Vision based Human Activity Recognition using Hybrid Deep Learning

Aishvarya Garg
Department of Physical Science
Banasthali Vidyapith
Rajasthan, India
aishvaryagarg@gmail.com

Swati Nigam
Department of Computer Science
Banasthali Vidyapith
Rajasthan, India
swatinigam.au@gmail.com

Rajiv Singh
Department of Computer Science
Banasthali Vidyapith
Rajasthan, India
jkrajivsingh@gmail.com

2022 International Conference on Connected Systems & Intelligence (CSI) | 978-1-6654-5815-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSI54720.2022.9924016

***Abstract*--Human activity recognition is a wide research area of computer vision that finds applications in smart surveillance system, healthcare, and human robotic interactions. Nowadays, deep learning methods have achieved more interest due to its ability of executing feature extraction and classification steps simultaneously. In this paper, we have focused on the vision based human activity recognition using deep learning algorithms. Long short term memory (LSTM) is a special form of recurrent neural networks (RNN), specifically designed for long term data dependencies. Also it is a known fact that among deep learning algorithms, convolutional neural networks (CNN) have earned high performance in image classification. To overcome the limitation of LSTM in case of classification of static images, a hybrid CNN-LSTM model is proposed in which features are firstly extracted through CNN and then feed to LSTM as a sequence by the means of time distributed layer. This model is utilized for classifying six activities from two datasets which have shown the accuracy of 96.24% and 93.39% on KTH and Weizmann datasets, respectively. We have also implemented the CNN and LSTM models separately on these datasets with same parameters as used in hybrid model to study their impact on accuracy and loss.**

***Keywords*-- Human activity recognition, convolutional neural networks, long short-term memory.**

I. Introduction

The advent of technology has made a significant difference in our daily lives. Human activity recognition (HAR) is considered as one of them. It is the most trending research topic that falls under the applications of computer vision. It basically deals with the identification of activities that a person is doing on the basis of his/her body movements. It focuses on the detection of 4 categories: gestures (hand waving, stretching), actions (running, sitting, walking, laying), interactions (human-human, human-objects, human-computer), and group activities [1][2]. It has a pivotal participation in healthcare that includes the providence of immediate medical services as well as keeping a close observation to a person in an ambient assisted environment with the applications of Internet of Things (IoT), Bluetooth and Wi-Fi, Intelligence surveillance systems that includes automated detection of abnormal or suspicious activities with the help of CCTV and drones, intelligent driving, robotics systems as in human-computer interactions (HCI) that includes grasping of gestures, improving in the collaborations of robot and human actions [3].

In a HAR system, there are four steps to be followed: data collection, data pre-processing, training, and activity recognition as shown in Fig. 1. According to several conducted surveys [2][5][8], many proposals were put forward for HAR on the ground of implementation and data collection. As for the data collection, raw data can be obtained with the help of two sources: sensor based and vision-based methods. In sensor based, data can be taken from environmental sensors, optical sensors, wearable sensors, smart home sensors, and many more whereas in vision based, data can be videos or images from movies, YouTube and cameras [4]. For implementation, many machine learning (ML) algorithms such as support vector machines (SVM), K-means, etc. are utilized and give good results. Due to certain limitations in machine learning approaches like inability in handling huge data and time complexity and in order to attain more precise features and accuracy for the improved version of the system, deep learning approaches came into light such as convolutional neural networks (CNN), long short-term memory (LSTM), auto encoders (AE) and many more [6][7][9].

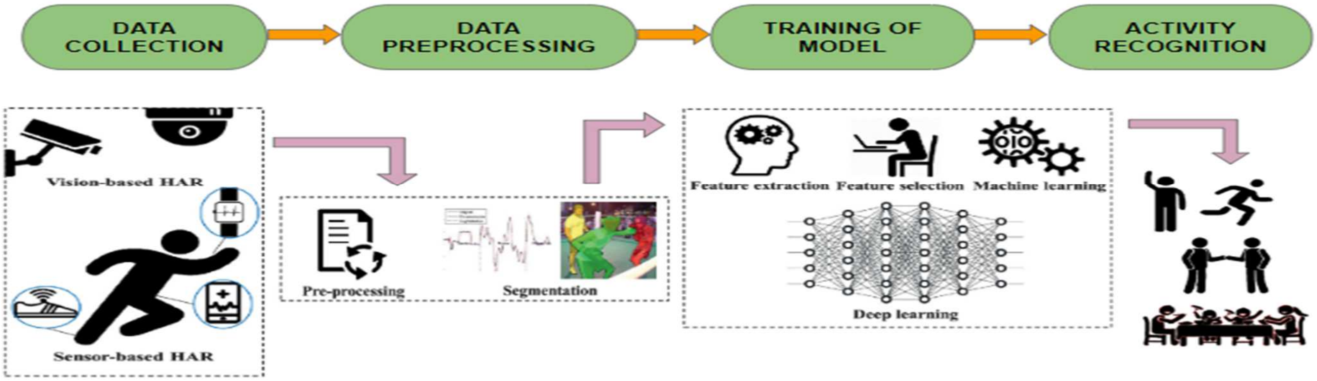


Fig.1. Human Activity Recognition [4]

978-1-6654-5815-3/22/$31.00 ©2022 IEEE

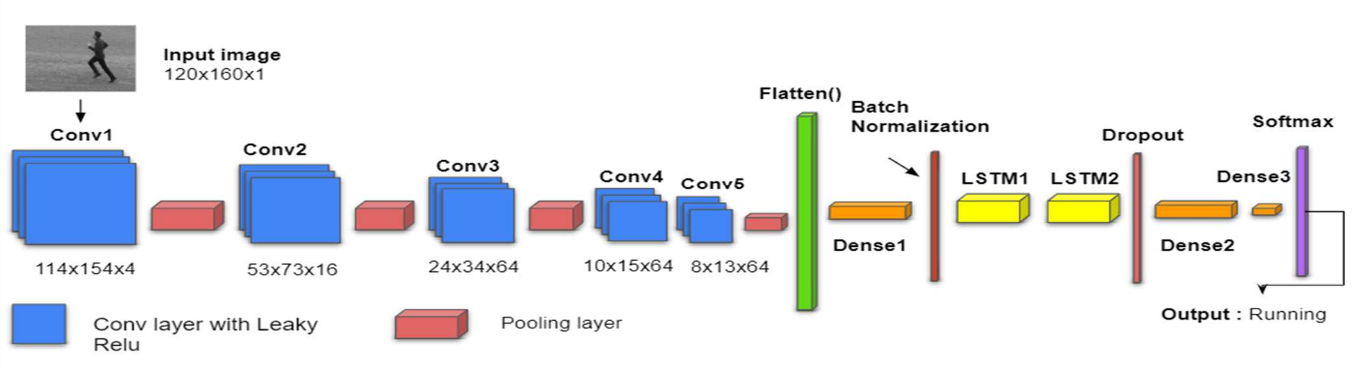


Fig. 2. Proposed model for Human Activity Recognition

The main aim of this work is to examine the performance of different models namely CNN and LSTM networks on the basis of accuracy and loss function and also to build a hybrid CNN-LSTM model for classifying six activities from two standard public vision-based datasets: Weizmann and KTH datasets.

In this paper, Section 2 presents the brief introduction and their related works of both deep learning algorithms that have been proposed for vision-based human activity recognition. Section 3 presents an overview of proposed work. Experimental setup for data and results of different models based on their training accuracy is presented in Section 4. Section 5 represents the conclusions and also the future scope on the basis of the study done in the paper.

II. Literature Review

In this section, a brief introduction of CNN and LSTM algorithms and their related works on the KTH and Weizmann datasets are provided.

The basic concept of working of the human brain is to acquire the unique features with the help of the net of neurons and give the accurate results according to its understanding. This concept serves as a building block of convolutional neural networks functionality. Like a net of neurons, it has layers for formulation of unique features. Its applications have been in object detection, speech recognition, text recognition, and image classifications. The network has three layers: convolution layer (consists of filters that goes through the input image and attains features), pooling layer (perform dimensionality reduction by reducing the parameters of input) and fully connected layer (performs the classification with accordance to the features extracted). The main criterion of the CNN is to extract unique features in order to create a crucial and flexible link between input and output of the model. [12][18].

Long short-term memory (LSTM) is an extended form of recurrent neural networks (RNN) that is designed for learning long term dependencies over long sequences of inputs. These are useful in sequence predictions, time series predictions, speech recognition. This network consists of four elements: cell (consists of a unique feature that runs through the whole block), input gate (responsible for adding information to the cell), forget gate (perform the duty of deleting information that is least important) and output gate (responsible for selecting important information in the cell as output) [12][19].

As we know, HAR has gained more interest and became a wide research topic since decades. Many researchers have proposed numerous deep learning-based models along with different techniques like transfer learning, active learning that attains a good accuracy. In [10][11], surveys were conducted that focuses on the utilization of CNN, its combination with other deep learning algorithms as well as machine learning in different classes of data such as sensor based, radar based, image and video, and many more along with their respective datasets for activity recognition. In [13], an attention based hybrid model of 2D/3D CNN and LSTM is proposed for HAR that attains an accuracy of 96.80% with KTH dataset. A 3D CNN based model is proposed for attaining the spatio-temporal features and then given as input to LSTM for classification of human actions with the two datasets KTH and Weizmann [14]. In [15], a combination of feature descriptors and gaussian mixture model (GMM) is utilized for the recognition of activities in which CNN algorithm is approached for the training and testing purposes. This model is applied on the public dataset Weizmann dataset and acquires an accuracy of 98.43%. A hybrid model of quaternion spatial-temporal CNN and LSTM is proposed which takes the input RGB image of three channels simultaneously without losing any spatial features. This model is implemented on Weizmann, UCF11 and UCF sports datasets [16]. A hybrid deep learning model is proposed for extracting the features from human motion in video using Kalman filters (KF) and GMM and other features from each frame on video sequences using recurrent neural networks (RNN) and gated RNN (GRNN) models. This approach is applied on three datasets, out of which KTH dataset attains an accuracy of 96.30% [17].

III. Proposed Model

As we know, LSTM is specially designed for sequence prediction therefore it will be difficult for LSTM for classifying activities as images have spatial features and LSTM derives the temporal information. Keeping this in mind, the proposed model is designed to perform human activity recognition with the help of a combination of 2D CNN and LSTM. In this approach, the features are extracted through a 2D CNN model and then given to the LSTM. For the setup, the input shape of the image is taken as 120 x 160 x 1. For training of model, 5 convolution layers having number of filters 4, 16, 64, 64, and

64 and size (7,7), (5,5), (3,3), (3,3) and (3,3) with stride (1,1) and padding as valid respectively are utilized. The pooling layer is taken as maxpooling2d with pool size (2, 2), and is added after conv2D layer. The first conv2d layer results with the dimensions 114x154 by convolving 4 filters of size 7x7 followed by maxpooling2d layer of size 2x2 results with the dimensions of 57x77. The conv2d_2 layer results with the dimensions 53x73 by convolving 16 filters of size 5x5 followed by maxpooling2d_1 layer of size 2x2 results with the dimensions of 26x36. The conv2d_3 layer results with the dimensions 24x34 by convolving 64 filters of size 3x3 followed by maxpooling2d_2 layer of size 2x2 results with the dimensions of 10x15. The conv2d_4 layer results with the dimensions 8x13 by convolving 64 filters of size 3x3. The conv2d_5 layer results with the dimensions 8x13 by convolving 64 filters of size 3x3 followed by maxpooling2d layer of size 2x2 results with the dimensions of 4x6. The CNN model is used for one image at a time and LSTM formulates the feature according to the observations made from the sequences, and distributed layer is introduced that wraps the whole CNN model at flatten and dense layer and gives it to LSTM as the sequence of images as desired input. Two LSTM layers with units 80 with activation function tanh are added. These layers are followed by 1 dense layer with 80 units with the activation function Leaky ReLU with alpha value 0.1. As there are 6 activities in both datasets such that last dense layer is added with the 6 units as input and activation function SoftMax that gives out the desired output as shown in Fig. 2. For the training, more than thousand images are given as input such that there may be chances of over fitting, to avoid this problem and to speed up we used batch normalization and dropout layers.

As we know ReLu is most popular activation function for its avoidance of occurrence of vanishing gradients, but there is one shortcoming of this function that is dying ReLU, happens when the derivative gradient is zero for negative value and there is no scope of learning or updating weights and hence affecting the accuracy. Taking this point as a note, we used Leaky ReLU as activation function as it introduced some slope which also called as alpha for negative input values and also speed up the training time. The model summary of proposed model is given in Table I.

TABLE I. Summary of Proposed Model

Layers	Output Shape
conv2d_1	(None, 114, 154, 4)
max_pooling2d	(None, 57, 77, 4)
conv2d_2	(None, 53, 73, 16)
max_pooling2d_1	(None, 26, 36, 16)
conv2d_3	(None, 24, 34, 64)
max_pooling2d_2	(None, 12, 17, 64)
conv2d_4	(None, 10, 15, 64)
conv2d_5	(None, 8, 13, 64)
max_pooling2d_3	(None, 4, 6, 64)
time_distributed	(None, 4, 384)
time_distributed_1	(None, 4, 80)
batch_normalization	(None, 4, 80)
lstm	(None, 4, 80)
lstm_1	(None, 80)
dropout	(None, 80)
dense_1	(None, 80)
dense_2	(None, 6)

IV. Experimental Results

A. Datasets

For the data processing, first frames are extracted from the video then these frames are given as the input image to the model. For the implementation of these models, two standard public datasets: KTH and Weizmann are used.

KTH dataset consists of 600 trimmed videos of 25 frame per second (fps) having 6 different activities namely, boxing, handclapping, handwaving, jogging, running, and walking. Each video has the dimensions of 120 x 160. These activities were performed by 25 individuals. Each video consists of a uniform background and a single person with different clothing, angle, and illumination conditions. For the training of data, all activities are taken into consideration.

Weizmann dataset consists of 90 videos of 30 fps having 10 different activities namely, handwave1, handwave2, walk, run, jump, skip, bend, gallop sideways, jumping jack, and jump in place. Each video has the dimensions of 144 x 180. These activities were performed by 9 individuals. For the training of data, out of 9 activities, 6 activities (handwave2, walk, run, bend, skip, and jump) are taken into consideration. The dataset statistics used for experiments are given in Table II.

TABLE II. Data Processing for Training

Parameters	KTH	Weizmann
Videos	600	54
Dimensions	120x160x1	120x160x1
Frames Taken per Video	100	35
Total Number of Frames	60000	1890

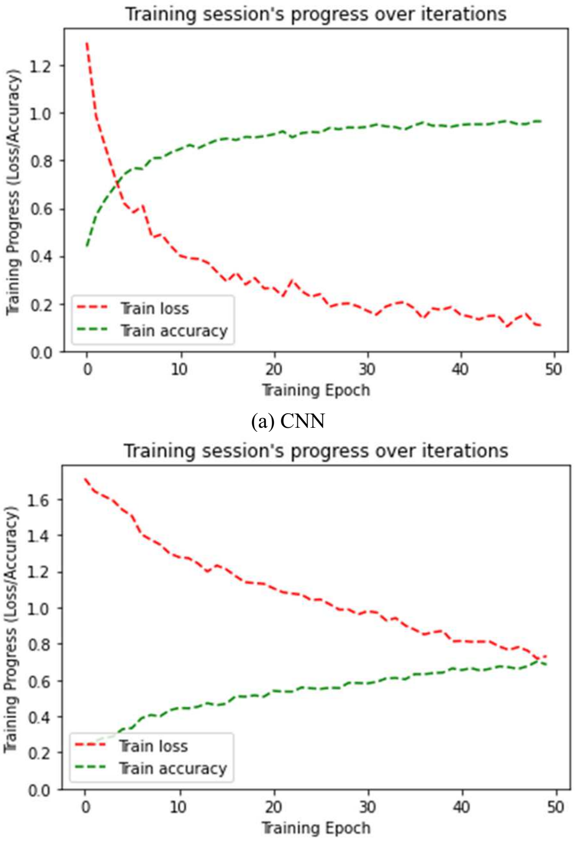
B. Results and Discussions

The implementations of models explained in previous sections were performed in the Jupyter Notebook environment of Google Colaboratory Pro with the Keras Deep Learning API. The models were compiled with the categorical cross entropy loss and optimizer RMSProp with activation function Leaky ReLU with alpha value 0.1. The data was divided in both datasets with the ratio of 8:2 for training and testing purposes. The batch size for KTH and Weizmann dataset were taken as 100 and 5 and run for 50 epochs respectively.

For the performance analysis, accuracy and confusion matrix is chosen. Accuracy can be defined as ratio of total number of correct predictions and total number of predictions. The confusion matrix is a tabular form of data which tell the performance of classification model on the basis of prediction that it makes on test data. The training accuracy is plotted along with the training loss with number of epochs is shown in Figs. 3, 4, and 5.

As mentioned, CNN model acquires good accuracy with any amount of data. It is cleared from the implementations of CNN model on both datasets KTH and Weizmann by acquiring the accuracy of 97.70% and 87.83% respectively and the training time is very less as compared to other two models. For the selection of number of layers, there are many factors like number of images, input dimensions, number of filters used and padding parameters. In view of these parameters, we first worked with 3 CNN layers then 5 layers, and results of latter are better as compared to the former's in classification of images. As we know more is number of layers, deeper is model and more distinguished features will be extracted and good accuracy is there, so we attempted to increase the layers by two, and results in crashing of the model.

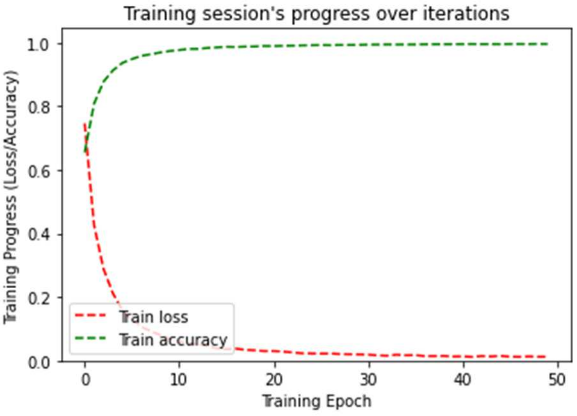
In the case of LSTM, as these images have spatial features such that there is an effect on accuracy as compared to CNN model in each dataset. It attains the accuracy of 93.53% and 51.05% in KTH and Weizmann dataset respectively and the training time is longer than others in both datasets.



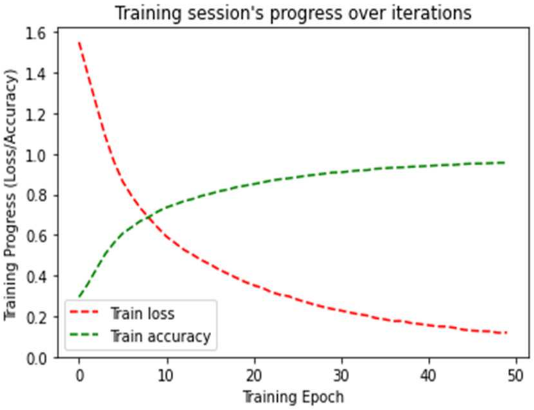
(a) CNN

(b) LSTM

Fig. 3. Training Loss/Accuracy on Weizmann Dataset

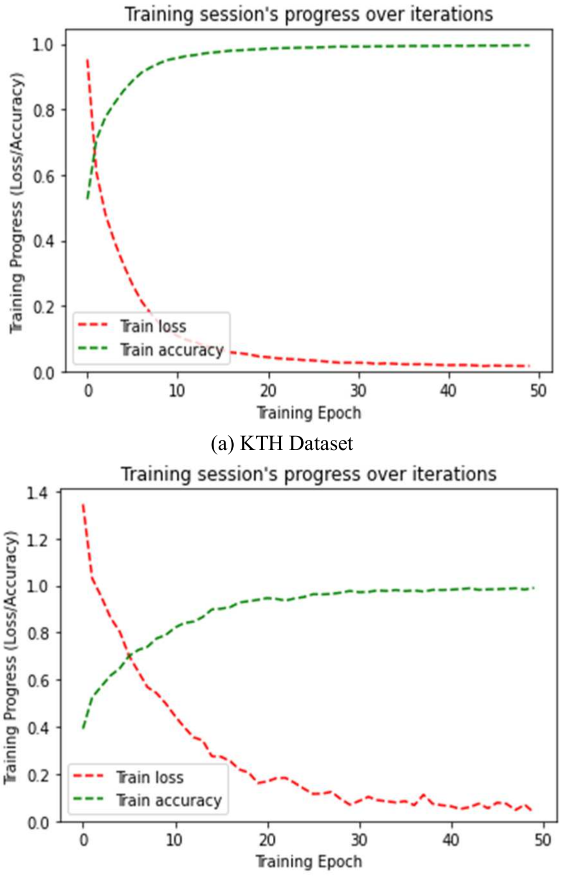


(a) CNN



(b) LSTM

Fig. 4. Training Loss/Accuracy and on KTH Dataset



(a) KTH Dataset

(b) Weizmann Dataset

Fig. 5. Training Loss/Accuracy of CNN+LSTM model

From the implementation of the proposed model i.e., combination of both CNN and LSTM, we attain the accuracy of 96.24% and 93.39% in KTH and Weizmann datasets respectively as shown in Tables III and IV. From these results, we can see that CNN help to overcome the limitation of LSTM in term of extracting spatial features of images and hence improves the accuracy in both datasets.

Table III. Performance metrics of the deep learning models on KTH Dataset

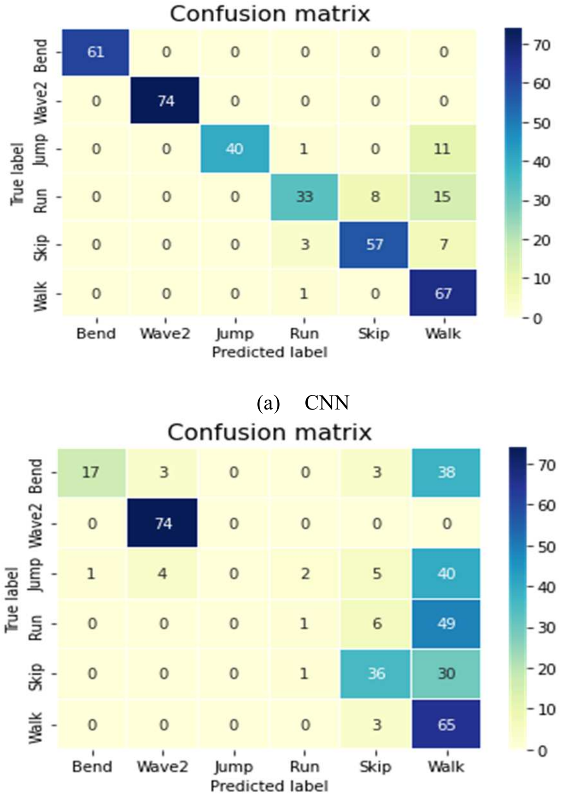
Metrics/Models	CNN	LSTM	CNN+LSTM
Accuracy	97.70	93.53	96.24
Precision	97.57	93.65	96.50
Recall	97.74	93.60	95.49
F1 Score	97.67	93.36	96.26

Table IV. Performance metrics of the deep learning models on Weizmann Dataset

Metrics/Models	CNN	LSTM	CNN+LSTM
Accuracy	87.83	51.05	93.39
Precision	90.20	51.32	92.80
Recall	86.57	41.83	93.64
F1 Score	85.92	46.09	93.22

The confusion matrix is also formulated with the graph as the performance metrics for both datasets and models as shown in Figs. 6, 7 and 8. The performance metrics, namely, precision, recall and F1 score are formulated with the help of confusion matrices of both datasets as shown in Tables III and IV.

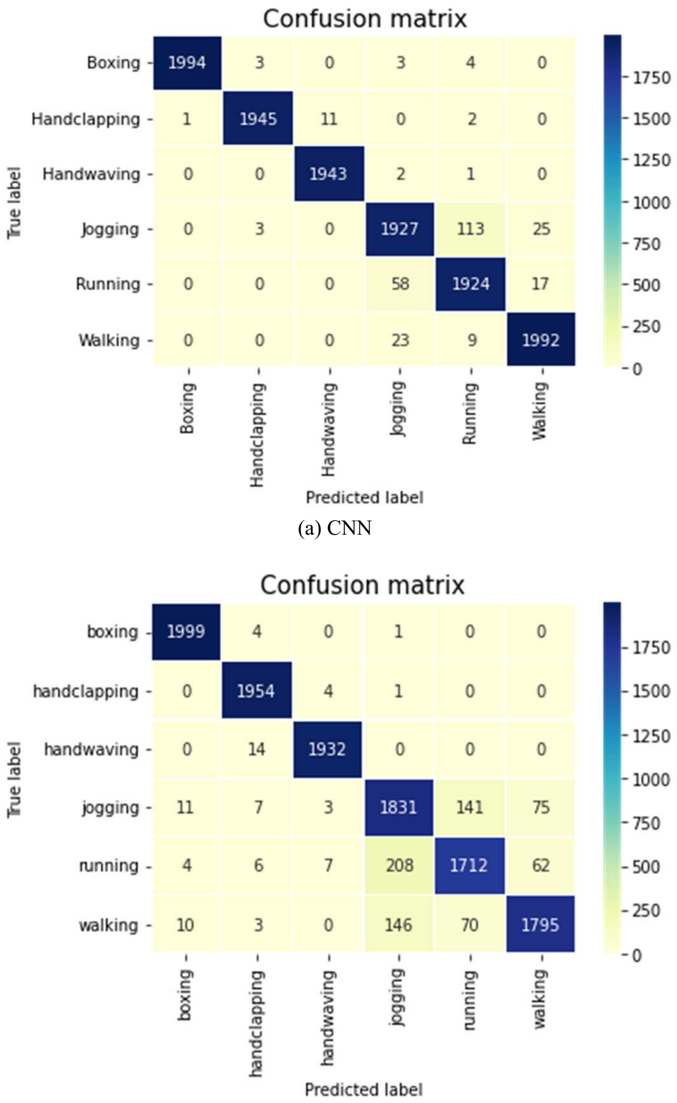
From confusion matrix of Weizmann dataset with LSTM and proposed models in Fig 6(b), we attain the precision and recall metrics which was unable to predict jump class which in turns affects the accuracy. But when we applied proposed architecture with same parameters, we get better results as shown in Table IV. From this result, we can say that LSTM model works with less efficiency in case of spatial features and CNN overcomes the limitation and give the better results. We can see that the loss is degrading with the increasing number of epochs. This is an indication of absence of over fitting of model.



(a) CNN

(b) LSTM

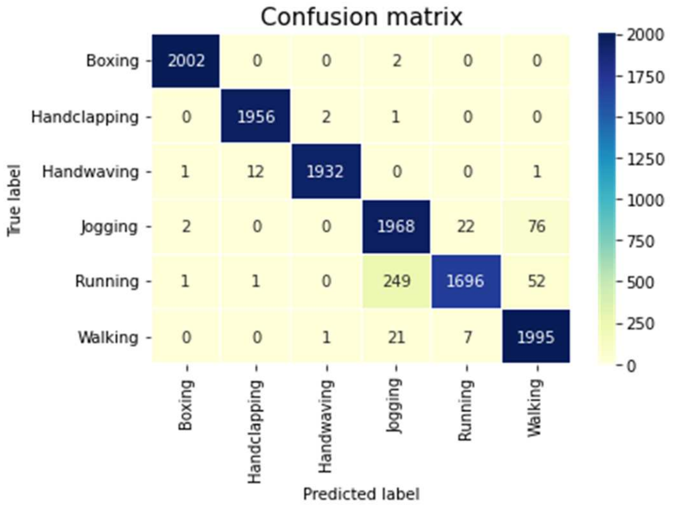
Fig. 6. Confusion Matrices for Weizmann Dataset



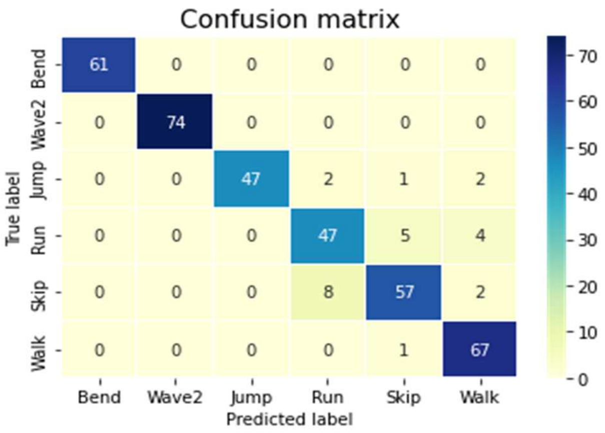
(a) CNN

(b) LSTM

Fig.7. Confusion Matrices for KTH Dataset



(a) KTH Dataset



(b) Weizmann Dataset

Fig. 8. Confusion Matrices for Proposed model

V. Conclusions

This paper presents the study of performance of two models CNN and LSTM on the static images with two standard public datasets for activity recognition namely, KTH and Weizmann. The results from both models showed that CNN give good accuracy for image classification. But in case of the LSTM, model took more time and but also accuracy is also less compared to CNN model with both datasets, this confirms that it is difficult for LSTM to acquire the spatial features as it is designed for sequence-based data or long data dependencies and hence it is unable to extract features more accurately than those models. From the studies made from implementations of these models separately with same parameters, a hybrid model, i.e., CNN-LSTM based HAR model is proposed that learns and extracts features through CNN and then further given as sequences to LSTM for recognition of activities. The deep learned models implemented on KTH and Weizmann datasets attain an accuracy of 96.24% and 93.39% respectively. From the performance metrics it can be seen that CNN overcomes the limitation of LSTM to some extent. This framework not only can be extended to recognize more complex activities from other public datasets but also be used for training and testing of models with machine learned features or descriptor based learned features in future works.

References

[1] S. Vishwakarma, and A. Agarwal, "A survey on activity recognition and behavior understanding in video surveillance." *The Visual Computer,* vol. 29, no. 10, pp. 983-1009, 2013.

[2] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey", *Multimedia Tools and Applications,* vol.79, no. 41, pp. 30509-30555, 2020.

[3] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning- based action representation approaches for human activity recognition", *Applied Sciences,* vol. 7 no. 1, pp. 110, 2017.

[4] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor based and vision based human activity recognition: A comprehensive survey", *Pattern Recognition,* vol. 108*,* pp. 107561, 2020.

[5] S. Zhang, Z. Wei, J. Nei, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision based method", *Journal of Healthcare Engineering,* 2017.

[6] M. A. R. Ahad, A. D. Antar, and M. Ahmed, "Deep learning for sensor based activity recognition: recent trends", *IoT Sensor-Based Activity Recognition,* pp. 149-173, 2021.

[7] A. A. Aguileta, R. F. Brena, O. Mayora, E. Molino-Minero-Re, and L. A. Trejo, "Multi-sensor fusion for activity recognition- A survey", *Sensors,* vol. 19, no. 17, p. 3808, pp. 5-7, 2019.

[8] Z. Hussain, M. Sheng, and W. E. Zhang, "Different approaches for human activity recognition: a survey" *arXiv preprint arXiv: 1906.05074*, 2019.

[9] A. Subasi, K. Khateeb, T. Brahmi and A. Sarrirete, "Human activity recognition using machine learning methods in a smart healthcare environment" *Innovation in Health Informatics*, pp. 131-134, 2020.

[10] M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, Human activity Recognition Using Tools of Convolutional Neural Networks: A State of the Art Review, Data Sets, Challenges and Future Prospects", *arXiv: 2202.03274*, 2022*.*

[11] Z. Malik, and M. I. B. Shapiai, "Human action interpretation using convolutional neural network: a survey" *Machine Vision and Applications,* vol. 33, no. 333(3), pp. 1-23, 2022.

[12] C. Shiranthika, N. Premakumara, H. –L. Chiu, H. Samani, C. Shyalika and C. -Y. Yang, "Human Activity Recognition Using CNN & LSTM", *5th International Conference on Information Technology Research (ICITR)*, pp. 1-6, 2020.

[13] K. Bayoudh, F. Hamdaoui, and A. Mtibaa, "An Attention-based Hybrid 2D/3D CNN-LSTM for Human Action Recognition", *2022 2nd International Conference on Computing and Information Technology (ICCIT),* pp. 97-103, 2022.

[14] S. H. Basha, V. Pulabaigari, and S. Mukherjee, "An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos", *arXiv preprint arXiv: 2002.02100*, 2020.

[15] J. Basavaiah, and C. M. Patil, "Human activity detection and action recognition in videos using convolutional neural networks", *Journal of Information and Communication Technology,* vol. 19, no. 2, pp. 157-183, 2020.

[16] B. Meng, X. Liu, and X. Wang, "Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos", *Multimedia Tools and Applications,* vol. 77, no. 20, pp. 26901-26918, 2018.

[17] N. Jaouedi, N. Boujnah, M. S. Bouhlel, "A new hybrid deep learning model for human action recognition", *Journal of King Saud University-Computer and information Sciences,* vol. 32, no. 4, pp. 447-453, 2020.

[18] *Introduction to Convolutional Neural Networks CNNs.* AIgents. Accessed on April 2, 2022. [Online]. Available: https://aigents.co/data-science-blog/publication/introduction-to-convolutional-neuaral-networks-cnns

[19] *Essentials of Deep Learning: Introduction to Long Short-Term Memory.* Analytics Vidhya. December 10, 2017. Accessed on April 2, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/