Identifying Human Actions Via Long-Term Recurrent Convolutional Network

K R Vignesh[1], J D Gowthm[2], Dr. K S Sivle[3]

[1,2]Research Scholar, Department Of Information Technology, K.S.R. College of Engineering (Autonomous), Tiruchengode, Tamilnadu
[3]Associate Professor, Department Of Information Technology, K.S.R. College of Engineering (Autonomous), Tiruchengode, Tamilnadu

Abstract

Automatic identification of human actions from videos has seen significant advancements. Typically, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are employed individually for this purpose. CNNs are trained on pre-existing models to extract visual features from video frames. These features are then utilized by LSTMs to predict outcomes. However, integrating CNN and LSTM layers into a unified architecture known as Long Short-Term Recurrent Convolutional Network (LRCN) yields superior performance. Our study illustrates that a unified LRCN model achieves higher accuracy compared to using CNN and LSTM models separately.

Keywords: Machine learning, LSTM, LRCN, CNN, Human Action Identification.

1. Introduction

Machine learning, a subset of Artificial Intelligence, has significantly reduced the gap in capabilities between humans and machines. Human Action Recognition (HAR) represents a crucial application of AI, employing deep learning techniques to comprehend human behavior. Video-based HAR poses challenges such as background clutter, partial occlusions, viewpoint variations, and lighting changes. Nevertheless, advancements in CNN and LSTM-based learning methods have demonstrated potential in action recognition. This study focuses on analyzing videos to identify specific actions occurring within them. Our findings consistently indicate that integrating CNN and LSTM layers into a unified LRCN model enhances overall performance compared to using CNN and LSTM models separately. This integrated approach leverages both spatial and temporal data to improve the accuracy of action recognition.

The ConvLSTM Model

Convolutional neural network (CNN) models are highly effective in image recognition tasks. However, traditional CNN architectures face challenges in recognizing temporal relationships within video data. This research investigates Convolutional Long Short-Term Memory networks (ConvLSTMs) for video classification. ConvLSTMs are a type of recurrent neural network specifically designed for predicting spatiotemporal data. They integrate convolutional structures into their internal gates, enabling them to learn spatial features from individual frames and temporal interactions between consecutive frames. This capability makes them particularly suitable for tasks like video classification, where capturing both spatial details within frames and temporal dynamics across frames is critical. In a ConvLSTM model, each cell considers inputs and the previous states of its neighboring cells to predict its future state.

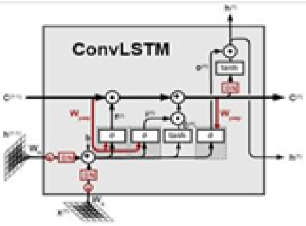

Fig. 1. ConvLSTM

Long-term Recurrent Convolutional Network (LRCN)

LRCNs combine the strengths of CNNs for visual feature extraction and LSTMs for sequential data processing. CNNs excel at extracting features from raw visual data such as images or videos, while LSTMs are adept at handling sequential information. In LRCNs, CNN layers first capture spatial features from the input data, which are then fed into LSTM layers to model temporal dependencies and predict outcomes.

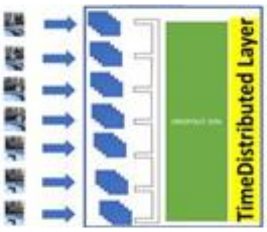

Fig. 2. Time Distributed wrapper layer

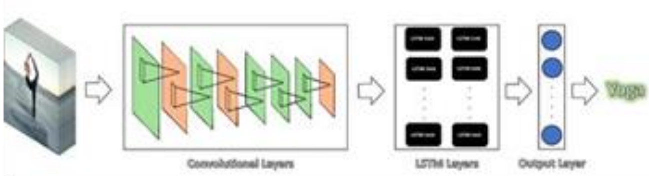

Fig.3.LRCN Structure

The architecture facilitates processing entire videos in a single pass using a Time Distributed wrapper layer. This layer applies the same CNN operation to every frame in the video sequence, allowing efficient processing of entire videos as a cohesive unit. This capability is particularly advantageous as it enables feeding the entire video into the model simultaneously.

LRCNs leverage the complementary strengths of CNNs and LSTMs. CNNs capture spatial characteristics from visual data, while LSTMs model the temporal relationships among these features. Importantly, the weights of both LSTM and CNN layers remain fixed throughout processing, enabling the model to handle sequences of varying lengths.

In summary, LRCNs harness the advancements in CNNs for visual recognition and address the growing demand for models capable of effectively managing data with inputs and outputs that evolve over time. This architecture processes variable-length visual inputs using CNNs followed by stacked LSTM layers, ultimately generating predictions of varying lengths. Such a design facilitates scalable representation learning for sequences of arbitrary length.

2. LITERATURE SURVEY

1. Ng et al. introduced a framework in "Beyond Short Snippets: Deep Networks for Video Classification" utilizing deep neural networks, including the Long-Term Recurrent Convolutional Network (LRCN), demonstrating its effectiveness on datasets such as UCF-101 and Sports-1M. [1]
2. Wang et al. in "Action Recognition with Improved Trajectories" combined hand-crafted and deep features acquired by LRCN, achieving state-of-the-art results on datasets like Hollywood2, HMDB-51, and UCF-101. [2]
3. Vu et al. provided an overview of deep learning techniques, including LRCN, for human action recognition in "Deep Learning for Human Action Recognition: Comprehensive Review," assessing different approaches and suggesting future research directions. [3]
4. Li et al. described a multiple streams bi-directional recurrent neural network for fine-grained action recognition in "A Multiple Streams Bi-Directional Recurrent Neural Network for Fine-Grained Action Recognition," achieving high performance on datasets such as JHMDB-21 and UCF101. [4]
5. Zhang et al. proposed a multi-modal fusion method in "Multi-modal Fusion Method for Human Action Recognition Based on IALC," combining LRCN with a multi-stream CNN to extract spatial and temporal data from videos effectively. [5]
6. Tran et al. proposed "Learning Spatiotemporal Features Using 3D Convolutional Networks," introducing a 3D CNN for video classification that incorporates long-term relationship detection (LRCN), achieving impressive results on datasets like HMDB-51 and UCF-101. [6]
7. Lea et al. proposed "Temporal Convolutional Networks for Action Segmentation and Detection," utilizing LRCN within a temporal convolutional network (TCN) framework to segment and detect actions in long videos, showcasing high performance on datasets such as ActivityNet and THUMOS14. [7]
8. Raj et al. suggested an improved human activity recognition technique using convolutional neural networks in "An Improved Human Activity Recognition Technique Based on Convolutional Neural Network," achieving 97.2% accuracy with their CNN-based model, surpassing previous methods. [8]
9. Donahue et al. recommended the use of LRCN for visual recognition and video description in "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," demonstrating its utility across various datasets. [9]
10. Todorovic et al. proposed "Temporal Deformable Residual Networks for Action Segmentation in Videos," introducing a temporal residual network (TRN) that combines two streams for enhanced frame classification accuracy, performing well on various datasets. [10]
11. Authors developed a spatio-temporal attention model integrating LRCN to improve recognition accuracy in "A Spatio-Temporal Attention Model for Action Recognition," achieving state-of-the-art performance on datasets like HMDB-51 and UCF-101. [11]
12. Peng et al. suggested a joint multiple feature learning approach integrating LRCN with deep belief networks for sensor-based human action recognition in "Learning Multi-level Features for Sensor-based Human Action Recognition," achieving cutting-edge performance across datasets including UCF-101 and HMDB-51. [12]
13. Tian reviewed deep learning-based action recognition methodologies, focusing on LRCN, in "Deep Learning-Based Action Recognition from Untrimmed Video Streams: A Survey," discussing recent advances

and future directions in the field. [13]

14. Xu proposed a recurrent convolutional neural network approach for video classification in "Recurrent Convolutional Neural Network for Video Classification," utilizing a two-stage LRCN to capture long-range dependencies effectively across datasets including UCF-101 and HMDB-51. [14]

15. Xia introduced "A Channel-Wise Spatial-Temporal Aggregation Network for Action Recognition," leveraging LRCN to capture spatial and temporal data with a unique filter bank technique, achieving advanced performance across datasets such as UCF-101 and HMDB-51. [15]

Table 1. *Literature review summary*

S.No	Title	Work Done	Techniques	Findings	Future Scope
1.	"Beyond Short Snippets: Deep Networks for Video Classification"	Suggested a deep network that uses LRCN for video categorization.	LRCN	Reached cutting-edge results on a number of datasets.ts.	Investigate the application of attention mechanisms for LRCN.
2.	"Action Recognition with Improved Trajectories"	Proposed an action recognition method using improved trajectories and LRCN.	Improved Trajectories, LRCN	Several datasets were utilized to attain innovative outcomes.	Apply LRCN to different kinds of trajectory characterstics.
3.	"Deep Learning for Human Action Recognition: A Comprehensive Review"	Reviewed the use of deep learning for human action recognition, including LRCN.	Literature Review	Provided an overview of current investigations in this area.	Investigate the usage of LRCN along with different varity of deep learning architectures.
4.	"A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Recognition"	LRCN was utilized to propose a multi-stream, bi-directional recurrent neural network for action recognition.	Bi-Directional RNN, LRCN	Achieved cutting edge performance on many datasets.	Investigate the utilization of LRCN with different types of multi-stream architectures.
5.	"Multi-modal fusion method for human action recognition based on IALC"	Proposed a multi-modal deep learning approach for human action recognition using IALC.	Multimodal Deep Learning, IALC	Attained remarkable outcomes across many datasets.	Investigate the use of IALC with other variety of modalities.

6.	"Learning Spatiotemporal Features using 3D Convolutional Networks"	Proposed a 3D convolutional neural network for learning spatiotemporal features for action recognition.	3D Convolutional Neural Network	reached cutting-edge results on multiple datasets.	Examine the application of LRCN to other 3D architectures.
7.	"Networks using Temporal Convolutions for Detecting and Segmenting Actions"	Proposed a temporal convolutional network for action segmentation and detection using LRCN.	Temporal Convolutional Neural Network, LRCN	reached cutting-edge results on a number of datasets.	Examine how LRCN can be used in conjunction with a variety of segments and detection approaches.
8.	"An improved human activity recognition technique based on convolutional neural network"	Proposed a method for Human Activity Recognition (HAR) using Convolutional Neural Networks (CNNs).	Human Activity Recognition (HAR) , Convolutional Neural Networks	reached cutting-edge results on a number of datasets.	further advancements in CNN architectures and sensor data fusion for even more accurate and diverse human activity recognition.
9.	"Long-term Recurrent Convolutional Networks for Visual Recognition and Description"	Proposed LRCN for visual recognition and description.	LRCN	accomplished novel conclusions on a number of datasets.	Examine how LRCN can be applied to different kinds of visual identification and descriptive tasks.
10	"Temporal Deformable Residual Networks for Action Segmentation in Videos"	proposed a novel deep learning model called Temporal Deformable Residual Network (TDRN) for segmenting human actions in videos.	residual networks, multi-scale processing, and deformable convolutions	reached cutting-edge results on a number of datasets, including THUMOS14 and ActivityNet	exploring further applications of TDRN for more complex action segmentation tasks and potentially adding additional modalities beyond just video data.
11	"Spatio-Temporal Attention Model to	Proposed an end-to-end spatio-temporal	LRCN, Attention	Showed cutting-edge results on a number of	In future research for improving the

	improve recognition precision.	attention model that uses LRCN to capture spatial and temporal information and incorporate attention mechanisms to improve recognition.	Mechanism	datasets,such as UCF-101 and HMDB-51	efficient working of attention mechanism
12	"Joint Multiple-Feature Learning for Action Recognition"	Devised a joint multiple feature learning strategy that combines LRCN with a deep belief network to capture both appearance and motion information.	LRCN, with a deep belief network	Achieved exceptional results on various datasets, including UCF-101 and HMDB-51.	Further research on improving the performance of the joint learning methodology.
13	"Deep Learning-Based Action Recognition from untrimmed Video Streams: A Survey"	Offered a complete evaluation of deep learning-based approaches for action recognition, including LRCN.	LRCN	Recent advances and potential future study areas in deep learning-based action recognition were analyzed.	More research is being done into expanding more effective deep learning-based systems for action recognition.
14	"Recurrent convolutional neural network for video classification"	Suggested an efficient and accurate video classification system which uses a two-phase LRCN to collect long-term dependencies.	Two stage LRCN	Achieved innovative results on various datasets, including UCF-101 and HMDB-51	More research is being done on improving the performance of the two-stage LRCN.
15	"A Channel-Wise Spatial-Temporal Aggregation Network for Action Recognition"	Proposed a spatial-temporal discriminative filter bank approach that uses LRCN to capture both spatial and temporal information	Filter bank technique that leverages LRCN	Achieved state-of-the-art performance on several datasets, including UCF-101 and HMDB-51	Additionally research in developing effective filter bank approaches for activity recognition

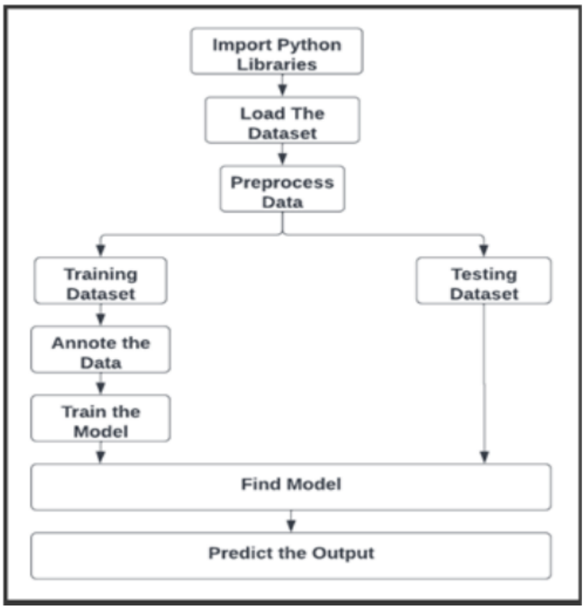

Fig. 4. Block Diagram of project

3. METHODOLOGY

The proposed approach for recognizing activities in videos is detailed below, involving four main processes: data collection, data preprocessing, model construction, and model evaluation.

Step 1: Data Collection

The first step involves gathering a comprehensive dataset of videos that depict the activities of interest. This dataset should encompass diverse camera movements, object appearances, postures, scales, perspectives, backgrounds, and lighting conditions to ensure the model's ability to generalize across various scenarios.

Step 2: Data Preprocessing

Next, the videos are segmented into individual frames. These frames are then converted into a suitable format for machine learning models, typically NumPy arrays. Class labels associated with each video are encoded using one-hot encoding, which represents categorical variables as vectors. Subsequently, the data is shuffled randomly and split into training and testing sets, often with a standard split of 74% for training and 26% for testing.

Step 3: Model Construction

The core of the method involves constructing a Convolutional Long Short-Term Memory (ConvLSTM) network to extract spatiotemporal information from the videos. This architecture combines the strengths of LSTMs, which capture temporal relationships between successive frames, and Convolutional Neural Networks (CNNs), which excel in extracting spatial features from individual video frames. The typical model architecture includes temporally stacked Conv2D layers for feature extraction, followed by MaxPooling2D layers and dropout layers to prevent overfitting. To capture temporal dependencies among frames, the extracted features are flattened and fed into an LSTM layer. Finally, a dense layer predicts the activity class for each video.

Step 4: Assessing the Model

Once the model is constructed, it is trained on the training dataset. After training, the model's performance is evaluated on the testing set using various metrics such as precision, recall, accuracy, and F1 score. If the evaluation results indicate poor performance, it may be necessary to explore more sophisticated model architectures or adjust hyperparameters to improve the model's accuracy and robustness.

4. Methods

Data Collection

In this study, we utilized the UCF50 dataset, which consists of real-world action recognition videos sourced from YouTube, encompassing 50 distinct action categories. The primary objective of using this dataset was to provide a valuable resource for action recognition research within the computer vision community. However, the dataset presents considerable challenges due to variations in lighting, background clutter, perspective, scale, posture, object appearance, camera motion, and other factors. Each category in UCF50 comprises 25 video groups, each containing more than four video clips. Videos within the same group may share similarities such as being filmed by the same person, having similar backgrounds, or exhibiting comparable viewpoints.

Design Approach

Initially, our data structure resembled a NumPy array, where each retrieved video frame was matched with a one-hot encoded class label. Subsequently, the collected data was split into training and testing sets, with 74.8% allocated to the training data and the remaining 25% to the testing data. To ensure unbiased splits that reflect the overall data distribution, the dataset was shuffled randomly before partitioning.

Two models were implemented for human activity recognition:

1. ConvLSTM Model: This model aimed to capture both spatial and temporal relationships among video frames. Unlike traditional LSTMs that handle only 1-dimensional data, ConvLSTM accepts 3-dimensional input (channel count, width, and height) owing to its convolutional structure, allowing it to directly represent spatiotemporal information. Recurrent layers from Keras' ConvLSTM2D were used to construct the model. The number of filters and kernel sizes for convolutional operations were specified within the ConvLSTM2D layer configuration. The outputs from these layers were flattened and fed into a dense layer with softmax activation to compute the probability for each action category. Dropout and MaxPooling3D layers were also included to mitigate overfitting. However, the model's performance was limited by its relatively simple architecture, inability to handle large datasets, and a smaller number of trainable parameters.
2. LRCN Model: This model provided a more effective approach by integrating convolutional and LSTM layers within the same framework. Convolutional layers extracted spatial features from video frames, while LSTM layers modeled temporal sequences over time. This integration enabled the network to learn spatiotemporal features comprehensively during training, resulting in a more robust model. Our LRCN architecture consisted of time-distributed Conv2D layers followed by MaxPooling2D and Dropout layers.

Table 2. ***ConvLSTM MODEL***

Layer type	**Filter size**	**stride**	**Outcome shape**	**parameters**
Conv_lst_m2d_8	4	-	(None,20,62,62,4)	1024
Max pooling 3	4	-	(None,20,31,31,4)	0
Time_dist	4	-	(None,20,31,31,4)	0
Conv_lstm_2d	4	-	(None,20,29,29,8)	3488
Max_pooling_3	4	-	(None,20,15,15,8)	0
Time dist	4	-	(None,20,15,15,8)	0
Conv-lstm_2d	4	-	(None,20,13,13,14)	11144
Max_pooling	4	-	(None,20,7,7,14)	0
Time_dist	4	-	(None,20,7,7,14)	0
Conv lstm 2d	4	-	(None,20,5,5,16)	17344
Max_pooling	4	-	(None,20,3,3,16)	0
Flatten	4	-	(None,2880)	0
dense	4	-	(None,4)	11524

The flatten layer reshaped the features extracted from the Conv2D layers before passing them into an LSTM layer. Following this, a dense layer with softmax activation predicted the action based on the output from the LSTM layer. The Time Distributed wrapper layer allowed the application of the same layer independently to each frame of the video, facilitating the processing of the entire video sequence within the model at once.

In comparison to the ConvLSTM model, the LRCN model had a significantly higher number of trainable parameters (73,060). As evidenced by the results, the LRCN model demonstrated superior performance, particularly on datasets with fewer classes.

Table 3. ***LRCN Model***

Layer type	**Filter size**	**stride**	**Outcome shape**	**parameters**
Time_distribution_22	4	-	(None,20,64,64,16)	448
Time distribution 23	4	-	(None,20,16,16,16)	0
Time_distribution_24	4	-	(None,20,16,16,16)	0
Time_distribution_25	4	-	(None,20,16,16,32)	4640
Time distribution 26	4	-	(None,20,4,4,32)	0
Time_distribution_27	4	-	(None,20,4,4,32)	0
Time_distribution_28	4	-	(None,20,4,4,32)	18496
Time distribution 29	4	-	(None,20,2,2,64)	0
Time_distribution_30	4	-	(None,20,2,2,64)	0
Time_distribution_31	4	-	(None,20,2,2,64)	36928
Time distribution 32	4	-	(None,20,1,1,64)	0

Time_distribution_33	4	-	(None,20,64)	0
Lstm		-	(None,32)	12416
dense		-	(None,4)	132

5. Result and Discussion

Human activity identification (HAI) is crucial for understanding human behavior by extracting insights from raw data to recognize objects, people, and their actions within videos. This study compared two approaches for HAI: ConvLSTM and LRCN models.

The ConvLSTM model achieved an accuracy of 81% with a corresponding loss of 9%. In contrast, the LRCN model demonstrated superior performance with an accuracy of 92.62% and a significantly lower loss of 2.24%. This indicates that the LRCN model effectively learned patterns from raw video data, achieving approximately 92% accuracy on a dataset of 50 videos trained over 50 epochs.

The direct utilization of raw data within the learning model contributed to the high accuracy achieved by the LRCN model. It is noteworthy that this approach could potentially achieve even higher accuracy with a larger dataset. The LRCN model consistently outperformed the ConvLSTM model, showing strong generalizability and high recognition accuracy without the need for complex feature extraction.

However, while the LRCN technique offers significantly higher accuracy, it requires increased processing memory and time resources. This is supported by experimental outcomes depicted in the graphs. These findings underscore the potential of deep learning models to significantly enhance human activity detection.

We achieved decent accuracy using Model 1, a ConvLSTM trained over 45 epochs. However, this model slightly overfitted, resulting in a loss of 8.9%.

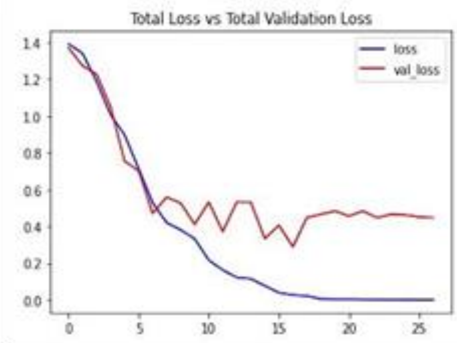

Fig. 5. Graph of Interval versus Loss

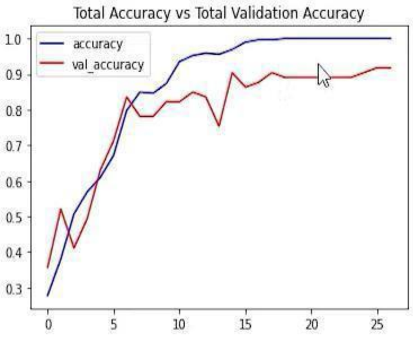

Fig.6. Graphs of Interval vs Accuracy

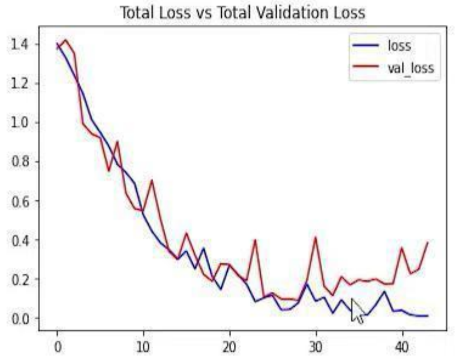

Fig. 7. Interval versus Loss Graph

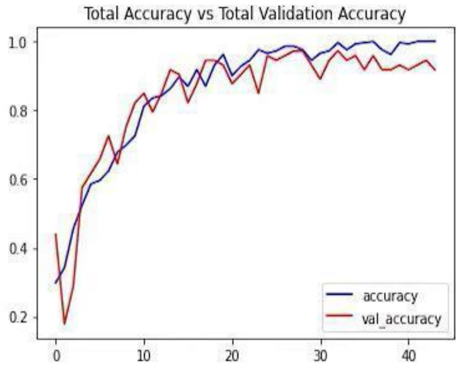

Fig.8. Interval versus Accuracy Graph

6. Conclusion and future work

Conclusion:

The current study proposes an LSTM-CNN approach for human activity recognition, aiming to achieve two primary objectives:

1. Enhanced Robustness in Feature Extraction: The CNN component is designed to extract robust features from the data, reducing susceptibility to input variations.
2. Improved Accuracy in Activity Detection: The LSTM component utilizes its capability to learn temporal dependencies, thereby enhancing the accuracy of activity classification.

Our results consistently demonstrate that integrating CNN and LSTM models outperforms their individual use. This integrated approach benefits from the spatial learning abilities of CNNs and the temporal learning capabilities of LSTMs.

Future Work:

Building upon this research, our objectives include:

- Evaluating the model's adaptability across diverse datasets: This will assess its capability to generalize effectively in various activity recognition tasks.
- Exploring its feasibility for real-time applications: Investigating methods like incremental learning or online training to overcome the current constraint of retraining on new data.

The successful deployment of this approach could pave the way for advancements in next-generation wearable technologies, such as smartwatches. These devices could monitor user activities in real-time, offering valuable insights into daily routines and potentially serving as health monitoring tools. Moreover, the application could extend to scenarios requiring continuous monitoring, such as elderly care or controlled environments. However, a limitation of this approach is its inability to handle real-time data for training, necessitating model retraining whenever new data is introduced.

References

1. J. Yue-Hei presented "Beyond Short Snippets: Deep Networks for Video Classification" at CVPR 2015.
2. H. Wang and C. Schmid discussed "Action Recognition with Improved Trajectories" at ICCV 2013.
3. D. Q. Vu, T. P. T. Thu, N. Le, and J. C. Wang conducted a comprehensive review titled "Deep Learning for Human Action Recognition" published in APSIPA Transactions on Signal and Information Processing, 2023.
4. X. Li et al. introduced "A Multiple Streams Bi-Directional Recurrent Neural Network for Fine-Grained Action Recognition" at CVPR 2016.
5. Y. Zhang et al. proposed a "Multi-modal Fusion Method for Human Action Recognition based on IALC" in IET Image Processing, 2022.
6. D. Tran et al. presented "Learning Spatiotemporal Features With 3D Convolutional Networks" at ICCV 2015.
7. C. S. Lea et al. discussed "Temporal Convolutional Networks for Action Segmentation and Detection" at CVPR 2017.
8. R. Raj and A. Kos developed "An Improved Human Activity Recognition Technique based on Convolutional Neural Network" published in Scientific Reports, 2015.
9. J. Donahue, L. A. Hendricks, and S. Guadarrama proposed "Long-term Recurrent Convolutional Networks for Visual Recognition and Description" at CVPR 2015.
10. P. Lei and S. Todorovic introduced "Temporal Deformable Residual Networks for Action Segmentation in Videos" at CVPR 2022.
11. L. Meng, B. Zhao, and B. Chang discussed "Interpretable Spatio-Temporal Attention for Video Action Recognition" at ICCVW 2019.
12. Y. Xu and Z. Shen developed "Learning Multi-level Features For Sensor-based Human Action Recognition" published in Pervasive and Mobile Computing, 2016.
13. E. Vahdani and Y. Tian conducted a survey titled "Deep Learning-Based Action Detection in Untrimmed Videos" published in TPAMI, 2022.
14. Z. Xu, J. Hu, and W. Deng proposed "Recurrent Convolutional Neural Network for Video Classification" at ICME 2016.
15. H. Wang, H. Li, and T. Xia introduced "A Channel-Wise Spatial-Temporal Aggregation Network for Action Recognition" in Mathematics, 2021.