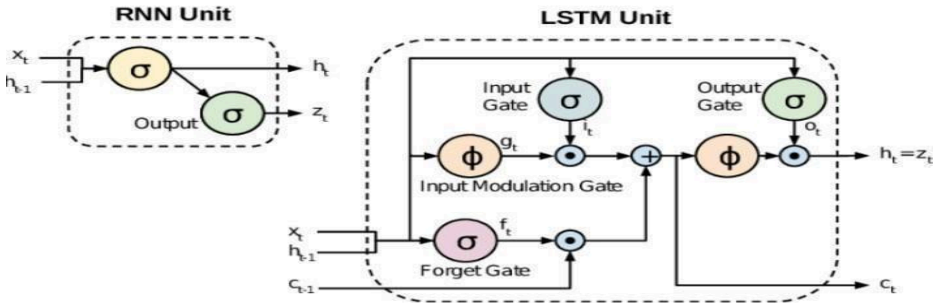

Fig. 2. Comparison of LSTM and RNN architectures, highlighting the key differences in handling sequential data.

4 Data Collection

The UCF101 dataset contains 101 action categories derived from authentic action videos found on YouTube, with the purpose of action recognition[28] (shown in Fig. 3). The Vimeo Activity set of data (UCF50) includes 50 distinct action categories; this dataset is an auxiliary one. Proper preprocessing is essential for the UCF101 dataset videos due to the wide range of video characteristics (camera motion, object appearance, pose, scale, viewpoint, etc.). Take still images from each video and adjust their dimensions so they all have the same resolution[29]. To make training samples more diverse, use data augmentation techniques like color jittering, random cropping, flipping, and rotation.

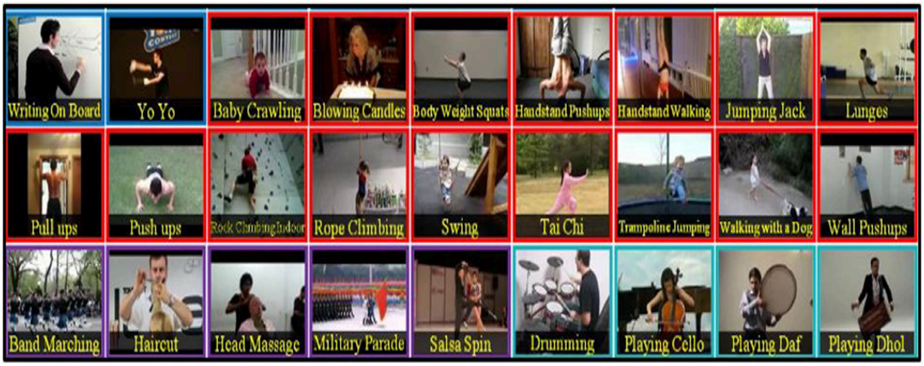

Fig. 3. Categories of human actions analyzed, including walking, running, jumping, and sitting

5 Model Creation

The proposed system was developed using Python and incorporates several widely-used libraries, including OpenCV, TensorFlow, and Keras. The TensorFlow library, developed by Google Brain, is an open-source tool employed for gradient computation, crucial for training deep learning models. Keras, a sophisticated neural network API, is employed to construct the architecture of deep learning models. OpenCV, a specialized library for real-time computer vision, is utilized for tasks such as reading, modifying, and performing various pre-processing operations on videos. The enhancement of the I3D + LSTM-Attention network for video classification can be achieved through the utilization of these libraries.

Performance metrics for the UCF101 dataset: The table summarizes the model's sensitivity and positive predictive value. The precision, recall, and F1-score for each dataset are displayed.

Table 1. Results of the UFC101 dataset, including accuracy, precision, recall, and F1-score for action recognition tasks.

Dataset	Precision	Recall	F1-Score
UCF101	95.17%	92.15%	93.16%

LRCN Model

We will employ time-circulated Conv2D layers, MaxPooling2D, and Dropout components to construct the LRCN architecture. A flattening surface will be employed to convert the features obtained from the Conv2D layers prior to transmitting them to an LSTM layer. The output from the LSTM layer will be utilized by the dense layer with hidden neurons to predict the operation being executed.

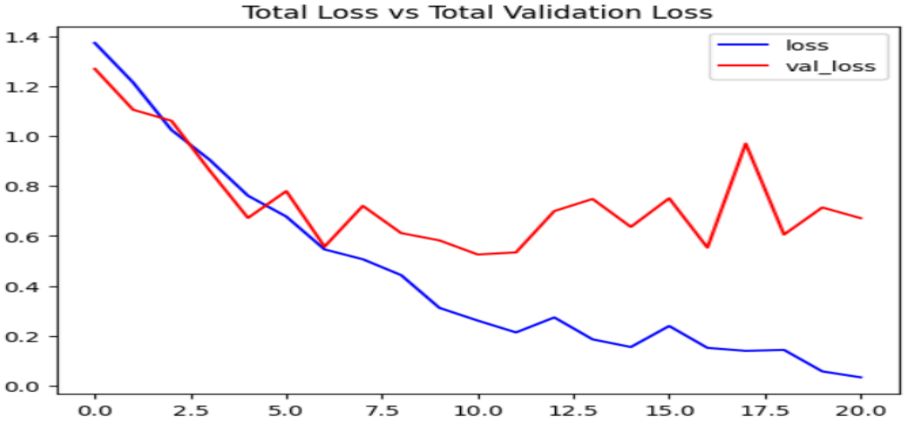

Fig. 4. Visualization of the ConvLSTM architecture, showcasing spatiotemporal feature learning from sequential data.

LRCN Model

We will employ time-circulated Conv2D layers, MaxPooling2D, and Dropout components to construct the LRCN architecture. A flattening surface will be employed to convert the features obtained from the Conv2D layers prior to transmitting them to an LSTM layer. The output from the LSTM layer will be utilized by the dense layer with hidden neurons to predict the operation being executed.
The ConvLSTM visualization is available in Fig. 4 , Fig. 5, Fig. 6.

Accuracy of ConvLSTM model

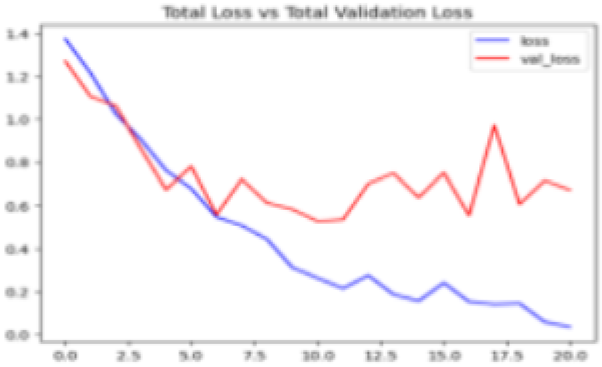

Fig.5. Total loss vs. Total Validation loss

Loss of ConvLSTM model

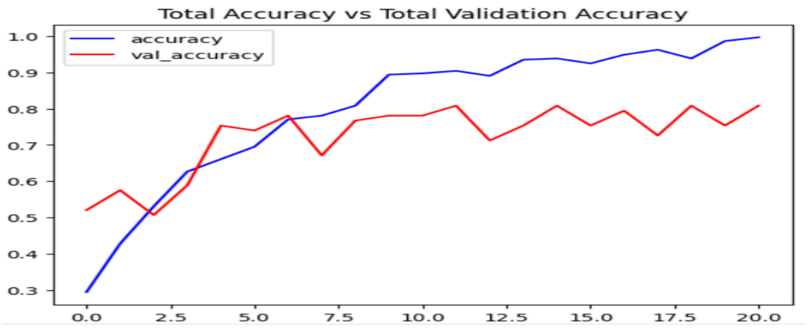

Fig.6. Total Accuracy vs Total Validation Accuracy

LRCN Model Evaluation

We'll analyze the model on testing data after it has been trained, and the accuracy rate is 92 %, with a loss rate of 22 % (Fig. 7, Fig. 8).

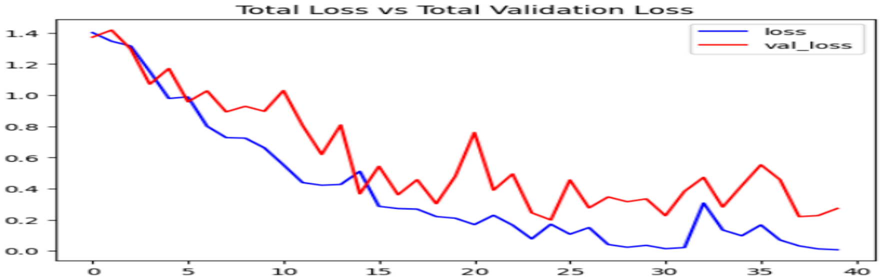

Fig. 7. Visualization of the LRCN approach, illustrating the integration of convolutional layers for feature extraction and recurrent layers for sequence modeling.

Loss of LRCN model

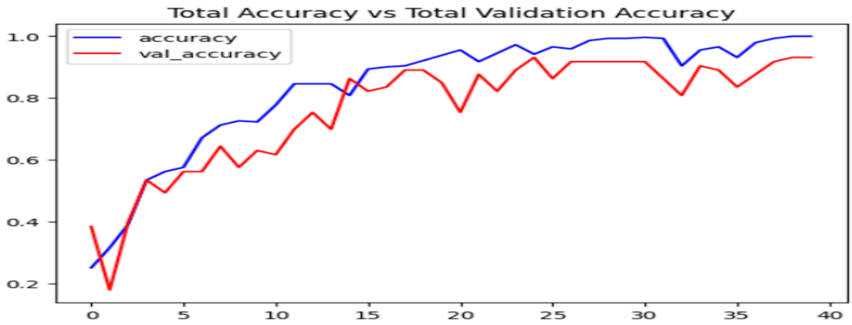

Fig. 8. Comparison of total accuracy and total validation accuracy across training epochs, highlighting the model's performance on training and validation datasets.

6 RESULTS

Our work in this paper draws from two distinct methods: long short-term memory and a combination of convolutional neural networks. Here, ConvLSTM and LRCN are the two methods that have been employed. Various System Configurations and Their Impact on Action Recognition Accuracy on the UCF101 Dataset.

Table 2. Accuracy of various algorithms on the UFC101 dataset, showcasing performance in action recognition tasks.

Method	UCF101 Accuracy (%)
Title (centered)	**Lecture Notes**
2D CNN + LSTM	79.12
3D CNN + LSTM	86.19
3D CNN + LSTM + Attention	91.80
3D CNN + LSTM + Attention + Center Loss	96.83

There is a 92% success rate for the LRCN model and an 80% success rate for the ConvLSTM model. It is clear that LRCN outperforms the ConvLSTM model in terms of accuracy (Table 2). For this reason, we will apply the LRCN Model to YouTube videos in order to make predictions.

7 CONCLUSION

This study demonstrates that our proposed method for video action recognition is effective by illustrating its utilization of an attention mechanism to enhance a Bidirectional LSTM architecture, 3D Convolutional Neural Networks (3DCNN), and transfer learning with Inflated 3D (I3D). Identifying human activities has numerous applications owing to its influence on individual well-being. Personalized medicine, encompassing weight management and geriatric care, has extensively utilized it.

Together, long- and short-term memory convolutional neural networks were employed in this research. We experimented with two distinct methods. For the first model, we used ConvLSTM; for the second, we combined LRCN and TensorFlow to create a model. For 80% accuracy with 89% loss, the ConvLSTM method is the way to go, while the LRCN method yields 92% accuracy with 22% loss. In our comparison of the two methods, we discovered that the LRCN method outperformed ConvLSTM. Predictions on videos on YouTube were thus made using the LRCN Model.

8 REFERENCES

1. Chen, Y., Xue, Y.: A deep learning approach to human activity recognition based on single accelerometer. 2015 IEEE International Conference on

Systems, Man, and Cybernetics (SMC), 1488–1492 (2015). https://doi.org/10.1109/SMC.2015.263
2. Alsheikh, M. A., et al.: Deep activity recognition models with triaxial accelerometers. CoRR, abs/1511.04664 (2015). Available at: http://arxiv.org/abs/1511.04664
3. Zeng, M., et al.: Convolutional neural networks for human activity recognition using mobile sensors. 6th International Conference on Mobile Computing, Applications and Services (MobiCASE), 197–205 (2014). https://doi.org/10.4108/icst.mobicase.2014.257786
4. Yang, J. B., et al.: Deep convolutional neural networks on multichannel time series for human activity recognition. Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI), 3995–4001 (2015).
5. Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. 2016 International Joint Conference on Neural Networks (IJCNN), 381–388 (2016). https://doi.org/10.1109/IJCNN.2016.7727224
6. Pienaar, S. W., Malekian, R.: Human activity recognition using LSTM-RNN deep neural network architecture. 2019 IEEE 2nd Wireless Africa Conference (WAC), 1–5 (2019). https://doi.org/10.1109/AFRICA.2019.8843403
7. Gupta, P., Dallas, T.: Feature selection and activity recognition system using a single triaxial accelerometer. IEEE Transactions on Biomedical Engineering 61(6), 1780–1786 (2014). https://doi.org/10.1109/TBME.2014.2307069
8. Jalloul, N., et al.: Activity recognition using complex network analysis. IEEE Journal of Biomedical and Health Informatics 22(4), 989–1000 (2018).
9. Bengio, Y.: Deep learning of representations: Looking forward. Proceedings of the International Conference on Statistical Language and Speech Processing, 1–37 (2013). Springer.
10. Zheng, Y., et al.: Time series classification using multi-channel deep convolutional neural networks. Proceedings of the International Conference on Web-Age Information Management, 298–310 (2014). Springer.
11. Ordóñez, F., Roggen, D.: Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1), 115 (2016).
12. Lin, Y., et al.: Human activity classification with radar: Optimization and noise robustness with iterative convolutional neural networks followed by random forests. IEEE Sensors Journal 18(23), 9669–9681 (2018).
13. Mario, M.-O.: Human activity recognition based on single sensor square HV acceleration images and convolutional neural networks. IEEE Sensors Journal 19(4), 1487–1498 (2019). https://doi.org/10.1109/JSEN.2018.2882943
14. Pavliuk, O., et al.: Transfer learning approach for human activity recognition based on continuous wavelet transform. Algorithms 16(2), 77 (2023). https://doi.org/10.3390/a16020077

15. Zebhi, S., et al.: Human activity recognition based on transfer learning with spatio-temporal representations. International Arab Journal of Information Technology 18(6) (2021). https://doi.org/10.34028/iajit/18/6/11
16. Xia, K., et al.: LSTM-CNN architecture for human activity recognition. IEEE Access 8, 56855–56866 (2020). https://doi.org/10.1109/ACCESS.2020.2982225
17. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2625–2634 (2015). https://doi.org/10.1109/CVPR.2015.7298878
18. Shanbhag, M. R.: Activity recognition in videos using deep learning. Dissertation (2018).