Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Enhanced Action Recognition through Deep Spatiotemporal Learning Using 3D CNN and GRU

Neha Bansal, Atul Bansal, Manish Gupta

Department of Electronics & Communication Engineering, GLA University, India
Email: neha.bansal_phdece20@gla.ac.in

The issues revolve around efficiently analyzing large video data streams while minimizing computer complexity and performing processing in real-time. On the other hand, it becomes more difficult to quickly react to unusual actions because of this. Also, smart homes, security systems, assisted living facilities, and health monitoring might all benefit from the ability to recognize events from video sequences. The techniques used to analyse data are still under constant scrutiny, even if sensing technology has advanced, especially with respect to 3D video. By combining 3D Convolutional Neural Networks (CNN) with gated recurrent units (GRU), we have created a new method for learning spatiotemporal features in movies. We found that 3D convolutional neural networks (CNNs) acquire spatiotemporal information better than 2D CNNs using the UCF50 dataset. Using smaller 3x3x3 convolution kernels in a uniform design also improves performance. Furthermore, we found that 3D CNN with GRU integrated yields better accuracy than 3D CNN alone. The results show that GRU outperforms LSTM in terms of accuracy (89.89%) and calculation time (less than LSTM) when compared.

Keywords: Action Recognition, GRU, CNN, LSTM, UCF50, Spatiotemporal learning.

1. Introduction

The abundance of multimedia content being shared on the internet, particularly in the form of videos, has resulted in a vast amount of data that needs to be analyzed and comprehended [1]. To address this issue, video analytics has been developed to examine the videos for research purposes. This field of study has been ongoing for several years, focusing on various aspects, such as recognizing actions, detecting anomalous events, and comprehending activities within videos [2][3]. There has been notable progress in dealing with specific challenges related to video analysis by implementing various solutions. However, the demand for a all-

encompassing video descriptor that can effectively handle real world videos in a consistent manner. A good video descriptor must have four important attributes to be considered effective [4][5]. Firstly, the representation should be generalizable and capable of effectively capturing diverse video content, while also maintaining distinctiveness. For instance, videos on the web can range from nature, sports, Firstly, the descriptor needs to be versatile and capable of effectively representing various categories of videos while maintaining its ability to differentiate between them. Secondly, the descriptor should be concise in order to handle massive amounts of video data efficiently during processing, storage, and retrieval [6]. Thirdly, the descriptor should be computationally efficient since real-world systems are expected to process thousands of videos per minute. A desirable feature of a good descriptor is that it should be straightforward to implement and not rely on complicated methods for feature encoding or classification [7]. It should be able to provide strong performance even when used with simple models such as linear classifiers [8] [18].

The advancement of deep learning in the realm of image processing has encouraged researchers to advance in feature learning. Although many pre-trained CNN models have been developed to extract image features that are useful for transfer learning tasks, they are not appropriate for videos as videos not only have spatial features but also time dependent temporal features. To solve this issue, the authors propose the use of deep 3D CNN to learn both spatial and temporal features for videos [10]. These features, along with a straightforward aligned classifier, greatly enhance the efficiency of video interpretation tasks. After rigorously testing of large datasets and modern 3D CNN architectures researchers proved that 3D CNN provide much better performance in case of video dataset [9]. The properties of 3D CNNs include the ability to extract spatiotemporal features from videos, which encode information about objects, scenes, and actions [19]. This makes them useful for multiple tasks without requiring the model to be adjusted for each individual task [14][15]. Additionally, 3D CNNs can be trained on large-scale datasets and modern deep architectures, which allow them to provide superior performance in various video analysis tasks. Overall, the properties of 3D CNNs make them an effective tool for video interpretation and analysis [16] [17]. 3D deep convolutional neural networks are effective in learning features from videos because they can capture both the spatial information of objects and the temporal information of movements at the same time [10].

2. Related works

For many years, computer vision experts have explored different problems related to videos, such as detecting anomalies, recognizing actions, recovering videos, and detecting events and actions. During this time, much of the research has been directed towards discovering efficient methods to represent videos. Several techniques have been proposed, such as spatiotemporal points of interest (STIPs) by Laptev and Lindeberg, extending SIFT and HOG to SIFT-3D and HOG3D for action recognition, cuboid features for action recognition like USD, and Action Bank for stock recognition by Sadanand and Corso. Li et al. [11] The research investigates Human Activity Recognition (HAR) with wearable sensors for health-related purposes. Traditionally, Artificial Neural Networks (ANNs) are employed; however, they are computationally intensive and exhibit constraints in temporal feature extraction. The

researchers advocate for the utilization of Spiking Neural Networks (SNNs), modeled after biological neurons, for Human Activity Recognition (HAR) applications. Spiking Neural Networks (SNNs) surpass Artificial Neural Networks (ANNs) and decrease energy usage by as much as 94%. Nafea et al. [12] presents a novel approach to feature capture combining bidirectional long short-term memory (BiLSTM) and convolution neural networks (CNN) with variable kernel size. What makes this study unique is that it successfully uses conventional convolutional neural networks (CNNs) and bidirectional long short-term memory (BiLSTMs) to extract spatial and temporal information from sensor input, and it successfully selects the appropriate video representation. This suggested approach utilizes data obtained from many sources, such as accelerometers, sensors, and gyroscopes, and is based on Wireless sensor data mining (WISDM) and UCI datasets. The outcomes prove that the suggested method effectively enhances HAR. Therefore, it was determined that the suggested technique outperformed other current methods in terms of accuracy, with a better score in the WISDM dataset (98.53% vs. 97.05%). Tasnim et al. [13] propose a method for 3D skeletal joint action discrimination using spatio-temporal image formation (STIF) that records both spatial information and changes in time. We assess the suggested approach using multiple fusion methods and perform transfer learning on pre-trained models—MobileNetV2, DenseNet121, and ResNet18—trained with the ImageNet dataset—in order to extract discriminative features. Human action recognition performance variance is the primary focus of our investigation into three fusion methods: element-wise average, multiplication, and maximum. Comparing our technique to previous research, which used publicly available benchmark 3D skeleton datasets with STIF representation, we find that it beats both UTD-MHAD (University of Texas at Dallas multi-modal human action dataset) and MSR-Action3D (Microsoft action 3D). Our results using MobileNetV2, DenseNet121, and ResNet18 for the UTD-MHAD and MSR-Action3D skeleton datasets are 96.00%, 98.75%, and 97.08% accurate, respectively.

3. Proposed Approach

This section provides a detailed explanation of the fundamental operations of 3D ConvNets, performs an empirical analysis of various 3D ConvNet architectures, and outlines the training process on large datasets to learn features.

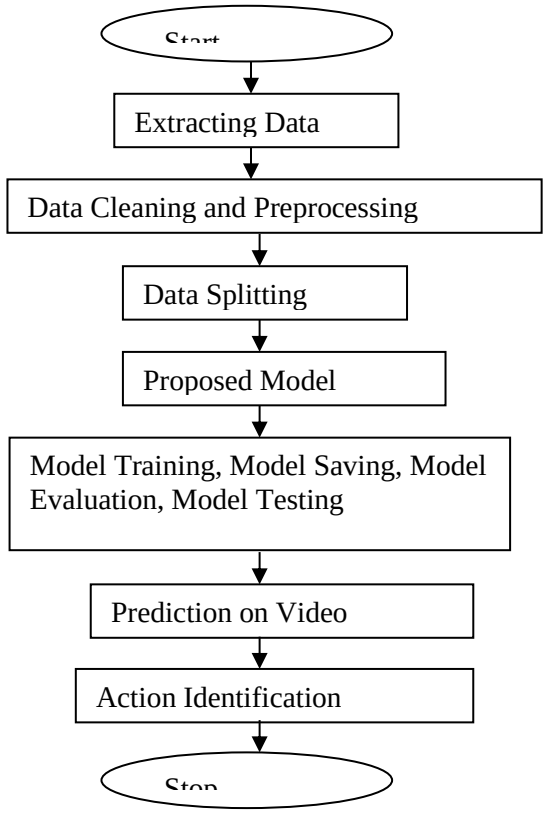

Fig. 1 Flow of proposed approach

The 3D convolutional neural networks (CNN) and RNN models such as gated recurrent units (GRU) has become a well-known research aspect in video analysis, as it authorizes for capturing spatiotemporal features and modeling temporal movement concurrently. In recent senility, various research documents have proposed distinct architectures for 3D CNN-GRU models and enforced them to tasks in the way that action recognition, video emphasize, and anomaly discovery. This assimilation has shown auspicious results in developing the preciseness of video analysis related to using only 3D CNN or GRU solely.

3.1 Gated Recurrent Unit: GRU Model

In 2014, Cho et al. introduced the GRU model as an alternative to standard RNNs, which struggle with vanishing gradients and are unsuitable for handling long-term dependencies. The GRU model has fewer parameters than the LSTM model, making it quicker to train and more efficient in terms of memory usage. The GRU architecture is a versatile model that can be employed for a range of sequence learning tasks such as speech modeling, machine translation, language recognition, and image captioning. The GRU model has shown to be effective in transfer learning settings, where it is trained on a very-large datasets and also can be fine-tuned on a smaller dataset for specific tasks. The GRU model perhaps stacked to form deep recurrent neural networks, which can further boost the model's performance on complex tasks. The GRU model is a strong appliance for processing sequential data, but it cannot suit for all types of data. It is essential to carefully analyze the features of the data and the necessities of the task

before selecting a model design.GRU is a RNN based model i.e. usually used in sequential data processing tasks in the way that the study of computers and time-series analysis. The GRU model is planned to address the question of vanishing gradients in traditional RNNs by utilizing gating methods that admit the network to selectively amend and ignore information. The GRU model subsists of a single layer of units that process sequential input data. Each unit has a private state h(t) that shows the current memory of the unit. The unit takes the previous private state h(t-1) and an input vector x(t) as inputs, and computes the current private state h(t) in this manner:

Update Gate: The update gate (z(t)) decides by virtue of what much of the prior memory (h(t-1)) bear be saved and how much of the current input (x(t)) should be amounted to the current memory.

Reset Gate: The reset gate (r(t)) decides how much of the former memory bear be overlooked when calculating the current memory.

Current Memory: The current memory (h~(t)) is the candidate memory that can be added to the current memory.

Final Hidden State: The final hidden state (h(t)) is the revised memory i.e. presented by linking the prior memory and the current memory.

By applying the gating systems of the reset gate and update gate, the GRU model is capable to selectively update and overlook information that admits it to handle unending dependencies and avoid the vanishing gradient problem. The output of the sigmoid function maybe elucidated as the chance of the neuron being "triggered" or "arousing". The sigmoid function is frequently used in neural network architectures to a degree logistic regression, feedforward neural networks, and recurrent neural networks like GRUs and LSTMs.

3.2 Convolution and Pooling

According to the authors, 3D ConvNets are more appropriate for learning spatiotemporal features than 2D ConvNets. This is because 3D convolution and pooling operations can better capture temporal information. When 2D convolution is applied to a single image, it generates a single image as the output, but when applied to multiple images; it creates multiple images as channels. This results in a loss of temporal information after each operation. While, The 3D CNN preserves this information and produces an output volume. The slow fusion model distinguishes itself by using 3D convolution in the first three layers, resulting in the best performance. Nonetheless, the temporal information is still lost after the third CNN layer in this model. The aim of this segment is to find the optimal architecture for 3D ConvNets by conducting experimental studies. However, since training models on large datasets is resource-intensive, the researchers first conduct experiments on a moderately-sized dataset, UCF 50, to identify the best architecture. They then validate their findings by performing a limited number of experiments on a large-scale dataset. To ensure consistency with the findings of 2D ConvNet, which suggest that a small receptive field of 3x3 nuclei with a deeper structure produces the best results, the researchers keep the spatial receptive field fixed at 3x3. The researchers only vary the temporal depth by 1,3,5,7 of the 3D convolution kernel in their architecture study. The authors believe that 3D ConvNets are superior to 2D ConvNets when it comes to learning spatial and temporal characteristics respectively. This is due to the fact

that 3D convolution and pooling procedures are capable of capturing temporal information more accurately. When applied to a single image, 2D convolution results in the generation of a single image as the output; however, when applied to many photos, it results in the generation of multiple images that are used as channels. Because of this, temporal information is lost after every operation that is performed. The 3D CNN, on the other hand, stores all of this information. The research focuses on a fusion model that largely makes use of 2D convolutions; nonetheless, the majority of networks suffer from a loss of information pertaining to the passage of time. The slow fusion model, on the other hand, differentiates itself from the others by employing 3D convolution in the first three layers of the network, which ultimately leads to the best performance. Despite this, after the third CNN layer in this model, the temporal information is still lost.

3.3 Integration of 3D CNN and GRU

The integration of 3D CNN and GRU is a authoritative approach for video interpretation that acknowledges for the modeling of spatiotemporal features and temporal dynamics synchronously. In this approach, the 3D CNN extracts spatial features, and the GRU model is used to extract the temporal features from the input video frames. The 3D CNN-GRU model conceivably presented mathematically in this manner:

Input image → X = {x1, x2, ..., xn} ,

State Sequence (GRU) → H = {h1, h2, ..., hn}

Height - h; Width - w; Depth - d

Reset - r_t; update - z_t

$r_t = \sigma(Wr [f_t, h_{t-1}] + br)$

$z_t = \sigma(Wz [f_t, h_{t-1}] + bz)$

$h_t = (1 - z_t) * h_{t-1} + z_t * \tanh(W[f_t, r_t * h_{t-1}] + b)$

Where r_t and z_t are the reset and update gates, σ is the sigmoid stimulus function, Wr, Wz, and W are the weight matrices, and br, bz, and b are the bias vectors.

The definitive output of the 3D CNN-GRU model is usually acquired by administering a completely connected layer to the last hidden state hn, trailed by a softmax stimulus function to achieve class probabilities for a regulation task or a regression function for a regression task.

Overall, the integration of 3D CNN and GRU determines a dominant framework for modeling spatiotemporal features and temporal dynamics in videos, and has existed proved expected potent in tasks to a degree activity recognition, video summarization, and anomaly detection.

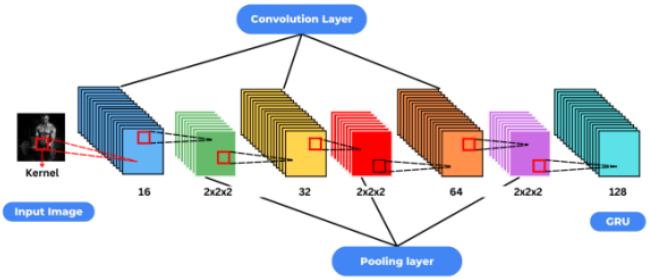

Fig. 2 Architecture of 3D CNN combined with GRU

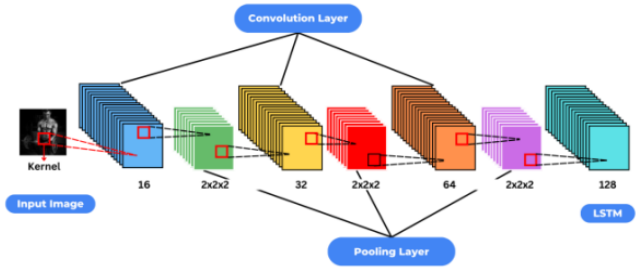

Fig. 3 Architecture of 3D CNN combined with LSTM

3.4. Exploring kernel temporal depth

In the UCF 50 dataset, we trained the aforementioned networks on split train 1 and evaluated their performance on the UCF 50 test partition. We observed that in the case of homogeneous networks, the one with depth 3 outperformed the others. However, depth 1 was significantly worse, which we attributed to its inability to model motion. For variable temporal depth networks, depth 3 performed the best, although the difference in performance compared to other depths was smaller than in the homogeneous case. The results are presented in a graph, with the left panel showing the performance of homogeneous networks and the right panel showing that of variable temporal depth networks. The researchers experimented with using larger spatial receptive fields and full input resolution for their ConvNet 3D architecture, but found that using a 3x3x3 kernel was the optimal choice according to their experiments. They also found that 3D ConvNets consistently outperformed 2D ConvNets in video classification. The researchers verified that their ConvNet 3D architecture consistently outperformed ConvNet 2D on their internal large-scale I380K dataset.

3.5 Spatial and temporal feature learning

Model Architecture: Based on our previous results, we have found that a homogeneous setup with 3x3x3 kernel size is the ideal choice for the 3D convolutional networks, which is consistent with similar findings in 2D ConvNets. For larger datasets, 3D ConvNet with 3x3x3 kernel size can be trained as deep as possible given hardware limitations. To implement this, we have designed 3D ConvNet with 3 convolutional layers, 3 pooling layers, a fully connected layer with 4 output units, and a softmax output layer. We will refer to this model as 3D CNN

for simplicity. All 3D convolutional filters are 3x3x3, and all pooling layers are 2x2x2.

Layer (type)	Output Shape	Param #
conv3d (Conv3D)	(None, 20, 100, 100, 16)	1312
max_pooling3d (MaxPooling3D)	(None, 10, 50, 50, 16)	0
dropout (Dropout)	(None, 10, 50, 50, 16)	0
conv3d_1 (Conv3D)	(None, 10, 50, 50, 32)	13856
max_pooling3d_1 (MaxPooling3D)	(None, 5, 25, 25, 32)	0
conv3d_2 (Conv3D)	(None, 5, 25, 25, 64)	55360
max_pooling3d_2 (MaxPooling3D)	(None, 2, 12, 12, 64)	0
reshape (Reshape)	(None, 2, 9216)	0
gru (GRU)	(None, 128)	3588864
dense (Dense)	(None, 4)	516

Fig. 4 Network Architecture

3D CNN model is trained on the UCF50 dataset. This comprises of 6,132 videos, with 50 different categories. Training: To handle the videos in the UCF50 dataset, we extracted 20 frames in a linear manner from each 5-second video clip, and resized them to a frame size of 64x64. During the training process, they used random cropping to generate 20x64x64 crops for spatial and temporal interleaving. We trained our 3D CNN networks with Gated Recurrent Unit (GRU) with a batch size of 4. The learning rate was at 0.001. The optimization process was stopped after roughly 13 epochs. The researchers used the deconvolution method discussed in a previous study to gain insight into the internal processes of the 3D CNN. They found that the network initially focuses on appearance in the first frames and then tracks important movements in the subsequent frames. By visualizing the deconvolution of two feature maps, conv5b with the highest activation, projected into the image space, they showed that the network selectively focuses on both motion and appearance, distinguishing it from standard 2D ConvNets. The team has provided more visualizations in the supplementary material to help readers better understand the process.

4. Experimental Results

4.1. Action Classifications

The collection includes 6,618 films from 50 human activity categories. This dataset's three segments are used. Classification model: GRU extracts temporal information from 3D CNN spatial data to train the model. We tested numerous 3D CNN combinations to discover the best one for spatiotemporal feature learning.

Baselines: Current greatest craft features are compared to 3D CNN characteristics. We evaluated the 2D CNN model with several others to test their accuracy, including LSTM 32,

GRU 32, LSTM+GRU, and 3D CNN, which is superior. That's why we employed a mix of 3D CNN models with LSTM 128, GRU 128, GRU 32, LSTM 32, LSTM 32 + GRU 32, GRU 128 + LSTM 128, LSTM 128 + GRU 128. We may compare their accuracy to choose the optimal spatio-temporal model. Results: We started with 2D CNN model combination. We reported the combined accuracy of all models in Table 1. 2D CNN with LSTM 32 offers 76.60% accuracy, 2D CNN with GRU 32 85.64% accuracy, and 2D CNN with LSTM & GRU 78.72% accuracy. According to the accuracy comparison, 2D CNN model outperforms GRU 32 model. 2D CNN model is worse than 3D. 3D CNN alone has 86.17% accuracy, while Table 2 shows the combined accuracy of all models. 3D CNN with LSTM 128 yields 89.36% accuracy, 3D CNN with GRU 128 gives 89.89% accuracy, 3D CNN with GRU 32 gives 88.83% accuracy, and 3D CNN with LSTM 32 gives 88.21% accuracy. Our analysis shows that 3D CNN model outperforms GRU 128 model with 89.89% accuracy. The 3D CNN model using LSTM 32 and GRU 32 achieves 23.64% accuracy. When paired with GRU 32 and LSTM 32, accuracy is 73.94%. Table 3 shows all model combination accuracies. Comparing accuracy rates, 3D CNN model outperforms GRU 32 + LSTM 32 model with 73.94%.

We have combined 3D CNN model with LSTM 128 and GRU 128 yields 23.64% accuracy. When paired with GRU 128 and LSTM 128, accuracy is 87.77%. In Table 4, we show the combined accuracy of all models. By comparing accuracy, 3D CNN model outperforms GRU 128 + LSTM 128 model with 87.77% accuracy.

Table 1 Model with Epochs

Model	Epochs
LSTM 32	76.60%
GRU 32	85.64%
LSTM + GRU	78.72%

Fig. 5 Accuracy Vs Loss

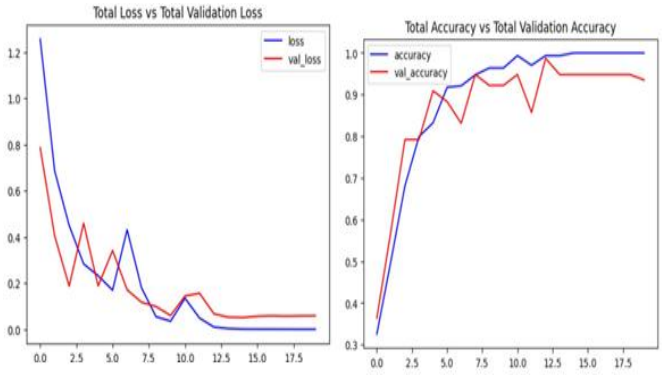

Fig. 6 Total Accuracy Vs Total Validation Accuracy

Table 2 3D CNN+ 20 Epochs

Model	
3D CNN+	20 Epochs
Taken only 3D CNN	86.17%
LSTM 128	89.36%
GRU 128	89.89%
GRU 32	88.83%
LSTM 32	88.21%
3D CNN+	20 Epochs
LSTM 32 + GRU	23.64%
GRU 32 + LSTM 32	61.17%
2D CNN+	20 Epochs
GRU128+LSTM128	87.77%
LSTM128+GRU128	23.64%

4.2. Scene and Object Recognition

We performed two assessments of the 3D CNN network to enhance situational awareness, utilising the YUPENN dataset, which comprises 420 films from 14 stage groups, and the Maryland dataset, which includes 130 movies from 13 stage groups. Furthermore, we evaluated the network's capacity to identify 42 commonplace items inside an egocentric dataset, whereby all films were recorded from a first-person viewpoint and exhibited unique visual and motion attributes in contrast to instructional videos.

Classification model: In assessing the efficacy of 3D CNN across two distinct datasets, we employed identical techniques for feature extraction and temporal feature extraction utilising Long Short-Term Memory and Gated Recurrent Unit. We adhered to the assessment procedure outlined by the dataset authors and employed a framework-based evaluation model for data products. We employed a 20-frame window for feature extraction via 3D CNN on each video, designating the most prevalent tag in each segment as the corresponding truth text. We eliminated any clip with a tag frequency below 10 frames, categorizing it as a deficient clip devoid of objects throughout training and testing.

The authors provide the outcomes of their 3D CNN model and juxtapose them with the

highest-performing algorithms for location categorization. Their 3D CNN model surpasses the prior state-of-the-art technique, which employed distinct approaches for spatial and temporal learning. In contrast, their 3D CNN model employs spatial learning in conjunction with GRU for temporal learning. Their 3D CNN model is compared to baseline models utilizing Imagenet characteristics, demonstrating comparable performance to the Imagenet-based model from the University of Maryland, however somewhat inferior to the 3D CNN model from YUPENN.

4.3. Runtime Analysis

We employed many models with diverse combinations of 2D CNN, 3D CNN, GRU (Gated Recurrent Unit), and LSTM (Long Short-Term RAM), utilizing a Tesla K80 GPU with 12GB DDR5 RAM and 2496 CUDA cores. We assessed our runtime on the UCF50 dataset, which has 50 action categories, each containing between 25 and 50 clips, with an average duration of 5 seconds each clip. The runtime for the 2D CNN with LSTM was measured at 220.54 seconds.

During the evaluation of the 3D CNN in isolation, we recorded a runtime of 26.412 seconds and an accuracy of 86.17%. Utilizing a 3D CNN in conjunction with LSTM, we achieved a runtime of 33.75 seconds and an accuracy of 88.83%, with a processing time of 0.0761 seconds per picture. The 3D CNN for spatial learning and GRU for temporal learning yielded a runtime of 32.14 seconds, an accuracy of 89.89%, and a processing time of 0.0719 seconds per image. The integration of 3D CNN and GRU demonstrates superior performance compared to both 3D CNN and the combination of 3D CNN and LSTM, in terms of time and accuracy metrics.

The method employed to calculate the processing time for each picture is t. The total number of frames in the video is T. The total time necessary for processing the entire video is f. Consequently, T = T/F.

5. Conclusion

This research aims to address the challenge of acquiring spatio-temporal features of videos through the application of 3D Convolutional Neural Networks trained on extensive video datasets. The study focuses on visualizing the long-term physical attributes of 3D ConvNets. We show that 3D CNN can model with LSTM and GRU and can outperform in models that use 3D CNN for learning both spatial and temporal features. For the spatial learning, 3D CNN is better and for the temporal learning GRU outperformed from all the models. We can see that 3D CNN features with spatial feature can now perform better on different video metrics. Finally, the 3D CNN with GRU model is more efficient, faster, and gives higher accuracy. Therefore, we can use our model in real time as it can process a frame in 0.07 seconds. The integration of 3D CNN and GRU in video analysis is gaining popularity due to its ability to capture spatiotemporal features and model temporal movement. However, challenges include limited training data and high computational costs. Researchers are exploring new techniques to improve performance and develop advanced architectures for better capture of spatiotemporal and temporal action. Future research will focus on expanding efficient training and inference techniques, such as knowledge distillation and model compression.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Statement The Authors consciously assure that for the manuscript is the authors' own original work, which has not been previously published elsewhere. The paper is not currently being considered for publication elsewhere. The paper reflects the authors' own research and analysis in a truthful and complete manner. The paper properly credits the meaningful contributions of co-authors and co-researchers. The results are appropriately placed in the context of prior and existing research. All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference. The violation of the Ethical Statement rules may result in severe consequence.

References

[1] Cheng C , Xu H (2024) A 3D motion image recognition model based on 3D CNN-GRU model and attention mechanism. Image and Vision Computing, 146, 104991. https://doi.org/10.1016/j.imavis.2024.104991

[2] Wang Y, Shen XJ, Chen HP, Sun JX (2021) Action Recognition in Videos with Spatio-Temporal Fusion 3D Convolutional Neural Networks. Pattern Recognition and Image Analysis, 31(3), 580–587. https://doi.org/10.1134/s105466182103024x

[3] Hosseini MS, Ghaderi F (2020) A Hybrid Deep Learning Architecture Using 3D CNNs and GRUs for Human Action Recognition. International Journal of Engineering. Transactions C: Aspects, 33(6). https://doi.org/10.5829/ije.2020.33.05b.29

[4] Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T (2016) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), 677–691. https://doi.org/10.1109/tpami.2016.2599174

[5] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2013) DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1310.1531

[6] Jain A, Tompson J, Andriluka M, Taylor GW, Bregler C (2013) Learning Human Pose Estimation Features with Convolutional Networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1312.7302

[7] Ji S, Xu W, Yang M, Yu K (2012) 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 221–231. https://doi.org/10.1109/tpami.2012.59

[8] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1408.5093

[9] Lan Z, Lin M, Li X, Hauptmann AG, Raj B (2014) Beyond Gaussian Pyramid: Multi-skip Feature Stacking for Action Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1411.6660

[10] Peng X, Qiao Y, Peng Q, Wang Q (2014) Large Margin Dimensionality Reduction for Action Similarity Labeling. IEEE Signal Processing Letters, 21(8), 1022–1025. https://doi.org/10.1109/lsp.2014.2320530

[11] Li Y, Yin R, Kim Y, PandaP (2023) Efficient human activity recognition with spatio-temporal

spiking neural networks. Frontiers in Neuroscience, 17. https://doi.org/10.3389/fnins.2023.1233037

[12] Nafea O, Abdul W, Muhammad G, Alsulaiman M. (2021) Sensor-Based Human Activity Recognition with Spatio-Temporal Deep Learning. Sensors, 21(6), 2141. https://doi.org/10.3390/s21062141

[13] Tasnim N, Islam MK, Baek J (2021) Deep Learning Based Human Activity Recognition Using Spatio-Temporal Image Formation of Skeleton Joints. Applied Sciences, 11(6), 2675. https://doi.org/10.3390/app11062675

[14] Senthilkumar N, Manimegalai M, Karpakam S, Ashokkumar S, Premkumar M (2021) Human action recognition based on spatial–temporal relational model and LSTM-CNN framework. Materials Today Proceedings, 57, 2087–2091. https://doi.org/10.1016/j.matpr.2021.12.004

[15] SaiRamesh L, Dhanalakshmi BKS (2024) Human Activity Recognition Through Images Using a Deep Learning Approach. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-4443695/v1

[16] Uddin MA, Talukder MA, Uzzaman MS, Debnath C, Chanda M, Paul S, Islam MM, Khraisat A, Alazab A, Aryal S (2024) Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN. International Journal of Cognitive Computing in Engineering, 5, 259–268. https://doi.org/10.1016/j.ijcce.2024.06.004

[17] Varshney N, Bakariya B, Kushwaha AKS, Khare M (2022) Human activity recognition by combining external features with accelerometer sensor data using deep learning network model. Multimedia Tools and Applications, 81(24), 34633–34652. https://doi.org/10.1007/s11042-021-11313-0

[18] Varshney N, Bakariya B, Kushwaha AKS (2021) Human activity recognition using deep transfer learning of cross position sensor based on vertical distribution of data. Multimedia Tools and Applications, 81(16), 22307–22322. https://doi.org/10.1007/s11042-021-11131-4

[19] Arjaria SK, Rathore AS, Bisen D, Bhattacharyya S (2022) Performances of Machine Learning Models for Diagnosis of Alzheimer's Disease. Annals of Data Science, 11(1), 307–335. https://doi.org/10.1007/s40745-022-00452-2