A ES
ISSN: 2096-3246
Volume 54, Issue 02, December, 2022
Special Issue

MULTI-FEATURE EXTRACTION BASED HUMAN ACTION RECOGNITION USING 3D-CNN

A.Rajaram
E.G.S Pillay Engineering College, Nagapattinam, India

C.Jegadheesan
cjega1987@gmail.com
Dhanalakshmi Srinivasan College of Engineering, Coimbatore, Tamil Nadu, India

Ashwini Kumar
Babu Banarasi Das University, Lucknow (U.P.) India

Abstract – In this paper, we discuss about using computer vision to identify human behaviour in security camera footage automatically. The majority of today's approaches calculate intricate characteristics by hand to use as the basis for classifiers. As a sort of deep model, convolutional neural networks (CNNs) are capable of operating on unprocessed data. However, at the moment, such models can only process 2D data. In this study, we create an innovative 3D convolutional neural network (CNN) model for recognising actions. By using 3D convolutions, this model is able to extract features from color, temporal, geometrical, and spatial dimensions effectively capturing the moving objects included in a sequence of consecutive frames. The input frame are used to produce various channels of data, and these channel are combined to construct the final visual features using the generated model. We propose regularizing the outputs using high-level characteristics and integrating the predictions of several algorithms to even further enhance the results. Models are then put to use in the real world, namely airport surveillance films, where they outperform state-of-the-art approaches in recognising human behaviours.

Keywords: Human Action Recognition, Surveillance Videos, 3D CNN, and Multiple Feature Extraction

1. Introduction

The most common and time-tested approach to activity identification is to set up security cameras around the building and observe people going about their daily routines. Both human (a person watching the videos and images coming from the cameras) and automated processes can be used for monitoring. Computer vision techniques are developed to automatically recognise activities based on the data (videos and images) captured by a camera.

In designed to safeguard against intruders by telling the difference between legitimate and illegitimate activity, today's smart burglar alert system and Human Activity Recognition systems comprise of network systems of advanced automated sensors and devices linked to a centralised control unit. The intellectual model employs machine learning approach and IoT (Internet of Things) sensor infrastructural facilities to detect and identify complex movement patterns, as opposed to traditional security devices that respond to a single sensor trigger. Motion detectors, ultrasonic

motion detectors, and passive infrared (PIR) sensors are just a few examples of the many types of sensors used to track human motion. Human Activity Recognition is similar in that it makes use of the same sensors and relies on deep data analysis and machine learning methodologies to characterise the motion based on the application needed. This idea can by implemented in many practical applications, including the early diagnosis of human ageing symptoms with dementia, the definition of oddities in employee transformation or motions within an indoor space, and even as part of burglary alarm systems that detect anomalies.

With the advent of deep learning techniques like convolutional neural networks for analysing video content, the field is making strides toward recognising and predicting complex human activities involving a wide range of people, objects, and events. This has resulted in the emergence of a number of important new areas for research and development, such as (I) developing methods for the reliable spatial-transient constraint of exercises, (ii) displaying the full complexity and significance chain of exercises from beginning to end, (iii) recognising and estimating the actions of groups of people, and (iv) expanding the size and scope of existing datasets and convolutional models. Recognizing human actions is an intriguing problem with many potential solutions.

Over the past few years, numerous strategies have been put into place. There are a number of approaches to identifying motion, so it's possible to tell whether someone is attempting to run, having to walk, trying to dance, jogging, or falling, for example. Data is collected in a wide variety of ways, from sensors to accelerometers to gyroscopes to cameras, and so on, and then analysed using a number of different methodologies and datasets. Comparisons are made between the intended results produced by the various methods and data sources. Hidden Markov models, K-nearest neighbours, decision trees, support vector machines, and unattended Deep Learning architectures are just some of the popular methods of machine learning used for categorization. Machines are becoming better at resolving some very complex problems (like recognising a picture) as a result of advancements in computer vision. Models are being managed to make wherein, if a photo is given to the model,it would then identify the action that is evident in the model. Deep learning is a branch of machine learning. Deep learning can be classified into two groups deep supervised learning as well as deep unsupervised learning. Human activity recolonization is a complicated situation. This issue can be fixed using a fusion supervised deep convolutional neural network and a deep unsupervised learning model. An exciting implementation of this issue could be trying to identify artefacts in video content in real-time.

Importance of Human Activity Recognition

The current challenge in human activity identification is discussed here. The future of this line of inquiry is addressed here, broken down into four bullet points.

- Safeguarding the public against opportunistic robbers in public spaces. Maintaining constant vigilance in public spaces is a real challenge. As a result, automated video surveillance is needed to gradually filter human activity and categorise it as typical or out-of-the-ordinary.
- Surveillance Cameras for Automatic Teller Machines: Money Vending Machine Safety There is already a distinct market for protecting ATMs. ATM withdrawals are fast and simple, but criminals may easily take advantage of this convenience if the machines and the areas surrounding them are not secure. HAR can be used to categorise any undesirable activity that

occurred in the ATM and transfer information to the police people for this reason still many researchers are designing a model to tackle these challenges.

- Staff performance reviews benefit from the availability of video footage of the workers in question. We need to create a deep learning algorithm that can tell whether employees are really working or just wasting time. Because of the number of persons engaged in this issue, this system faces several obstacles, and the video quality may not be suitable for usage in the workplace. The Medical Treatment of the Elderly Innovations in the healthcare sector have had a profound effect on our level of contentment, and this has propelled the future forward. This factor alone has contributed to a rise in the elderly population. One factor in the rising price of medical treatment is the growing number of elderly individuals who need to be cared for. Due to the ever-increasing cost of providing medical treatment, many hospitals, clinics, and charitable organisations are exploring cost-cutting alternatives. Human activity recognition must be improved because of the various real-world difficulties that older persons experience.

2. Related Work

When it comes to handling data, convolutional neural networks (CNNs) excel at grid frameworks but struggle with dynamic skeletons, which are often described as graphs. For the purpose of feeding such data into CNNs, the authors of this paper presented a skeleton-based square grid (SSG) for converting dynamic skeletons into 3D grid-structured data. Each SSG is made up of a JSG and an RSG, the former of which is based on the intrinsic interdependence of the body's distinct sections, and the latter on their extrinsic dependencies. Next, a two-stream 3D CNN is built to train spatiotemporal features utilising the JSG and RSG sequences, which improves the capacity of feature representation to detect the relationships among 3D grid-structured data. At last, the authors presented a soft attentiveness model that zeroes down on just the relevant skeletal features in the sequences. We used three datasets (NTU RGB+D, Kinetics Motion, and SBU Kinect Interaction) to test and verify the accuracy of our proposed model for action recognition. The empirical findings validate the efficacy of the suggested strategy and show that it outperforms the current gold standard in the field [11]. Humans' ability to recognise actions relies heavily on spatial and temporal data. In this task, researchers propose a new deep learning architecture called a recurrent 3D convolutional neural network (R3D) to capture long-range temporal dependencies by grouping the entries from a 3D convolutional network to be utilized as input to a Long Short-Term Memory (LSTM) architecture, which is then used to retrieve efficient and discriminatory spatial characteristics for use in action recognition. Extracting temporal information is facilitated by the 3D convolutional network and LSTM. In order to capture short-term spatial-temporal characteristics into the LSTM, the suggested R3D network combined the two approaches by using a shared 3D convolutional network in sliding windows during video streaming. Long short-term memory (LSTM) encodes spatial-temporal information representative of a high-level abstraction of human behaviours in its output characteristics. Traditional, state-of-the-art, and deep learning algorithms are contrasted with the suggested algorithm. Experiment findings showed that the suggested approach works well as a smart monitoring tool for distant healthcare [12].

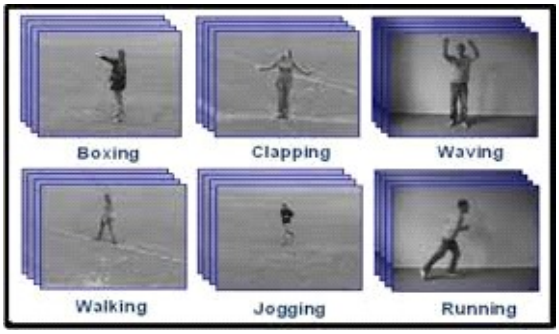
For video-based human action detection from trim and untrimmed films, the authors provide a simple but successful network that incorporates an unique Discriminative Feature Pooling (DFP)

mechanism and a novel Video Segment Attention Model (VSAM). Our VSAM ensembles the most important features from all the video content and understands (1) class-specific attention weights to categorise the videos into the corresponding action categories, while the DFP module introduces a selective attention pooling method for 3D Convolutional Neural Networks, which attentionally pools 3D fully convolutional maps to emphasise the most critical spatial, time and space, and channel-wise features towards the actions inside a video segment. The action recognition networks may be trained either completely supervised or weakly supervised, using trimmed or untrimmed movies, respectively. Without the need for accurate temporal annotations of action occurrences in movies, our network is able to acquire attention weights for untrimmed videos with weak labels. When compared to the current gold standard technique [13], our network gets encouraging results when tested on the untrimmed video datasets of THUMOS14 and ActivityNet1.2, as well as the trimmed video datasets of HMDB51, UCF101, and HOLLYWOOD2. In [14], the authors presented a better spatiotemporal attention model for action recognition based on a two-stream structure. To be more specific, we begin by isolating the spatial information present inside each frame as well as the optical flow patterns present between frames in each film. We next put into practise a powerful attention module that, in turn, successively infers focus maps across channel, geographic, and temporal domains. Researchers next apply a temporal pooling approach to compress the dynamic nature after performing adaptive feature refinement using the attention maps. Next, the spatial LSTM and the temporal LSTM get the accomplished spatial and temporal information. After that, we combine the spatial feature, the temporal feature, and the two-stream fusion feature to label the activities in movies. For the sake of future human-robot interaction tasks, we additionally gather and create a fresh Ping-Pong action dataset from YouTube. There are 2400 videos split into 4 categories with labels. On the Ping-Pong action dataset and the HMDB51 dataset [14], we conduct a comprehensive analysis of the proposed technique and compare it to existing action detection algorithms to prove its viability and efficacy. Significant advancements have been achieved in video-based human activity detection with the use of convolutional neural networks (CNNs). Convolutional neural network (CNN) characteristics that are both spatial and channel-wise may provide substantial data for accurate picture description. However, CNNs cannot effectively zero in on the informative motion regions of actions, and they cannot analyse the long-term temporal dependence of a whole movie. In this study, we offer a new framework for action detection based on video data in an effort to solve these two issues. First, we use dynamic image sequences (DISs) to characterize films, which effectively characterise videos by modelling their local texture dynamics and dependencies. To zero in on the most informative spatial motion areas of human activities and the networks' most discriminative channels, a CNN-based channel and spatial-temporal interest points (STIPs) attention model (CSAM) is presented. To be more specific, channel attention (CA) is accomplished by automatic identifying channel-wise fully convolutional and assign specific values for distinct channels. In order to encode STIPs attention (SA), the frames of dynamic visual sequence with identified STIPs are projected into the domain of the relevant convolutional feature map. To provide efficient feature representations for movies, the suggested CSAM is inserted following CNN convolutional layers to enhance the feature maps, then global average pooling is used. Finally, an LSTM is used to extract temporal dependencies and produce classifications based on frame-level

video representation. Our strategy beats the state-of-the-art depth-only methods [15] in experimental results on three demanding RGB-D datasets.

3. Proposed Work

Due to the exponential growth in the amount of training examples with increasing size of input window, the input to 3D CNN models are restricted to a small number of consecutive video frames. However, many human acts take place over many frames. Encoding high-level motion information into the 3D CNN models is therefore desirable. In order to regularise the 3D CNN models, we suggest calculating motion characteristics from a large number of frames and utilising these features as auxiliary outputs. Different 3D CNN architectures may be developed based on the 3D convolution processes. The suggested architecture outperforms the others evaluated in this research when applied to the YouTube dataset. This may or may not hold true for different data sets, however. Choosing the best architecture for a given task is difficult since it is context-dependent. Alternately, one might build many models and use a combination of their forecasts in making predictions. In addition to modern neural network combinations, this method has been used to combine conventional neural networks. However, research on the impact of model combination in the context of convolutional neural networks for activity recognition is lacking. In this work, we suggest building a number of 3D CNN models, each with a slightly different design to capture potentially complimentary information from the inputs. During the prediction stage, all of the models are fed the same data and their results are averaged. The experimental findings show that the 3D CNN models' effectiveness is much improved by using this model combination scheme while tackling activity recognition tasks.



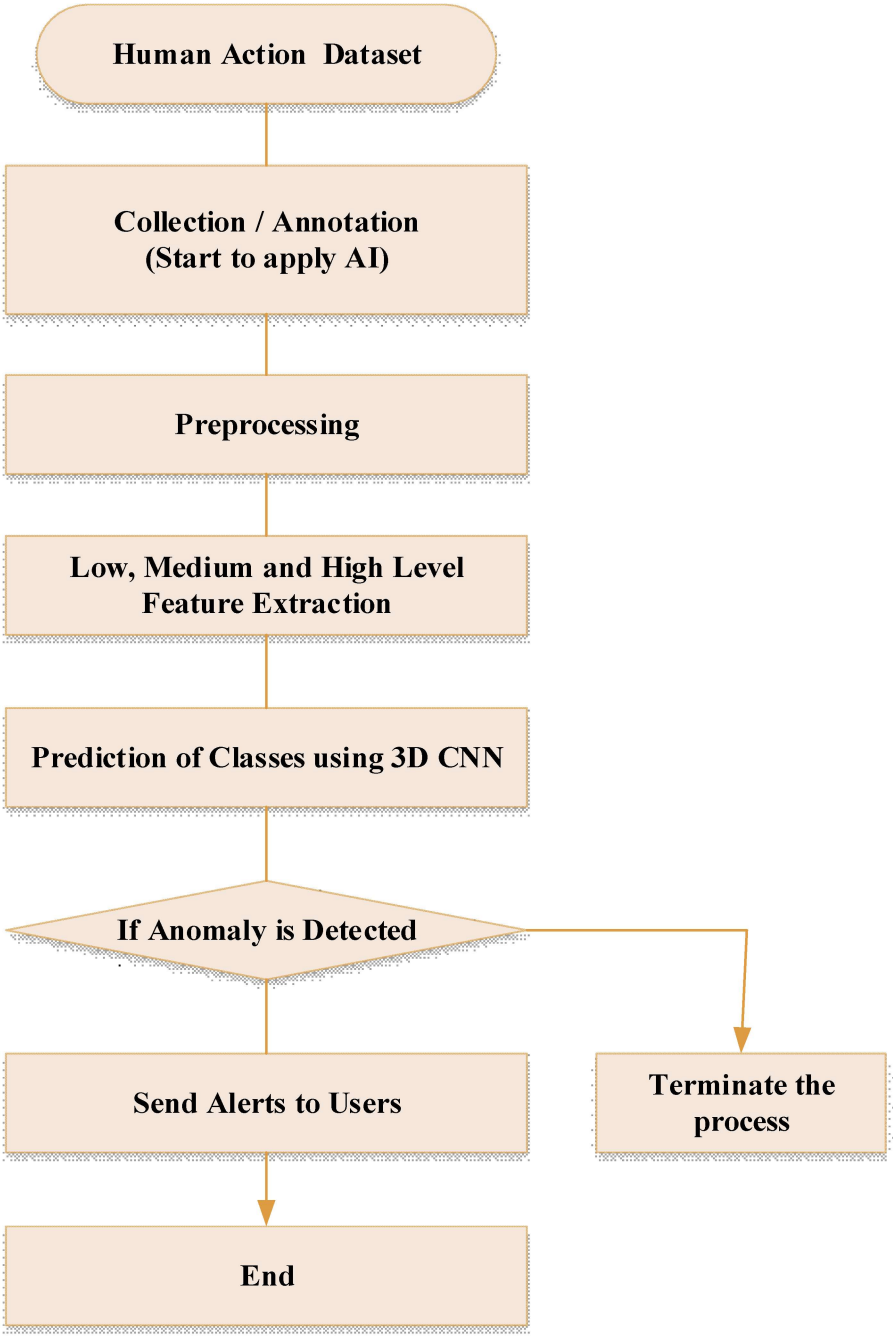


Fig. 1. System Architecture

We employed five different deep learning models (CNN, LSTM, RNN, and GRU) and compared the classification performance to different burglary techniques. There was support for three-dimensional data in classification models; this included batch size, features, and time steps. The models were trained using data from videos of varying durations that were masked, allowing them to better understand and assist actions in the actual world. Previous work on time series issues claimed that this might be modelled as a regression, thus we tested the models using both the binary cross entropy and mean squared error loss functions. When studying sentiment analysis, voice recognition, or behaviour recognition, LSTM is often used because of its effectiveness in dealing with long-term dependencies. This study challenge necessitates the persistence of earlier frames in order to comprehend the whole event, since the sequence may include overlapping behaviours. Due to the overlap among both normal and burglar sequences caused by the sub-action of opening the door using the grasp, a recurrent type of neural network is chosen to be the most ideal algorithm for this study.

To get rid of the padding that was included to make the sequences even, the first layer of the 6-layer sequence LSTM model serves as a masking layer. Since the face and foot key joints were eliminated, the masked value was set to -1, and the input number of features is 45. Then the 3D input shape was preserved by an LSTM layer, which was followed by a Dense layer. Then, a second LSTM layer was employed after a dropout layer with a rate of 0.5 was implemented. Finally, using the Adam optimizer's 0.0001 learning rate and mean squared error nonlinear function, a dense layer consisting of a single neuron and the sigmoid transfer function as the network output was constructed. Experiments were run with varying values for hyper parameters, dropout rates, and the number of neurons, and the best possible classification results were achieved.

The second kind of classification algorithm was a simple RNN network with the ability to learn consecutive connections. The GRU model was created with two GRU layers, following by one with a ReLU activation function, and a drop out layer with a frequency of 0.2 was inserted between the first and second GRU layers. This classification effort made use of a fully connected network with an activated softmax function. The training was carried out using an Adam optimizer and a training rate of 0.001.

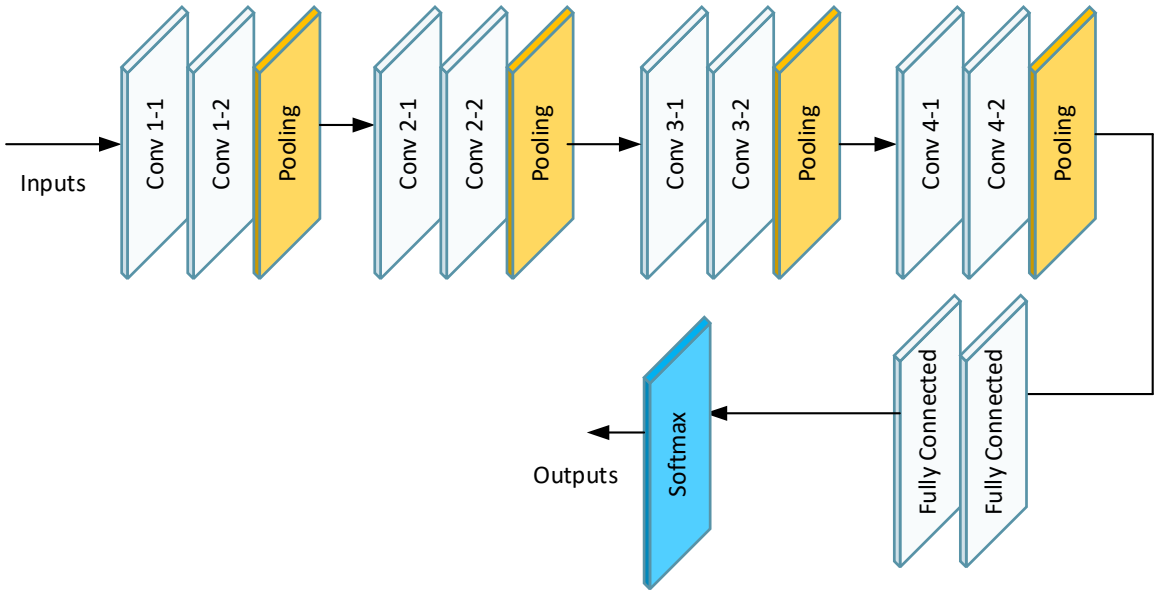


Fig.2.3D CNN for Video Action Recognition of Humans

One subset of deep neural networks is the convolutional neural network (CNN) which is often used to examine pictures. Here, instead of a vector as is typical in neural networks, we get a multi-channel picture as the input. A variant of multilayer perceptrons used in CNNs is optimised to function with little preprocessing. Typical CNNs can only identify the kind of items, not their precise location. Actually, it is feasible to regress bounding boxes straight from a CNN, however this can only be done for a single item at a time. If there are a number of objects in view, the interference between them will prevent the CNN bounding box regression from functioning well. The Convolutional Neural Network (CNN) is a subset of the larger ANN (Artificial Neural Network) family, often used for visual data challenges. Images and movies are the only kind of visual data accepted by computers. Layered CNNs are possible. There is a connection between every neuron in each successive layer. The output of one layer is used as the input for the next layer. An activation function might be shared by several layers, including any combination of convolution, pooling, and fully connected layers. At long last, a Softmax layer may be used for categorization. Therefore, CNN may be seen as an onion with several levels. Image and video analysis, object localization, semantic segmentation, optical character recognition, etc. are only few of the many uses for CNN. CNN has numerous potential uses in the medical field. CNNs can perform object categorization, detection, and segmentation in medical imaging applications. VGGNet is one example of a popular CNN like this.

In order to recognise human actions inside videos, we employed 3D convolutional neural networks (CNNs). CNNs in this instance were trained using labelled datasets, necessitating a significant number of labelled examples. Additionally, a tracking algorithm was required to choose a sub-window inside a video sequence prior to performing the action recognition, and the action-success recognition's was contingent on the tracking method's effectiveness. 3D Convolutional Neural Networks (CNNs) are able to learn both the spatial and temporal characteristics of films.

4. Results & Discussion

In all, 240 films, each 25 seconds long and tagged as either "burglar" or "regular," were captured at 30 frames per second for this study. From these videos, a total of 228,640 frames were gathered.

The UCF Crime dataset was the only alternative that met our needs. Unfortunately, we had to create a dataset with just six topics since the great majority of movies were either too obscure or of too low a quality, and they included a mixture of both residential and business burglaries. This research took into account the five most popular methods of entry used by burglars: smashing the door with a hammer, picking the lock, bashing on the door with shoulder, kicking the door with the foot, and attempting to open it with a lever. The total number of datasets was evenly split into training and testing sets. An array of feature combinations were used for training and testing the model throughout the experiment. After analysing the dataset's characteristics, researchers found that the eyes, nose, and ears, along with the toes and balm, were hidden in the vast majority of the frames reporting 0 values. Using face traits to train the model has been ruled out since there is 0% confidence for almost 17,000 images. Previous studies that used a skeletal approach to action detection found that using raw dimensions of the joints added little value to the model. Thus, it was decided to use joint co-occurrence rather than a comprehensive comparison of attributes. To achieve this, the recurrent neural network was designed as a fully linked network, with each layer's outputs feeding into the next layer, and the inputs of the first layer representing features feeding into the inputs of the second layer, and so on.

This research was conducted in part because the present system suffers from a false alert rate of between 94% and 99.9%. The precision of the method was very variable, necessitating several iterations before settling on an optimal model outcome. A second pass through the dataset eliminated the deformed and flickering skeletons and included trimming all the films and reanalyzing them. The sequence length was capped at 350 time steps since running the model for longer than that degraded its accuracy. When compared to a binary cross entropy loss function, the findings of this study improved upon those of previous studies that used LSTMs to experiment with time series forecasting issues and found success utilising a mean squared error loss function. The experimental findings for all five models are summarised in Table 1, and the suggested LSTM model performed particularly well in the test of intruder recognition. The results of the implemented system's performance testing By carefully examining the timeframes involved, we can deduce that reading the JSON files and using the model to create predictions takes just 0.3 - 0.5 seconds, thus the issue must have been elsewhere. Further investigation revealed that this lag was brought on by the fact that MATLAB not only writes JSONS but also pictures. There was a four-minute lag because the output photos were not ready. Fig 3 to Fig 10 describes the performance of the human action recognition in terms of accuracy, precision, recall, f-score, respectively. Then the AUC curve is presented in Fig 11 to Fig 15 for five different deep learning models.

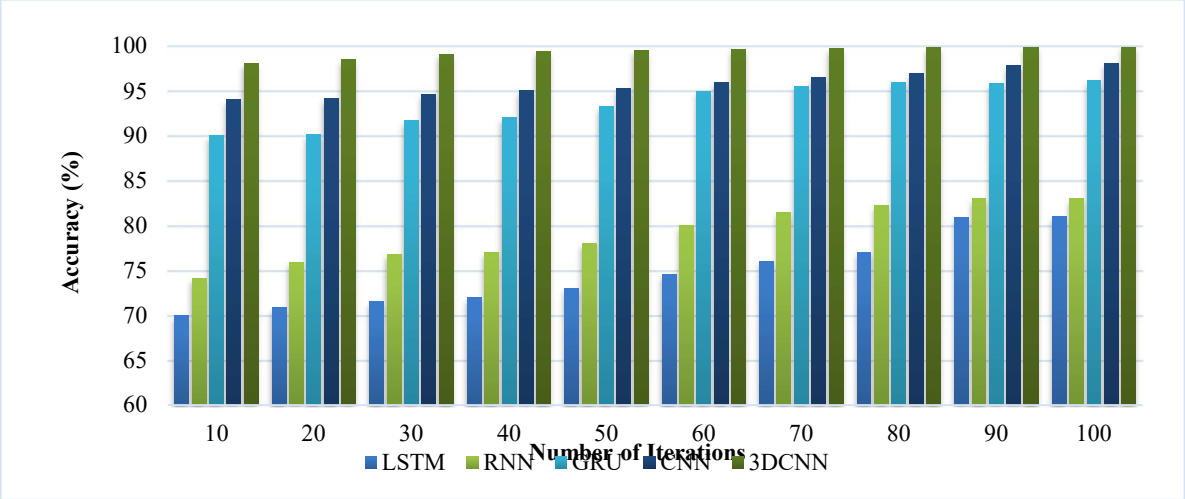


Fig.3. Performance of Accuracy vs. No of Iterations

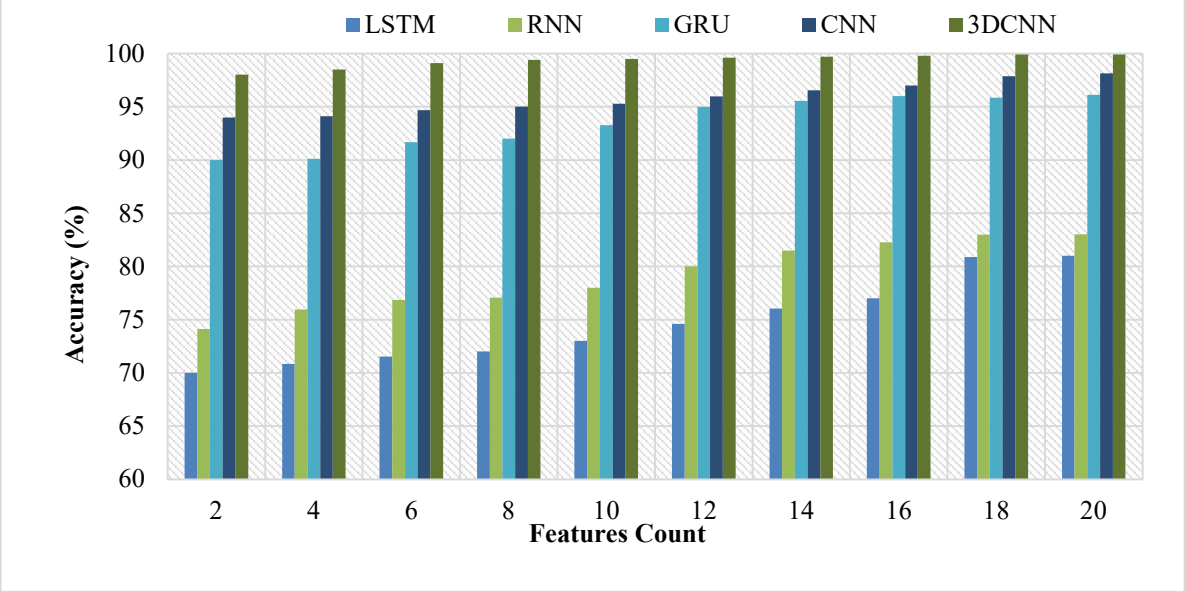


Fig.4. Performance of Accuracy vs. No of Features

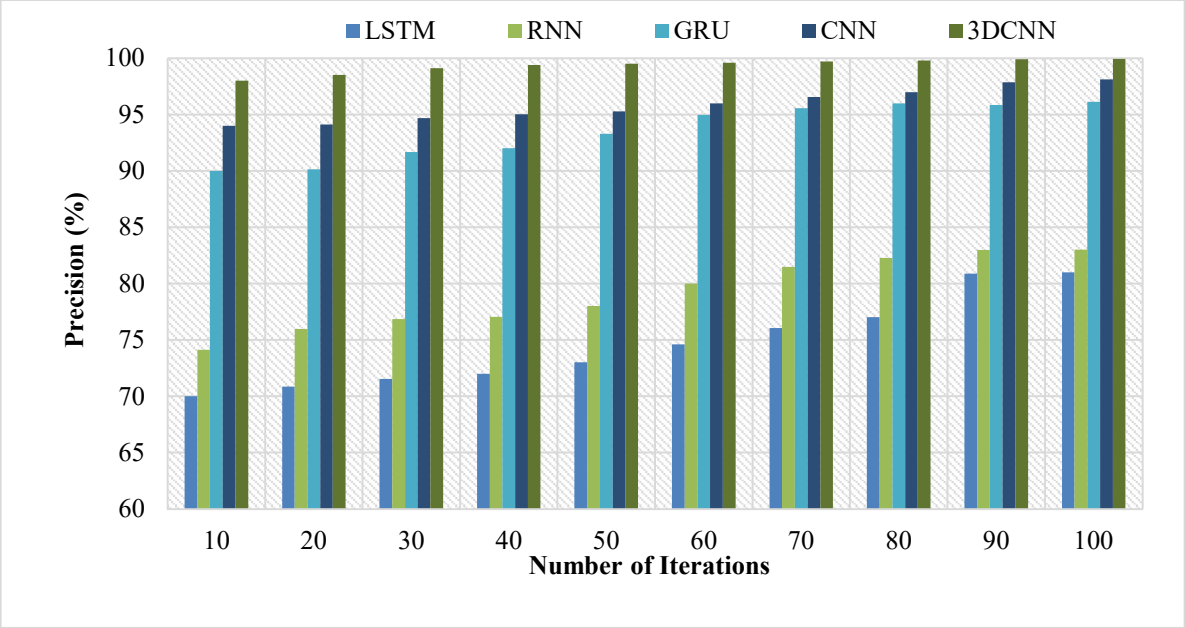


Fig.5. Performance of Precision vs. No of Iterations

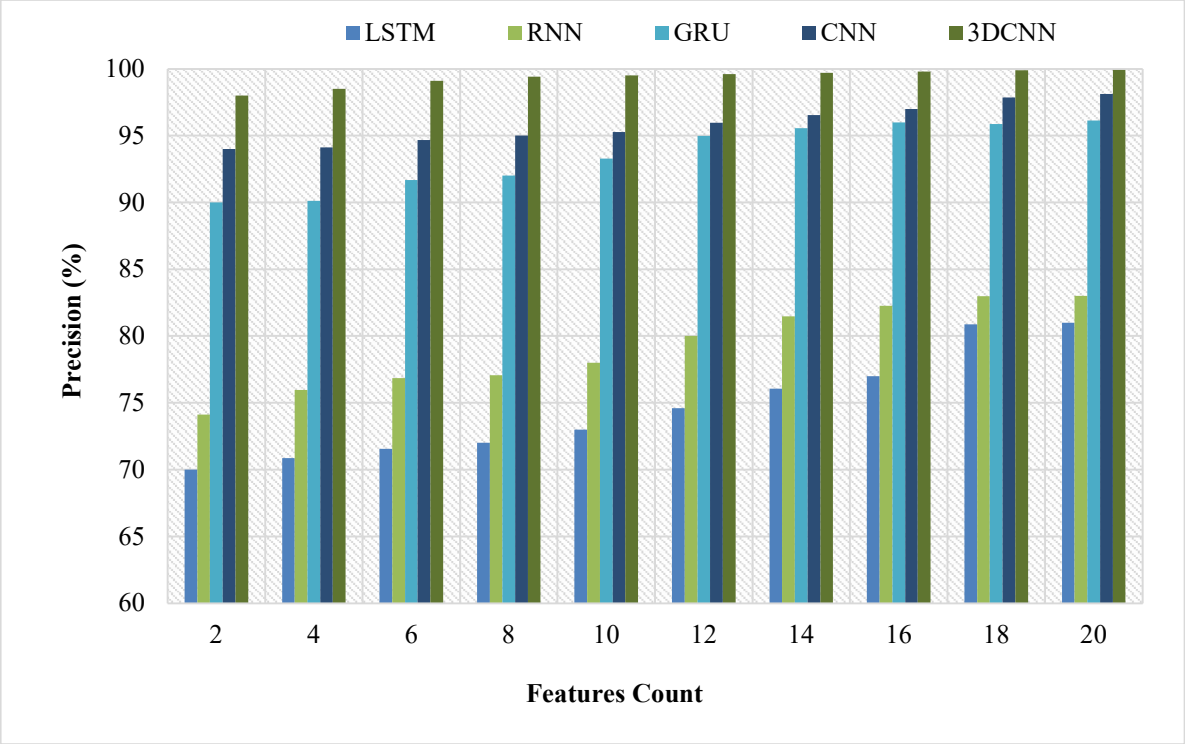


Fig.6. Performance of Precision vs. No of Features

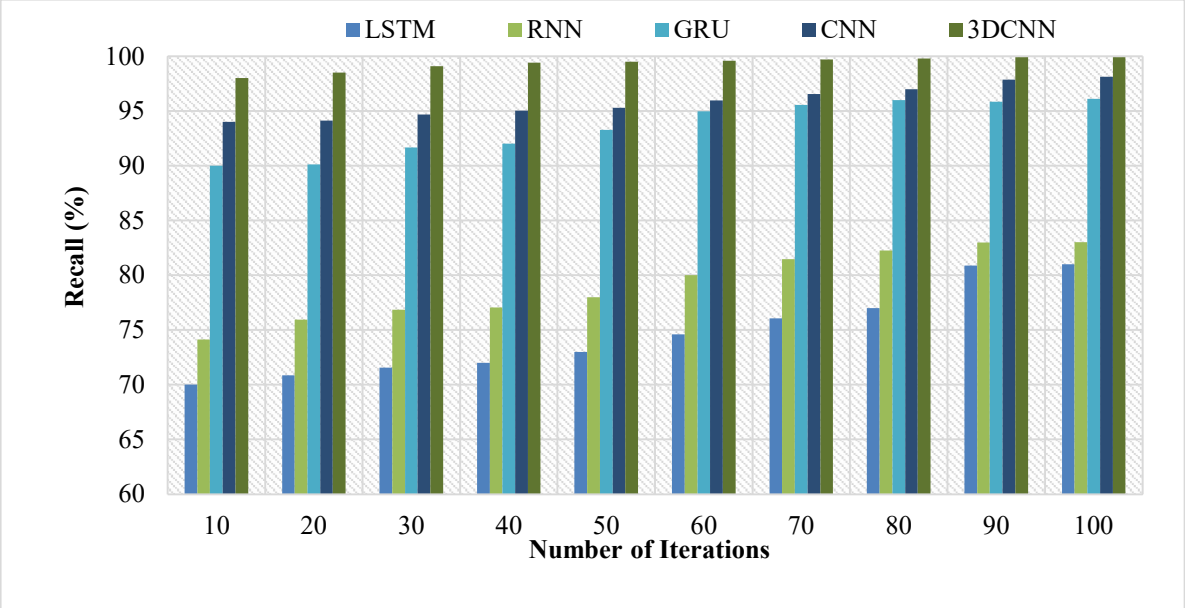


Fig.7. Performance of Precision vs. No of Iterations

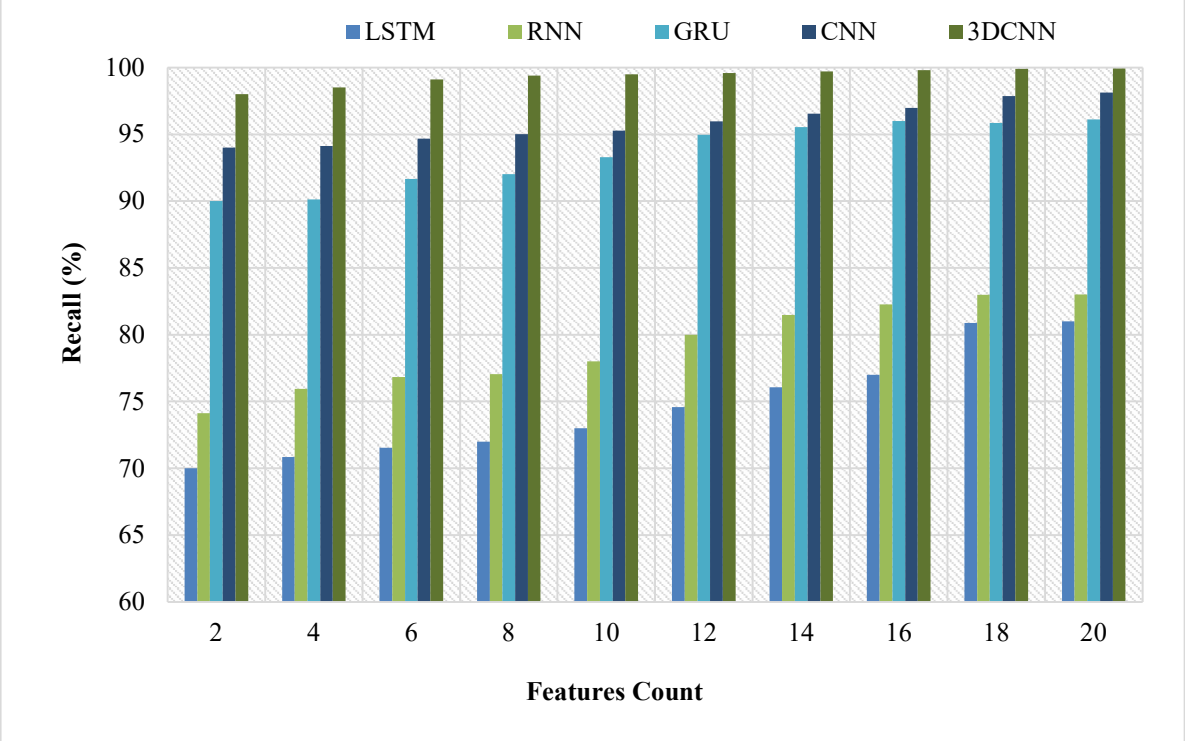


Fig.8.Performance of Precision vs. No of Features

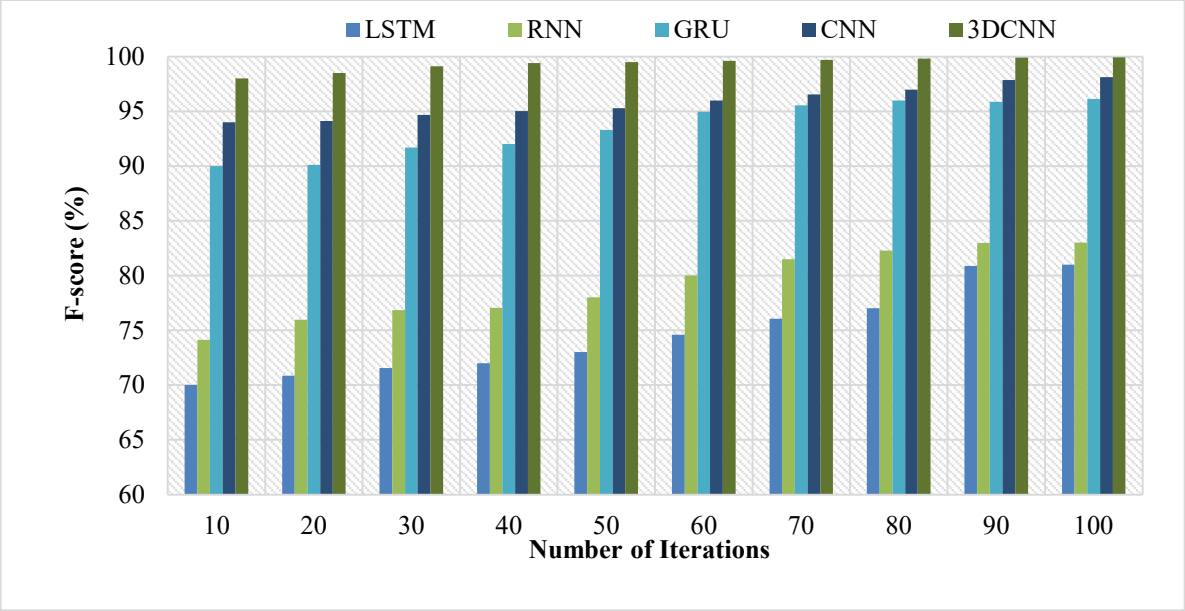


Fig.9. Performance of F-score vs. No of Iterations

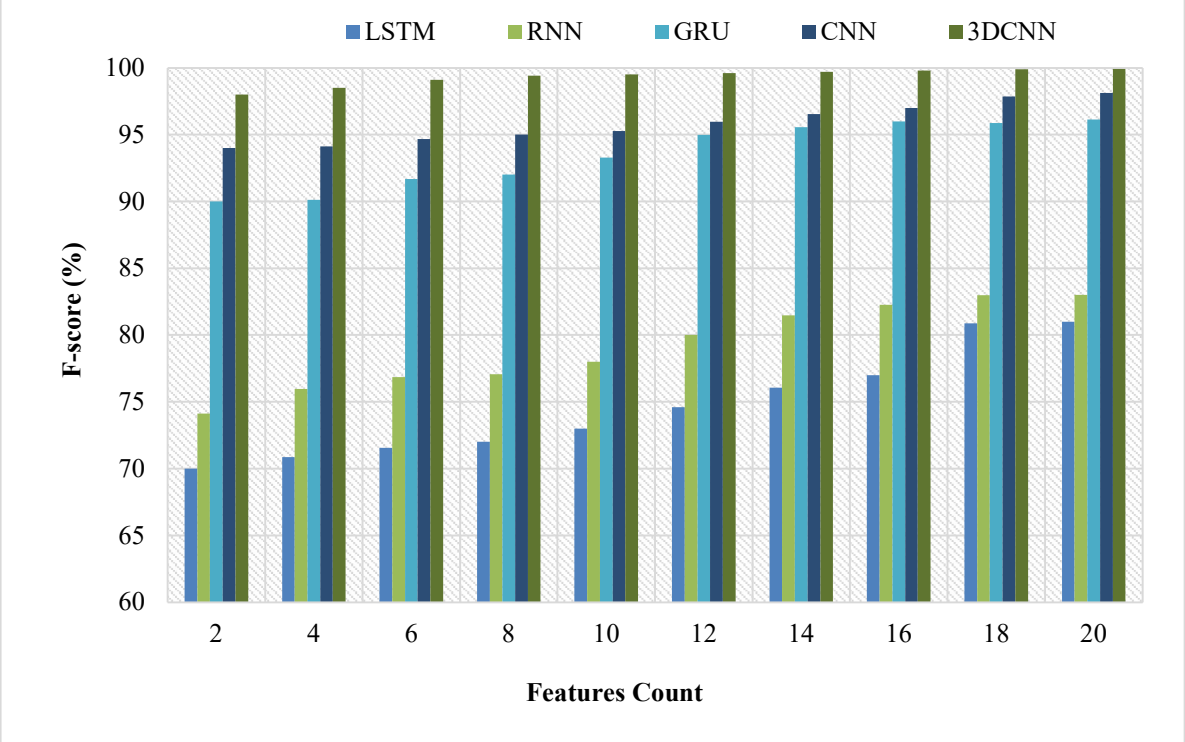


Fig.10. Performance of F-score vs. No of Features

Table.1.Performance of Human Action Recognition (Proposed vs. Existing)

Metrics	LSTM	RNN	GRU	CNN	3DCNN
Accuracy	86%	81%	84%	89%	99.4%

Precision	85%	81.5%	83.5%	88%	99.5%
Recall	84%	81%	83%	87%	99%
F-score	84.5%	81.2%	85%	86%	97%

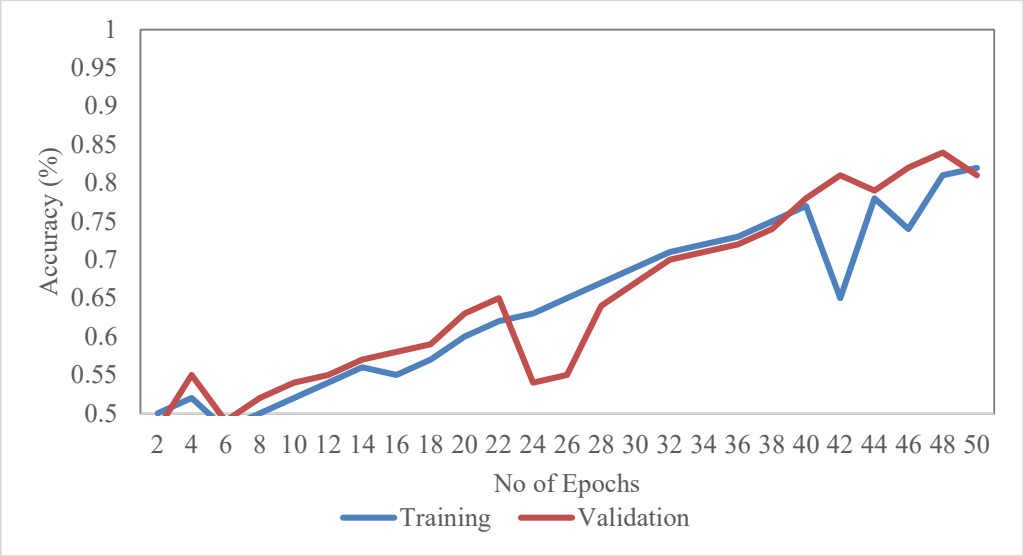


Fig.11.Comparison of LSTM Accuracy vs. No of Epochs

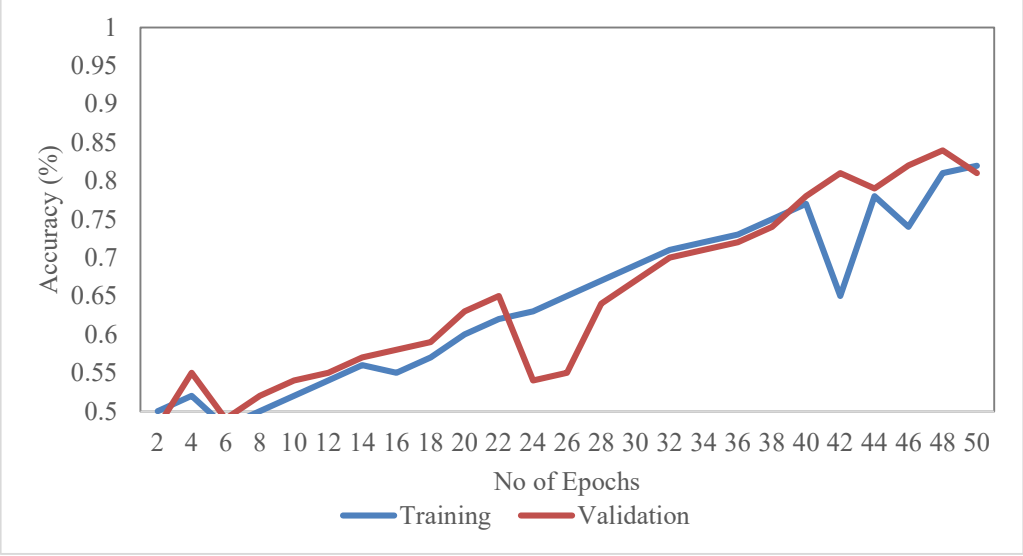


Fig.12.Comparison of GRU Accuracy vs. No of Epochs

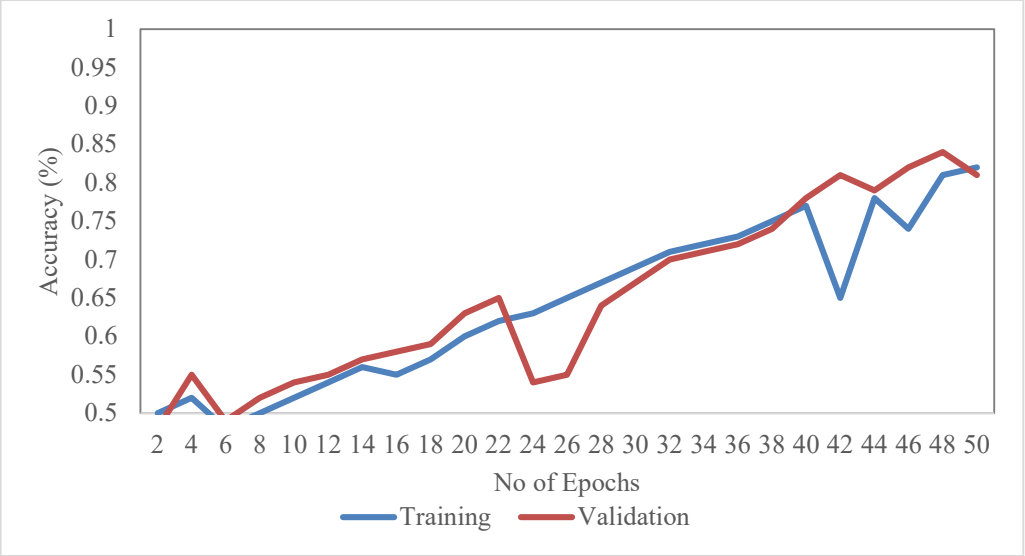


Fig.13.Comparison of RNN Accuracy vs. No of Epochs

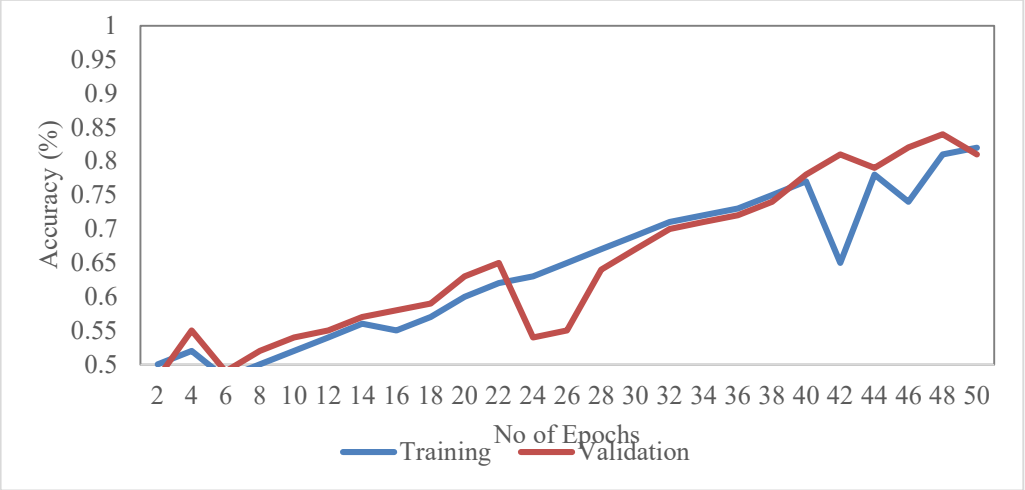


Fig.14.Comparison of CNN Accuracy vs. No of Epochs

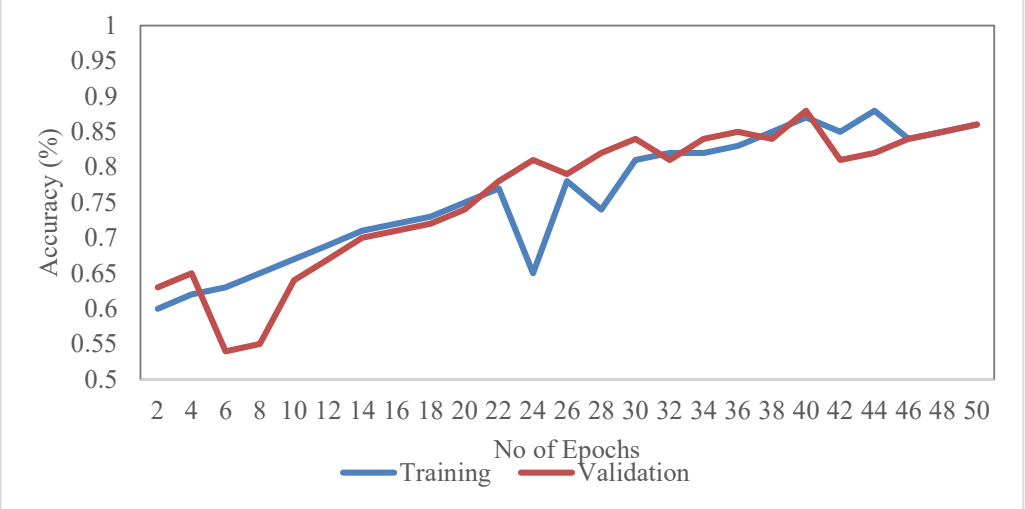


Fig.15.Comparison of 3D-CNN Accuracy vs. No of Epochs

5. Conclusion

In this work, we created 3D CNN models for performing such tasks. To create their features, these models conduct 3D convolution operation over many dimensions. The created deep architecture produces many channels of data from nearby input data, with each channel carrying out its own convolution and subsampling. Information from all channels is combined to generate the final feature representation. To further improve the model performance, we devised regularization and combination strategies for the system. The YouTube surveillance footage datasets were used to assess the 3D CNN models. The results reveal that the 3D CNN model performs better than competing approaches on video surveillance data, and reaches superior results on video data, indicating its superiority in real-world settings. In this study, we looked at using a CNN to recognize actions. In this article, a supervised technique was used to train a 3D CNN model, which necessitates a high number of labelled examples. Pretrained utilizing unsupervised techniques, this kind of model has been shown to drastically minimise the amount of labelled samples required. In the future, we want to investigate 3D CNN models that have been trained autonomously.

References

1. Tufek, N., Yalcin, M., Altintas, M., Kalaoglu, F., Li, Y., & Bahadir, S.K. (2020). Human Action Recognition Using Deep Learning Methods on Limited Sensory Data. IEEE Sensors Journal, 20, 3101-3112.
2. Berlin, S.J., & John, M. (2020). Particle swarm optimization with deep learning for human action recognition. Multimedia Tools and Applications, 79, 17349-17371.
3. Gutoski, M., Lazzaretti, A.E., & Lopes, H.S. (2020). Deep metric learning for open-set human action recognition in videos. Neural Computing and Applications, 1-14.
4. Tasnim, N., & Baek, J. (2022). Deep Learning-Based Human Action Recognition with Key-Frames Sampling Using Ranking Methods. Applied Sciences.
5. Gutoski, M., Lazzaretti, A.E., & Lopes, H.S. (2020). Deep metric learning for open-set human action recognition in videos. Neural Computing and Applications, 33, 1207 - 1220.
6. Wei, H., Jafari, R., & Kehtarnavaz, N. (2019). Fusion of Video and Inertial Sensing for Deep Learning–Based Human Action Recognition. Sensors (Basel, Switzerland), 19.
7. Le, V. (2022). Deep learning-based for human segmentation and tracking, 3D human pose estimation and action recognition on monocular video of MADS dataset. Multimedia Tools and Applications.
8. Ghosh, S.K., M, R., Mohan, B., & Guddeti, R.M. (2022). Deep learning-based multi-view 3D-human action recognition using skeleton and depth data. Multimedia Tools and Applications.
9. Jegham, I., Khalifa, A.B., Alouani, I., & Mahjoub, M.A. (2021). Soft Spatial Attention-Based Multimodal Driver Action Recognition Using Deep Learning. IEEE Sensors Journal, 21, 1918-1925.
10. Tsai, J., Hsu, C.J., Wang, W., & Huang, S. (2020). Deep Learning-Based Real-Time Multiple-Person Action Recognition System. Sensors (Basel, Switzerland), 20.
11. Ding, W., Ding, C., Li, G., & Liu, K. (2021). Skeleton-Based Square Grid for Human Action Recognition With 3D Convolutional Neural Network. *IEEE Access, 9*, 54078-54089.

12. Gao, Y., Xiang, X., Xiong, N.N., Huang, B., Lee, H.J., Alrifai, R., Jiang, X., & Fang, Z. (2018). Human Action Monitoring for Healthcare Based on Deep Learning. IEEE Access, 6, 52277-52285.
13. Moniruzzaman, M., Yin, Z., He, Z., Qin, R., & Leu, M.C. (2022). Human Action Recognition by Discriminative Feature Pooling and Video Segment Attention Model. IEEE Transactions on Multimedia, 24, 689-701.
14. Liu, D., Ji, Y., Ye, M., Gan, Y., & Zhang, J. (2020). An Improved Attention-Based Spatiotemporal-Stream Model for Action Recognition in Videos. IEEE Access, 8, 61462-61470.
15. Wu, H., Ma, X., & Li, Y. (2020). Convolutional Networks With Channel and STIPs Attention Model for Action Recognition in Videos. IEEE Transactions on Multimedia, 22, 2293-2306.