Vision based Human Activity Recognition using Deep Neural Network Framework

Jitha Janardhanan[1]*
Research Scholar, Department of Computer Science
Dr.G.R. D College of Science
Coimbatore, Tamil Nadu, India

Dr.S.Umamaheswari[2]
Associate Professor, Department of Computer Science
Dr.G.R. D College of Science
Coimbatore, TamilNadu, India

***Abstract*—Human Activity Recognition (HAR) has become a well-liked subject in study as of its broad application. With the growth of deep learning, novel thoughts have emerged to tackle HAR issues. One example is recognizing human behaviors without exposing a person's identify. Advanced computer vision approaches, on the other hand, are still thought to be potential development directions for constructing a human activity classification approach from a series of video frames. To solve this issue, a deep learning neural network technique using Depthwise Separable Convolution (DSC) with Bidirectional Long Short-Term Memory (DSC-BLSTM) is proposed here. The redeeming features of the proposed network system comprises a DSC convolution that helps to reduce not only the number of learnable parameters but also computational cost in together training and testing method The bidirectional LSTM process can combine the positive and the negative time direction. The proposed method comprises of three phases, which includes Video data preparation, Feature Extraction using Depthwise Separable Convolution Neural Network algorithm and DSC-BLSTM algorithm. The proposed DSC-BLSTM method obtains high accuracy, F1-score when compared to other HAR algorithms like MC-HF-SVM, Baseline LSTM Bidir-LSTM algorithms.**

***Keywords*—Activity recognition; long short-term memory (LSTM); deep learning; feature extraction**

I. INTRODUCTION

The Human Activity Recognition (HAR) system, a broadly used pattern recognition system discussed by [1], can be separated into numerous parts such as feature extraction, sensing segmentation, post-processing and classification ([11], [15]–[19]). HAR systems can be classified into two categories acceleration-based and time-based. Acceleration-based techniques need several accelerometers to be used for data gathering, but time-based techniques normally require the use of additional cameras to gather data. The drawback of the acceleration technique is that it can cause discomfort to the human body when performing behavior such as running, lying down and walking.

The different human activities to be observed in this paper include hand washing, punching, kicking, yoga, riding a bike, curling hair, ice skating etc. The benefit of a vision-based system is that the sensor works without attaching to the body. The detection of performance depends on illumination surroundings, screening angle, and extra factors. The paper proposed a system that uses a kinetics data set [3] and a MobileNetV2 structure with Bidirectional Long Short-Term Memory (Bidir-LSTM) classification ([4], [25]) to solve this problem. This can decrease the actions of the handcraft procedure and boost the accuracy [20].

Human activity recognition is the problem of recognizing and classifying specific human actions executed in video frames. An instance of such a human action could be kicking or pull-ups. A classification can be trained on specific instances of an activity (training set) and then tested on a specific instance of an activity (test set). The aim of the system is to recognize the proper class of action to which a video frame belongs, or further generally, to recognize and appreciate what the human is doing in the video frames [13].

MobileNetV2 is 2D resource efficient architecture. It implements of MobileNetV1 using depth-wise separable convolutions. It establishes 2 novel sections: 1) linear bottlenecks among the layers, and 2) shortcut connections among the bottlenecks. The design is following the dimension that reduces amount of channels and extracts as much as information by depth-wise convolution after decompressing the data. This convolutional module permits reducing memory usage during inference ([21], [23]).

The objective of this Human activity recognition is to extend the deep learning approaches currently being developed, specifically targeting classification and potentially training/retraining on constrained computing environments. However, an integration of Depthwise Separable Convolution with LSTMs incurs significant computational expenses and prevents the network from running in real-time. To solve this problem, this paper introduced a Depthwise Separable Convolution with Bidirectional-LSTM (DSC-BLSTM) principles to reduce computational costs.

The rest of the paper is organized as follows: Related work is detailed in Section 2. In Section 3, proposed methodology of video frame extraction and Depthwise Separable Convolution with Bidirectional-LSTM (DSC-BLSTM) are described. In Section 4, experimental results and discussion are described finally conclusion portion is in Section 5.

*Corresponding Author.
Paper Submission Date: May 11, 2022
Acceptance Date: June 13, 2022

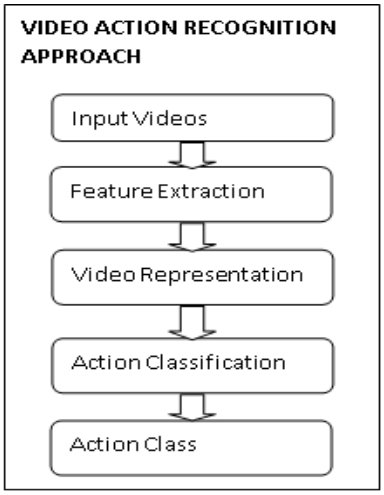

Fig. 1. Overview of Typical Video Action Recognition Approach.

II. Related Work

Das and Chakrabarty [5] have presented human gender detection system approach. The Silhouettes from Center for Biometrics and Security Research (CASIA) step are portioned to recognize main body locations and to produce equivalent point-light demonstrate. The attributes such as two dimensional coordinates of main body locations and combined positions are mined from the point-light present. The attributes are categorized using Support Vector Machines (SVM) and Hidden Markov Model (HMM). The revision performs a detection rate of 76.79 percent and 69.18 percent with 100 subject data using SVM and HMM correspondingly.

Hammerla et al., [6] authors use sensor data that contains different motions, explored deep, convolution and recurrent approaches on these datasets. A new proposal has been put forth for regularization of recurrent networks. They have concluded that discovered recurrent networks outstripped the state-of-art and illustrated sample by sample prediction of physical activity.

Munzner et al [7], has evaluated PAMP2 and RBK dataset with convolutional neural network. The article illustrates the outperformance of early fusion technique over late -hybrid fusion by improvising F1 -score on RBK dataset.

M. Panwar et al., [8], discussed various developments that had taken place in human activity recognition using different machine learning approaches. Though, feature engineering has conquered conventional techniques connecting the complicated procedure of best feature selection. This difficulty has been mitigated by using a new method based on deep learning framework which automatically mines the positive features and decrease the processing cost.

A. Jain and V. Kanhangad [9] proposed a descriptor-based approach for action prediction using built-in sensors of smart-phones. Gyroscope and accelerometer sensor signals are obtained to recognize the behaviors achieved by the client. The authors described a histogram of gradient and centroid signature-based Fourier descriptor that are utilized to mine feature or attribute sets from these signals. Attribute and gain level synthesis are discovered for in order fusion.

A. Ignatov [10] presented a user-independent deep learning-based approach for online human activity classification. The authors proposed Convolutional Neural Networks for local attribute extraction jointly with plain arithmetical attributes that conserve information regarding the large-scale form of time series. Moreover, they investigated the crash of time series length on the identification accuracy and boundary it up to one second that builds potential and permanent real-time action prediction.

III. Proposed Methodology

The proposed Depthwise Separable Convolution with Bidirectional-LSTM (DSC-BLSTM) technique performs to test the experimentations that were accomplished with kinetics-400 dataset [3]. The DSC-BLSTM algorithm successfully detects the human activity in video frames and is grouped into different activity classes. The overall DSC-BLSTM workflow is illustrated in Fig. 2.

A. Video Frame Extraction

Video frame extraction is executed with Kinetics 400 action recognition dataset of action videos, accumulated from YouTube. With 306,245 short trimmed videos from 400 action categories. It is one of the largest and most widely used dataset in the research community for benchmarking state-of-the-art video action recognition models [24], described in Fig. 1. Since some YouTube links are expired, so could only download 234,584 of the original datasets, thus missing 11,951 videos from the training set, which are about 5%. This leads to a slight drop in performance of about 0.5%. The example video frame extraction result of riding camel video illustrated in Fig. 3.

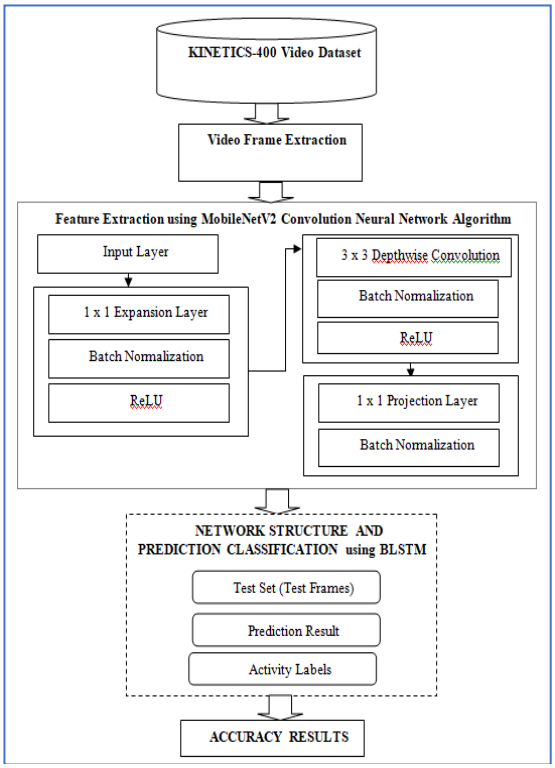

Fig. 2. Proposed DSC-BLSTM Algorithm Flow Diagram.

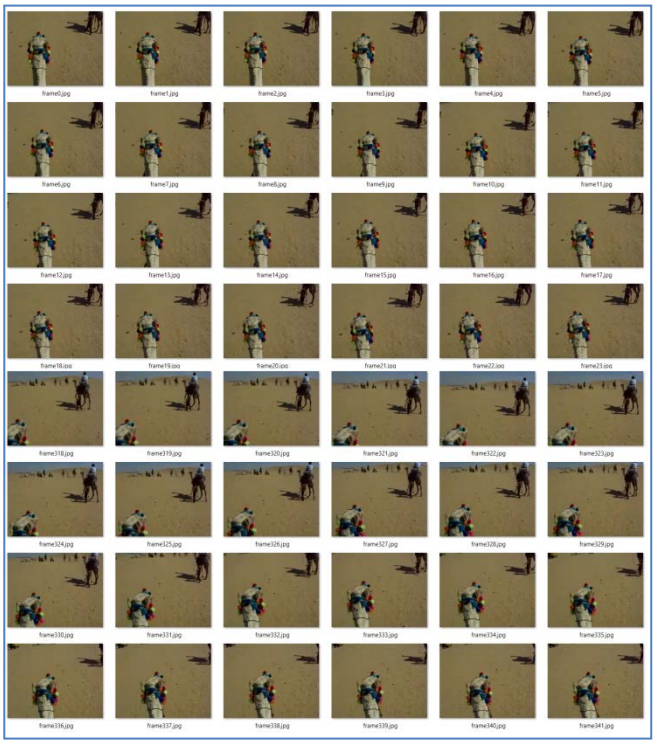

Fig. 3. Video Frame Extraction Result.

B. *Feature Extraction using Depthwise Separable Convolution Neural Network Algorithm*

The Feature extraction using MobileNetV2 convolution network using DSC is a factorized form of the standard convolution. The initial feature extraction layer is a 1 x 1 expansion layer. It increases the data (enhancing the number of channels) that flows through it. It does the opposite of the projection layer. The each video frame gets expanded based on the expansion factor [22]. This is a hyper parameter to be found from different architecture trade-offs. The default expansion factor is 6. A normal convolution is separated into a DC and a 1 x 1 PC. Instead of applying each filter to all the channels of the input like the standard convolution, the DC layer applies one filter to one input channel, then a 1x1 PC is employed to combine the outputs of the DC. DSC helps to reduce not only the number of learnable parameters but also the computational cost in both training and testing process. The flow diagram of DSC is described in Fig. 4.

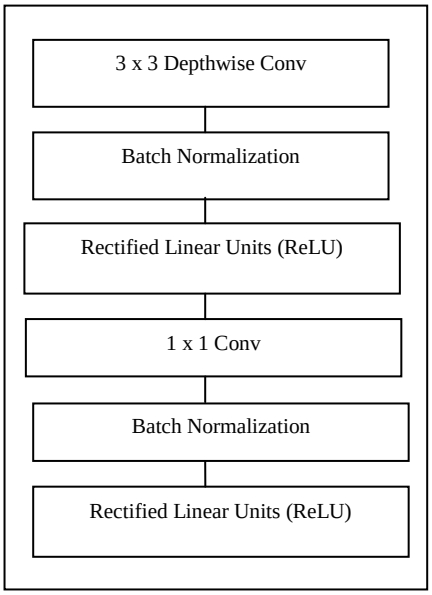

Fig. 4. Flow Diagram of DSC.

The proposed system of DSC is performed by 2 layers: 1) Depthwise Convolutions (DC) and 2) Point wise Convolutions (PC). The every input (contribution) channel (contribution depth) is applied by a filter with DC PC is a simple 1 x 1 convolution, is used to build linear permutation of the result of DC layer. MobileNets utilize equally batchnorm and ReLU nonlinearities layers. DC with single filter for each contribution channel can be written as:

$$Fmap_{k,m,n} = \sum_{a,b} Dw_{i,j,n} \cdot F_{k+i-1,m+j-1,n} eqn. \quad (1)$$

where Dw is the DC kernel size $Dw_{k*}Dw_{K*}$ N where n^{th} filter in Dw is functional to the n^{th} channel in F to create n^{th} channel filtered result of feature map Fmap.

DC cost is defined by,

$$Dw_k \cdot Dw_k \cdot N \cdot D_F \cdot D_F eqn. \quad (2)$$

The grouping of DC and 1 x 1 PC is called DSC which was formerly introduced by [11].

DSC cost is defined by equation 3,

$$Dw_k \cdot Dw_k \cdot N \cdot D_F \cdot D_F + N \cdot M \cdot D_F \cdot D_F eqn. \quad (3)$$

By stating convolution as a two-step procedure of filtering and combining to obtain decrease in calculation is,

$$\frac{Dw_k \cdot Dw_k \cdot N \cdot D_F \cdot D_F + N \cdot M \cdot D_F \cdot D_F}{Dw_k \cdot Dw_k \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{Dw_k^2} \; eqn \quad (4)$$

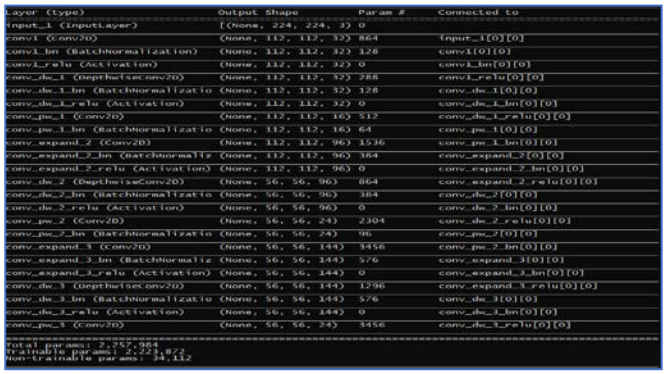

Fig. 5. Depthwise Separable Convolutions base Model Result.

MobileNetV2 uses 3 x 3 DSC take on among eight to nine times less computation than normal convolutions having less accuracy. By exploring the network in simple terms that are proficient to simply explore network topologies to discover an excellent network. The proposed training model is defined in Fig. 5. The entire layers are trailed by batchnorm and ReLU nonlinearity with the exclusion of the last completely connected layer which has no nonlinearity and brought into a SoftMax layer for classification.

C. *Depthwise Separable Convolution with Bidirectional-LSTM (DSC-BLSTM)*

According to [2], [12], authors described LSTM as an expansion of recurrent neural networks. Appropriate to unique architecture, which conflicts the disappearance and ignition gradient issues, it is fine at managing time series issues up to a positive depth. LSTM conserve information from inputs that has previously passed through it via the hidden state. Unidirectional LSTM only conserve information of the precedent because of the simple inputs it has observed from the past. The bidirectional LSTM will process the inputs with one from past to future and one from future to past and what varies this approach from unidirectional is that in the LSTM processes backwards to protect information from the future and using the two hidden states that are combined in some point to protect information from together past and future. Bidirectional LSTM (see Fig. 6) comprises of 2 LSTM cells, and the result is resolute. The bidirectional LSTM presented result is not only associated to prior information but also connected to consequent information. The overall DSC-BLSTM flow diagram is illustrated in Fig. 7.

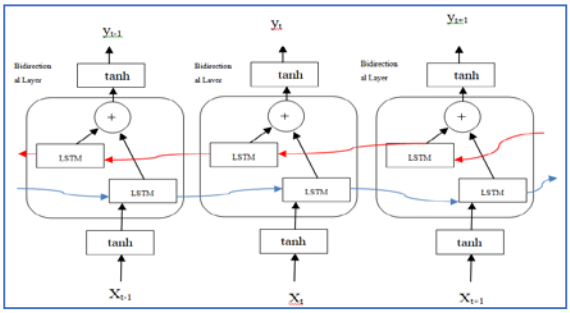

Fig. 6. Bidirectional LSTM Construction.

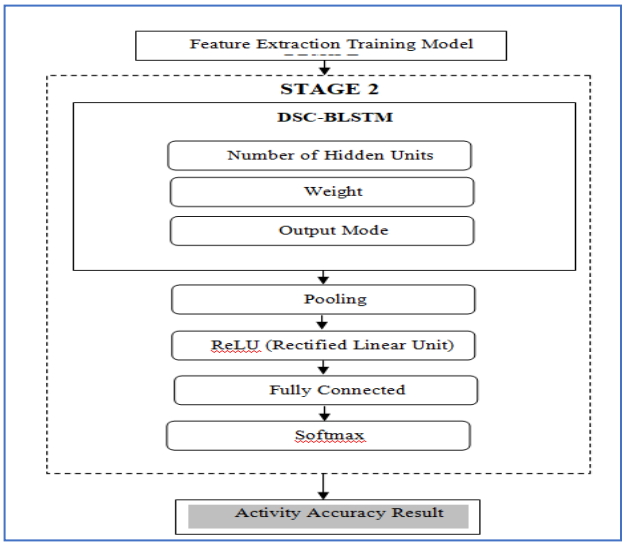

Fig. 7. DSC-BLSTM Flow Diagram.

Algorithm: DSC-BLSTM Pseudo code

Input: Continuous Video Frames *f*, Class Labels C
Output: Predicted activity class with accuracy score
Preparation:

1. Video Data Preparation
2. Feature Extraction using Depthwise Separable Convolution Neural Network (DSC)
3. Depthwise Separable Convolution with Bidirectional-LSTM (DSC-BLSTM)

Steps:

While (video frame)

1. Frame *f* ← Extract frames from videos

   //Fix sample duration (i.e., frames taken per loop/iteration) = 16 frames per iteration and sample size (i.e., Frame width size) = 112pixels wide
2. *M* ← Create trained model using DSC method

End While

3. **for** t = 1 to n **do** // where *n* represents number of video frames
   a. frame *f*(t) ← Read the test video frame.
   b. Apply *f*(*t*) to DSC Model // *Calculate Similarity matrix value of test frame f(t) with trained model*
   c. Predict activity Class label with frame *f*(*t*) using (DSC-BLSTM)
4. Label Predicted activity ← Result class label
5. Display predicted activity class with accuracy score

End for

The DSC-BLSTM algorithm takes these three inputs Model M, classes C and test video frames f. Initially, load the trained human activity recognition DSC model M and contents of the class labels C file. After that, test frames are grouped and resized with defined sample duration (i.e., number of frames for classification) and test size (i.e., the spatial dimensions of the frame). Next, test input video frames are looped over the amount of essential test frames (i.e., duration of 16 frames per loop) and read a test frame from the video stream. The test frame streams are forwarded to the network model for checking the distance matrix (i.e., similarity) between train model M and test frame f model outputs. The outputs matrix is passed through Bidirectional LSTM process to get the activity label. Finally, the maximum classes of label are the predicted activity for the processed frame. The Table I shows the parameters, symbol with the corresponding value of the proposed system implementation.

TABLE I. PARAMETERS DESCRIPTION

Parameter Name	Symbol	Value
Total Classes	C	400 action classes
Frame Duration	Sample_duration	16 frames per iteration
Sample window Size	Sample_size	112 pixel wide
Training Model	*M*	2000 Videos model
Test video frame	*f*	-
Total Number of Frames	*n*	Total no of frames in a video

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Environment

The results have been estimated with the proposed DSC-BLSTM algorithm. The results are implemented with Intel I5-6500U series 2.71 GHz, x64-based processor, 8GB main memory, and run on the Windows 10 operating system using python 3.8 simulations. This paper is implemented with Kinetics 400 action recognition dataset which consists of 2000 videos as training dataset with 21 classes that includes massage, Ice Skating, Yoga, Playing, Pull ups, Pushing, Reading, Tasting, Skating, Side Kick, Filling, Crawling, Waiting, Washing, Making Pizza, Kicking, Jumping, Curling, Dancing, Massage, Shaving and Shooting. The resulting parameters of Ice-Skating activity are described in Fig. 8 to Fig. 10.

B. Discussion

This section presents a detailed analysis of experimental outcomes through the proposed method on the basis of accuracy measures such as precision, recall, accuracy, and F1-score. The proposed algorithm consists of three main stages. These three main stages include the video frame extraction which is performed first in which datasets are normalized to get better results. The accurate results will give more accuracy. In the second step, feature extraction is implemented using Depthwise Separable Convolution Neural Network algorithm. In this step, features is implemented separately based on DSC to get the best features are stored in trained dataset. In the third step, features are fused, while in the final stage, results are taken through the classification learner. In video frame extraction, all individual frames are stored as images to detect the activity from the image. After the frame extraction, MobileNetV2 convolution network using DSC method to create a feature model from extracted images. Finally, Depthwise Separable Convolution with Bidirectional-LSTM (DSC-BLSTM) classifier to discover the similarity matrix value of test frame with trained model to attain the prediction activity result. Combining DSC-BLSTM have achieved higher accuracy than the other classification learners on the "MC-HF-SVM, Baseline LSTM and Bidir-LLSTM algorithms", respectively.

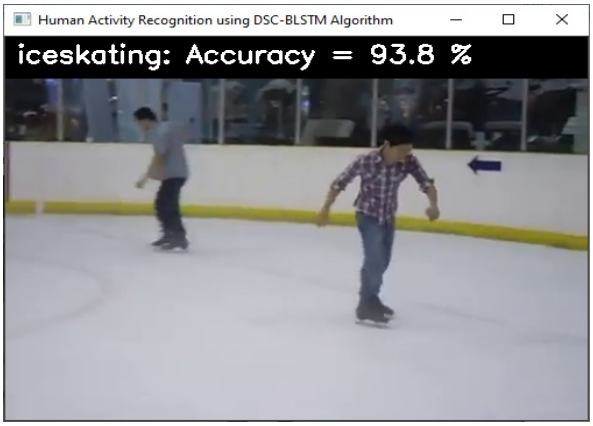

Fig. 8. HAR Result of Ice Skating with Accuracy.

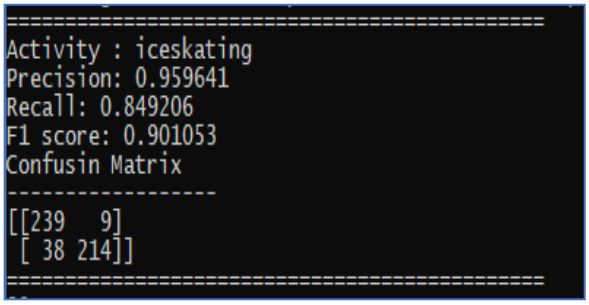

Fig. 9. HAR Result of Ice Skating Performance Measure of Precision, Recall, F1 Score and Confusion Matrix.

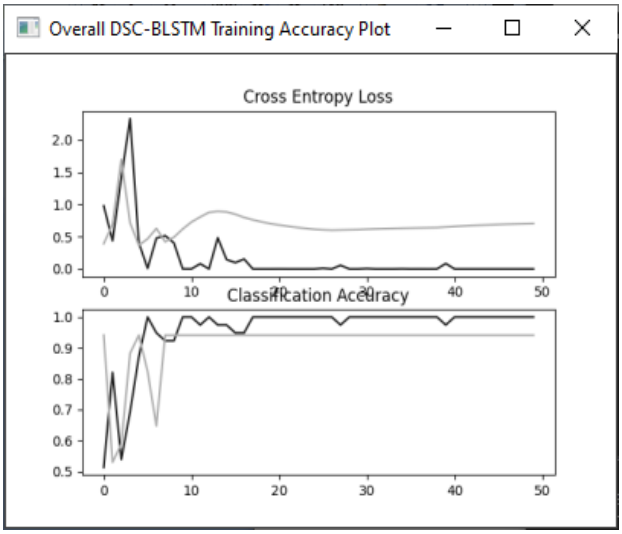

Fig. 10. Overall DSC-BLSTM Training Loss and Accuracy Plot Result.

C. Evaluation Index

To evaluate the performance of the proposed model for HAR, the followed metrics [26] were used for evaluation generally.

$$Accuracy = \frac{tp+tn}{tp+fn+fp+tn} \quad (5)$$

$$Precision = \frac{tp}{tp+fp} \quad (6)$$

$$Recall = \frac{tp}{tp+fn} \quad (7)$$

$$F1-score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (8)$$

The proposed DSC-BLSTM method substantiates with high Accuracy and F1-score ratio when compared to other HAR algorithms like Multiclass Hardware-Friendly Support Vector Machine (MC-HF-SVM) [7], Baseline LSTM [14], Bidirectional Long Short Term Memory (Bidir-LSTM) [4] algorithm. As a result of the improved human activity recognition presentation, there is a higher accuracy. The proposed DSC-BLSTM method proves high Accuracy and F1-score ratio when compared to other HAR algorithms like Multiclass Hardware-Friendly Support Vector Machine (MC-HF-SVM) [7], Baseline LSTM [14], Bidirectional Long Short Term Memory (Bidir-LSTM) [4] algorithm are described in Table II and Fig. 11 shows the comparison chart.

In Table III shows the comparison of precision, recall, accuracy and F1 score with kinetics 400 dataset of test videos prediction activity measures and Fig. 12 shows the comparison chart.

TABLE II. COMPARISON OF ACCURACY AND F1 SCORE WITH EXISTING AND PROPOSED DSC-BLSTM ALGORITHM OF KINETICS 400 DATASET

Methods	MC-HF-SVM	Baseline LSTM	Bidir-LSTM	DSC-BLSTM
Accuracy	89.3	90.8	91.1	**93.8**
F1-Score	89.0	90.8	91.1	**92.637**

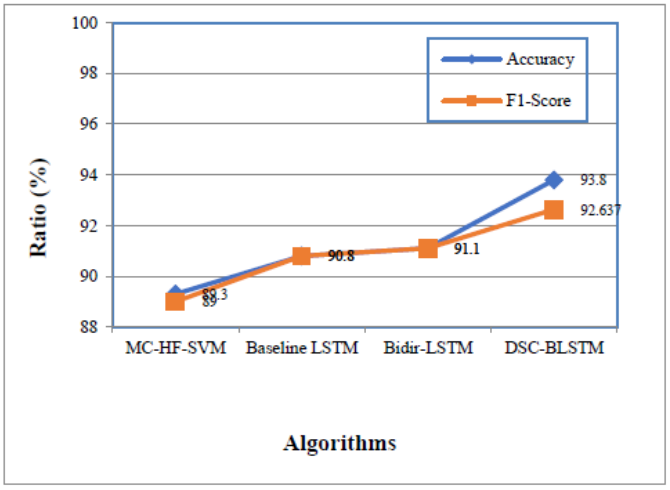

Fig. 11. Comparison Measures of Accuracy and F1 Score Chart.

TABLE III. COMPARISON OF TEST VIDEOS PREDICTION ACTIVITY MEASURES OF PROPOSED DSC-BLSTM ALGORITHM OF PRECISION, RECALL, ACCURACY AND F1 SCORE WITH KINETICS 400 DATASET

Prediction Activity	Precision	Recall	Accuracy	F1-Score
Ice Skating	96.25	91.66	94	93.90
Yoga	94.23	90.87	92.60	92.52
Playing	91.56	90.47	91	91.01
Pull ups	95	90.47	92.80	92.68
Pushing	95.08	92.06	93.60	93.54
Reading	91.49	89.68	90.60	90.58
Tasting	92.68	90.47	91.60	91.56
Skating	93.56	86.50	90.20	89.89
Side Kick	94.34	86.11	90.40	90.04
Filling	95.08	84.52	90.00	89.49
Crawing	96.39	84.92	90.80	90.29
Waiting	89.23	92.06	90.40	90.62
Washing	97.83	89.68	93.80	93.58
Making	94.97	90.07	92.60	92.46
Kicking	91.76	92.85	92.20	92.30
Jumping	95.79	90.47	93.20	93.06
Curling	93.00	89.68	91.40	91.31
Dancing	91.20	90.47	90.80	90.83
Massage	90.38	93.25	91.60	91.79
Shaving	91.39	88.49	90	89.91
Shooting	94.44	89.69	91.20	90.94

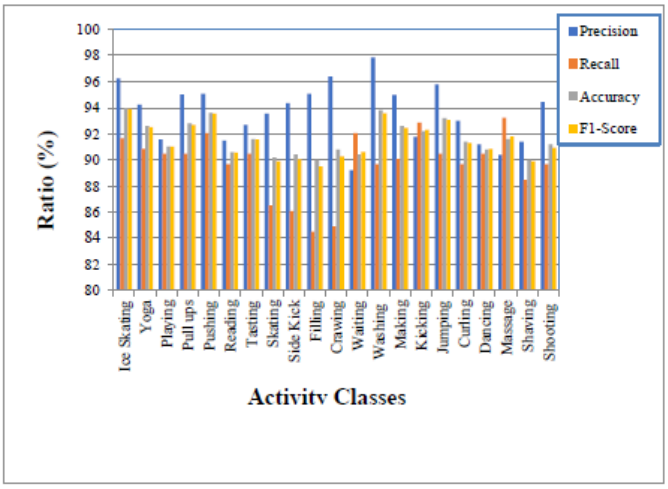

Fig. 12. Comparison of Test Video Class Prediction Activities of Performance Measures of Kinetics 400 Dataset.

V. CONCLUSION

This paper analyzed the advancement of Human Activity Recognition (HAR) concepts in the field of deep neural network technique. The proposed work presents a Depthwise Separable Convolution with Bidirectional long short-term

memory (DSC-BLSTM) algorithm adapted to the HAR task. This system seeks to improve the accuracy of activity recognition by leveraging the robustness in feature extraction and classification model. The results were impressive and worked well after we used the DSC-BLSTM method in the HAR system. The result is shown as 93.8%, and is more recognizable than the other HAR algorithms like MC-HF-SVM, Baseline LSTM, Bidir-LSTM algorithms.

VI. Conflict of Interest

The authors declare no conflict of interest

References

[1] J. Schmidhuber, “Deep Learning In Neural Networks: An Overview”, Neural Networks, vol. 61, pp. 85-117, 2015.

[2] D. C. Ciresan, U. Meier U, and L. M. Gambardella, “Deep, Big, Simple Neural Nets For Handwritten Digit Recognition”, Neural computation, vol. 22, no. 12, pp. 3207-3220, 2010.

[3] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, “The Kinetics Human Action Video Dataset”, 2017.

[4] Yu Zhao, Rennong Yang, Guillaume Chevalier, Ximeng Xu, and Zhenxing Zhang, “Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors,” Mathematical Problems in Engineering, Volume 2018.

[5] D. Das and A. Chakrabarty, “Human Gait-Based Gender Identification System Using Hidden Markov Model And Support Vector Machines,” in Conf. Comput. Commun. Autom. ICCCA 2015, pp. 268–272, 2015.

[6] N. Y. Hammerla, S. Halloran, and T. Ploetz, “Deep, Convolutional, And Recurrent Models For Human Activity Recognition Using Wearables,” arXiv preprint arXiv:1604.08880, 2016.

[7] S. M¨unzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. D¨urichen, “Cnn-Based Sensor Fusion Techniques For Multimodal Human Activity Recognition,” in Proceedings of the 2017ACM International Symposium on Wearable Computers, ser. ISWC ’17. New York, NY, USA: ACM, 2017, pp. 158–165.

[8] M. Panwar et al., “CNN Based Approach For Activity Recognition Using A Wrist-Worn Accelerometer,” in Proc. Annu. Int.Conf. IEEE Eng. Med. Biol. Soc. EMBS, no. July, pp. 2438–2441, 2017.

[9] A. Jain and V. Kanhangad, “Human Activity Classification in Smartphone’s Using Accelerometer and Gyroscope Sensors,” IEEE Sensors Journal, vol. 18, no. 3, pp. 1169-1177, 1 Feb.1, 2018.

[10] A. Ignatov, “Real-Time Human Activity Recognition From Accelerometer Data Using Convolutional Neural Networks,” Appl. Soft Comput. J., vol. 62, pp. 915–922, 2018.

[11] L. Sifre, “Rigid-motion Scattering for Image Classification,” PhD thesis, Department of Informatics, CMP Ecole Polytechnic, France., 2014.

[12] A. Krizhevsky, I. Sutskever, andG. E.Hinton, “Imagenet Classification With Deep Convolutional Neural Networks,” in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS ’12), pp. 1097–1105,Lake Tahoe, Nev, USA, December 2012.

[13] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “Energy Efficient Smartphone-Based Activity Recognition Using Fixed-Point Arithmetic,” Journal of Universal Computer Science, vol. 19, no. 9, pp. 1295–1314, 2013.

[14] N Srivastava, E Mansimov, R Salakhudinov,” Unsupervised Learning Of Video Representation Using LSTM”, International conference on machine learning, pp 843–852, 2015.

[15] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, “Action Classification In Soccer Videos With Long Short Term Memory Recurrent Neural Networks”, in Proceedings of ICANN, 2010.

[16] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential Deep Learning For Human Action Recognition,” Human Behavior Understanding, 2011.

[17] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell., “Long-Term Recurrent Convolutional Networks For Visual Recognition And Description”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2625–2634, 2015.

[18] A. Karpathy, G. Goderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li., “Large-Scale Video Classification With Convolutional Neural Networks,” In Proceedings of CVPR, 2014.

[19] K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks For Action Recognition In Video.” In arXiv preprint arxiv:1406.2199, 2014.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen., “Mobilenetv2: Inverted Residuals And Linear bottlenecks.” In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4510–4520. IEEE, 2018.

[21] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, “Pruning convolutional neural networks for resource efficient inference,” arXiv preprint arXiv:1611.06440, 2016.

[22] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, “Convnet Architecture Search for Spatiotemporal Feature Learning,” arXiv preprint arXiv:1708.05038, 2017.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. “Densely Connected Convolutional Networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708, 2017.

[24] Heng Wang and Cordelia Schmid. “Action recognition with improved trajectories”,in Computer Vision (ICCV), 2013 IEEE International Conference on, pp 3551–3558. IEEE, 2013.

[25] F. J. Ordonez and D. Roggen, “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition,” Sensors, vol. 16, no. 1, pp. 115–140, 2016.

[26] Yang, D.; Huang, J.; Tu, X.; Ding, G.; Shen, T.; Xiao, X, “A Wearable Activity Recognition Device Using Air-Pressure and IMUSensors,” IEEE Access 2019,7, 6611–6621.