International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Convolutional Long Short-Term Memory Hybrid Networks for Skeletal Based Human Action Recognition

K. Vijaya Prasad, P.V.V. Kishore, O. Srinivasa Rao

Abstract: ***The objective is to develop a time series image representation of the skeletal action data and use it for recognition through a convolutional long short-term deep learning framework. Consequently, Kinect captured human skeletal data is transformed into a Joint Change Distance Image (JCDI) descriptor which maps the time changes in the joints. Subsequently, JCDIs are decoded spatially well with a Convolutional (CNN). Temporal decomposition is executed on long short term memory (LSTM) with data changes along*** x, y ***and*** z ***position vectors of the skeleton. We propose a combination of CNN and LSTM which maps the spatio temporal information to generate a generalized time series features for recognition. Finally, scores are fused from spatially vibrant CNNs and temporally sound LSTMs for action recognition. Publicly available action datasets such as NTU RGBD, MSR Action, UTKinect and G3D were used as test inputs for experimentation. The results showed a better performance due to spatio temporal modeling at both the representation and the recognition stages when compared to other state-of-the-arts.***

Index Terms: ***Human action recognition, convolutional neural networks, long short term memory, Joint Change Distance Image, Microsoft Kinect.***

I. INTRODUCTION

Human actions in a video sequence have shown to have far reaching applications in security, surveillance and identification. Consequently, these applications add value, if a certain level of end to end automation is achieved. However, convolutional neural networks (CNN) has shown an increasing presence in this area of research. Previously, CNNs performed exceptionally well on images and fixed length video sequences. However, for variable length video sequences they showed a decrease in performance due to the traditional one versus all prediction models. Here, we propose an end to end trainable recurrent CNN with long short-term memory (LSTM) network to improve spatio temporal image recognition task on skeletal human action recognition.

Human action recognition has been researched extensively in the past decade using images, videos, depth and skeletal data as inputs. The processing algorithms usually used sailency detection, human extraction, tracking with kalman and particle filters; modeling the extractions using features and finally recognition with models in machine learning such as trees, support vector machines (SVMs), kernels and hidden markov models (HMMs). However, the processing is quite intense and requires a lot of algorithms to work synchronously resulting in accurate recognition. On the other hand, this computationally intensive approach has given away to efficient deep learning frameworks that accepted the raw video data to make decisions that are more accurate respectively. CNNs were widely considered for image processing tasks such as recognition and classification. They rapidly extend their presence into video data Search and identification applications. However, video RGB data seems to contain sensor irregularities, which became a hindrance in the recognition of human actions. Hence, the researchers turned their focus towards multi modal data obtained from Kinect sensor which contained RGB video, depth and skeletal data. The use of all three data has improved the recognition with CNNs, RNNs and other deep learning architectures.

Eventually, the large multi modal data has increased computational complexity of the algorithms which upscaled hardware requirements. In contrast to the previous models on human action recognition with multi modal data, we propose to use single modal data to achieve higher accuracies. This work presents a CNN based recurrent network on images generated from skeletal joint distance changes. The proposed CNN based LSTM architecture has the potential to mark spatial and temporal changes in the spatio temporal images.

To test the proposed CNN -LSTM architecture, we use NTU RGBD, MSR Action, UTKinect and G3D human action datasets. The rest of the paper is organized as follows. Section 2 describes the literature review related to the proposed framework. Section 3 gives the CNN - LSTM architecture followed by results and discussion in section 4. Finally, section 5 concludes the work.

II. LITERATURE REVIEW

Human action recognition has gained popularity in the last two decades due to widespread applications such as behavioural analysis, surveillance, security, interactive environments and animation synthesis [1]. Action recognition involves some generic steps such extraction of low-level features in spatial domain or mid-level action representations using motion or high-level semantic features [2].

Revised Manuscript Received on January 30, 2020.

* Correspondence Author

K. Vijaya Prasad *, is Research Scholar, Department of ECE, JNT University Kakinada, Kakinada, India. E-mail: vijayaprasad835@gmail.com

P.V.V. Kishore is Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, E-mail: pvvkishore@kluniversity.in

O. Srinivasa Rao, Professor, Department of CSE, UCEK, JNT University Kakinada, Kakinada, India. E-mail: osr_phd@yahoo.com

Whatever are the features, most of the works used a learning algorithm to learn the exacted features. HMMs [3], SVMs [4] and decision trees [5] were most popular in HAR.
The data for the above-mentioned methods were videos consisting of actions.

However, the video data has shown sensitivity towards various factors such as lighting, motion blurring and inter subject occultation. The results of most of the algorithms performed randomly due the above sensitive factors on evaluations such as cross subject and cross views. On the other hand, sensors such as Kinect provided multi modal data in the form of depth maps and skeletons additionally, to RGB videos. This additional information has transformed human action recognition greatly in the current situation [6].

Initially, depth maps in spatial and motion domains were used as features for recognition with SVM or ANN or HMM [7]. These methods were performing better than the traditional RGB video-based models due to ease in identification of the human subject which is at a variable depth compared to other video background objects. Despite their success, the depth data was inaccurately represented by the sensors for intra subject occultations and orientations during the action process [8].

The skeletal data has transformed the mediocre recognition human action systems into an exceptionally good system. The skeletal data was used in raw (x, y, z) position points [9] as features as inputs to different machine learning techniques. The raw skeletal Joints were used for action recognition as the points on a lie group [10], covariance Joint descriptors of 3D positions [11] and most frequently as the inputs to a LSTM networks [12,13]. Despite their better performances, LSTMs alone weren't able to decode spatio temporal patterns in raw position vectors of joints.

Hence, modeling joints as distance maps will bring the entire action sequence into a RGB image of the skeletal data. This map representation has improved both the recognition accuracy as well as training time latency [14,15]. These distance maps produce patterns on the image representing the underlying changes between Joint distances during an action sequence.

However, these maps contain distance variations from as low as a zero towards a very large value. Most of the previous study's used CNNs [16] to decode by using the joint distance maps (JDM) [17]. In order to extract meaning from the spatio temporal maps, we apply hybrid model of CNN followed by LSTM [18]. These hybrid networks have shown extraordinary accuracies related to skeletal based action recognition [19]. In this work, we apply serial CNN LSTM network which decodes the spatio temporal information embedded in the joint change distance image (JCDI). This work is different from previous study's in two ways:

1. The proposed network can model spatio temporal time series data from an image.
2. The training process is quite simple and faster due to fewer operations.

In the next section, we describe the proposed framework for developing an end to end action recognition system.

III. PROPOSED METHODOLOGY

The methodology has been developed in three phases as shown in figure1. The first phase is feature extraction for the four deep networks. The second and third phases are designing and training the proposed action recognition framework respectively. Two types of features are available from skeletal data: spatial and temporal. Spatial data features (SDF) are joint change distance features (JCD) represented as an RGB color image. Similarity, the temporal data features (TDF) are position vectors in the action video sequence represented as (x_1, y_1, z_1) to $(x_{20}, y_{20}, z_{20})(t)$. The SDF is the input to the CNN and the TDF becomes inputs to 3-layer LSTMs. The TDF are Joint trajectories represented as time series x, y and z axis data of the 20 joints on the human skeleton.

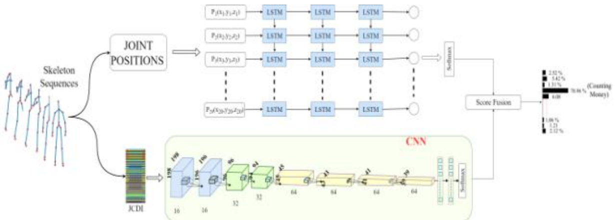
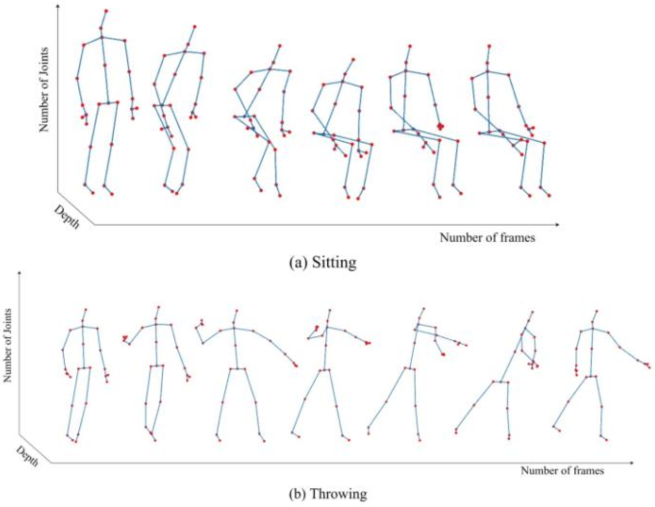


Fig.1. Flowchart of the proposed CNN-LSTM architecture

A. Feature Extraction

The actions of the human subject are embedded in the skeletal system represented by 20 joints. These joints are uniquely accessed through their location information in 3D world coordinates as (x_j, y_j, z_j), where j is joint number $\{j = 1\ to\ 20\}$. Consequently, spatial and temporal changes occur between the joints during an action. In order to obtain TDF, the position vectors form a set of time series x, y and z across actions. Fig.2 shows the time series TDFs for actions such as walking, running, jumping and waving. These time varying positions in an action are to be learned by a network that preserves the relationship between joints across action frames. Hence, we use LSTM for learning these time varying position vectors in an action.



(a) Sitting

(b) Throwing

Fig.2. Visualization of skeleton action sequence

However, these time varying joints find difficulty in modeling spatial features, which are equally effective in determining an action. Therefore, we create an RGB image that represents the space changes between joints in the action video.

The N^{th} frame of a 3D skeleton with J joints is being represented with $\frac{J(J-1)}{2}$ unique joint pairs where joint is a position vector in 3D space given by $(x_i, y_i, z_i) \in \mathbb{R}_{3\times J} \forall i = 1\ to\ J$. The joint change distance (JCD) in a frame is a vector with $\frac{J(J-1)}{2}$ entries. For N frames, the JCD is a matrix of size $N \times \frac{J(J-1)}{2}$. Intuitively JCD matrix characterizes the spatial changes in the skeletal action joints, which are then encoded into a color image of the same size to form joint change distance image (JCDI). However, for uniformity, the variable size JCDIs are resized to color images of uniform size of $\frac{J(J-1)}{2} \times N_u \times 3$, where N_u is the modified image resolution for training. The 3D labelled skeletal model used in our system for action recognition is shown in Fig. 3.

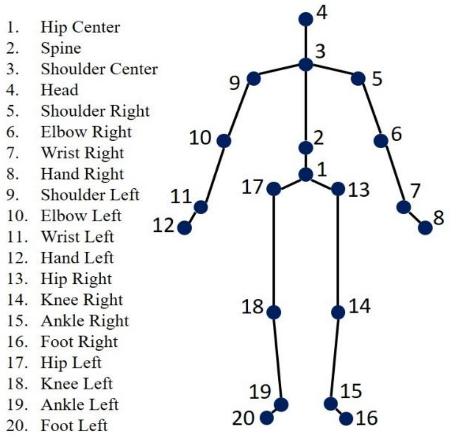


Fig.3. Skeleton joint representation

These J joints form a joint set with position vectors $p_i = [p_1, p_2, ..., p_J]$. A 3D human action sequence H on N frames is expressed as $H = \{P_1, P_2, ..., P_T\} \in \mathbb{R}_{J} \times 3 \times N$. Generally, the l_2 distance norm dn_{ij} between i^{th} and j^{th} joint pair in H [14] for n^{th} frame is formulated as

$$dt_{ij} = abs\left((P_i^n - P_j^n)^2\right) \quad (1)$$

Finally, the color-coding is done simply by using the 'jet' color map to encode the JCDs with the following standard mapping procedure in [15] to create JCDI spatial color maps for the CNN. Fig.4 shows some color coded JCDIs of common actions across datasets. The next section gives an insight into the architecture of the proposed CNN-LSTM combination.

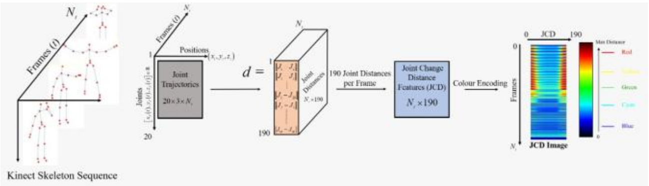


Fig.4. Color encoding process

B. ML Architecture

Unlike the previous models [14,15] where only CNN or LSTMs were used for action recognition, this paper proposes a CNN-LSTM integrated network named as Memory CNN or MCNN. The proposed MCNN is being adopted to process different types of features. The CNN compliments LSTM in a way to improve their discovery on cross subject and cross view data. LSTMs have shown the ability to effectively process temporal data, whereas CNNs are good at decoding spatial changes. Thus, improving the performance of the complimentary network MCNN.

Traditional RNNs perform well for discovering short term relationships in time series data. However, for long term relationships, LSTMs have shown better performance over RNNs. In this work, we propose to design our network with LSTMs. A LSTM block consists of 4 gates, an output state and a memory unit state. A schematic of LSTM block is shown in figure 5. The state transition expression of an LSTM block is

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ u_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \right) \quad (2)$$

$$c_t = i_t \circ u_t + f_t \circ c_{t-1} \quad (3)$$

$$h_t = o_t \circ \tanh(c_t) \quad (4)$$

where the parameters $i_t, f_t, o_t, u_t, h_t, c_t$ denote input gate, forget gate, output gate, input modulation gate, output state and internal memory cell state respectively [20]. The input to the network at time step t is given by variable x_t and the symbol $\circ$ points to an element-wise product operation. The symbol σ denotes the sigmoid function.

The LSTMs in the proposed framework forms a three layer network as shown in figure1. The input to the 1^{st} set of LSTMs is a 3D vector at x_t. The output of 1^{st} LSTM layer i.e. h_t becomes input to the 2^{nd} LSTM and its output as input to the 3^{rd} stage LSTM. Finally, the output of all 3^{rd} stage LSTMs feed the softmax that translates the coded outputs to probability scores for class identification.

In this paper the CNN architecture has been inspired from shallow VGG – 16 [14]. The CNN has 8 convolutional layers, 2 fully connected and a softmax layer. The filter resolutions were constant at 3×3, whereas the number of filters increased from 16 to 64 respectively. In between two convolutional layers, a max pooling is introduced with a factor 2. In the end we have dropouts to generalize the feature vector.

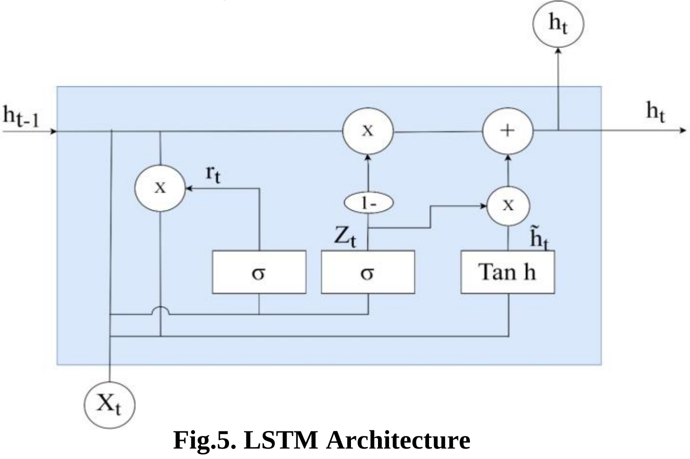


Fig.5. LSTM Architecture

C. Training and Score Fusion

The goal of training is to optimize the multinomial logistic regression objective function by means of mini-batch gradient descent with momentum of 0.5. The penalization multiplier is set to 0.002 to regularize weight decay during training for a video frame size of 200. The RGB JCDIs are resized to 200×200 using down sampling. The max pooling in the CNN stream is set to 0.5 for all 8 conv net layers. Weights in each layer are randomly initialized with a gaussian distribution function with zero mean and 0.01 variance for all layers. Intermediate validation showed overfitting in the dense layers during training for all the datasets. This is handle by inducing dropout of 0.5 before the dese layer, which improved the validation score.

For the three LSTMs the training algorithm used was Root Mean Square Propagation with a max dropout of 0.5. The input skeletal vectors are grouped into batches of 20 frames or sub sequences to feed into LSTMs. The initial learning rate was set at 0.001 and was decreased by 10% whenever there the error became constant. The time step on LSTM was set to 20 in each epoch. Finally, the outputs of the SoftMax layers in the CNN and LSTM streams are fused using average the similarity scores for making a decision on the output class labels.

IV. RESULTS AND DISCUSSION

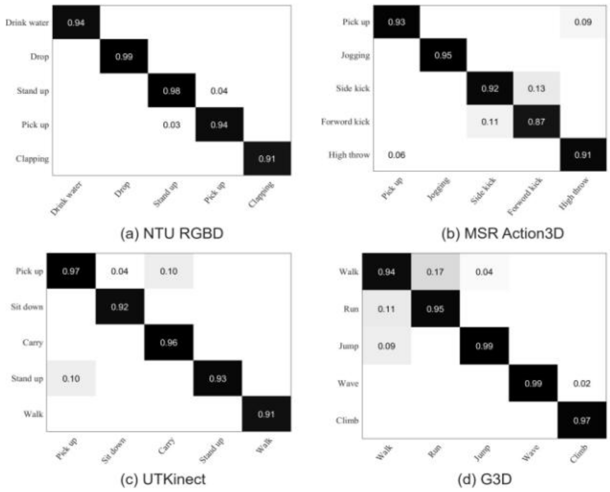
The proposed Memory CNN (MCNN) network was evaluated over skeletal action data in publicly available datasets such as NTU RGB-D, MSR Action 3D, UTKinect and G3D. Experimentation on the proposed framework were conducted using the following datasets to evaluate the effectiveness in recognizing actions with respect to cross subject and cross view. We also check the robustness of the proposed hybrid framework against singular CNNs and LSTMs.

A. Action Datasets

All the datasets were recorded using Microsoft Kinect producing RGB, depth and skeletal video frames of 3D human action. However, in this work the focus will be on using skeletal data only. The NTU RGB D has a rich variety of actions recorded with 40 subjects in 5 multiple viewing angles consisting of around 56k video samples [21]. MSR Daily Action 3D is an action dataset with 20 actions collected from 10 subjects, with a total of 567 videos [22]. UTKinect action dataset has 10 actions from 10 subjects [23]. Finally, the G3D dataset contains 10 subjects performing 20 gaming actions such as punch right, punch left, kick right, kick left, defend, golf swing, tennis swing forehand, tennis swing backhand, tennis serve, throw bowling ball, aim and fire gun, walk, run, jump, climb, crouch, steer a car, wave, flap and clap. This is the most naturalistic dataset with 80,000 frames of skeletal video data, where the subjects were free to perform an action based on their comfort [24]. Each action was performed by 10 subjects and was repeated 3 times. All datasets are available with 20 joints on the skeleton. The hardware used in all experiments has been a 8GB GPU from NVIDIA with a 16GB memory running on 2.4GHz Intel core process.

B. CNN LSTM Performance

Firstly, we evaluate the performance of the proposed hybrid CNN LSTM architecture against the state of the arts CNN, LSTM and deep learning frameworks. The evaluating criteria is based on recognition percentage, precision and recall on each of the datasets. Testing the networks was based on one left training model, where one sample was set aside for testing and remaining were used for training. In this experiment only cross subject testing was performed. Training of each network chosen has been accomplished using 5 subjects and tested with remaining subjects, one at a time. Figures 6 shows the confusion matrices for few actions in the datasets using the proposed MCNN architecture. The averaged results of the experimentation were presented in table 1 (table 2 in word).



(a) NTU RGBD

(b) MSR Action3D

(c) UTKinect

(d) G3D

Fig.6. Confusion matrix of the proposed method on (a) NTU RGBD dataset (b) MSR Action 3D dataset (c) UTKinect action dataset (d) G3D dataset.

The comparison set is a combination of CNN and LSTM models from literature. The models VGG, AlexNet, ImageNet and CNN are based on convolutional neural network which were imputed with JCDI RGB images. The recurrent models tested were RNN and LSTM with time series skeletal data from 20 frames per step. Finally, our proposed MCNN uses both the inputs and average score fusion. The table 1 shows that the proposed framework performs better when compared to the other deep networks. This is because of the hybrid architecture which decodes spatial and temporal information embedded within the action sequence. In contrast with the other methods which use only one type of features for recognition.

Table-1
Performance of proposed algorithm against state-of-the-art algorithms

Datasets	Networks	Precision	Recall	Recognition
NTU GRB-D	VGG	82.1	81.4	84.3
	AlexNet	79.1	78.6	80.7
	ImageNet	77.6	77.8	79.2
	RNN	81.9	82.7	83.6
	CNN	85.2	85.4	86.2
	LSTM	84.1	83.5	85.1
	Proposed	86.7	86.1	87.6
MSR Action 3D	VGG	87.1	86.6	86.4

	AlexNet	82.9	82.4	83.1
	ImageNet	81.2	80.6	82.1
	RNN	84.2	83.9	85.6
	CNN	86.1	85.4	87.3
	LSTM	86.4	86.2	87.1
	Proposed	88.9	88.3	89.2
UT-Kinect Action 3D	VGG	92.1	92.7	92.9
	AlexNet	87.4	86.9	90.4
	ImageNet	86.4	86.7	90.1
	RNN	89.5	89.2	91.1
	CNN	92.4	92.5	93.4
	LSTM	91.9	91.7	92.7
	Proposed	93.7	93.1	94.5
G3D	VGG	89.7	89.2	88.4
	AlexNet	84.2	84.7	85.2
	ImageNet	83.6	83.2	85.7
	RNN	87.8	87.1	88.7
	CNN	91.1	90.7	90.9
	LSTM	89.4	89.1	89.9
	Proposed	92.1	91.7	92.4

C. Spatial Maps Evaluation

In this study, spatial representation of skeletal information is being evaluated keeping the temporal data constant. Table 2 shows the recognition rates on all datasets using various methods of representing spatial skeletal data. It shows raw joints, edges, surfaces and joint angular displacement maps (JADM) against our proposed JCDIs. The experimentation was done on our network to test its performance across subjects and views. The recognition percentages indicate that the proposed JCDIs are a better choice along with time series data for MCNN architecture. The superiority of JCDIs can be attributed to the change representation among joint changes across action frames compared to direct modeling of distance features in the remaining representations. Hence, JCDIs model spatial changes in data to accommodate joint variations in space.

Table-2
Comparison of different skeleton features on proposed algorithm

Datasets	% Recognition									
	Raw Skeleton Data		Edge		Surface		JADM		JCDI	
	Cross view	Cross Subject	Cross view	Cross Subject	Cross view	Cross Subject	Cross view	Cross Subject	Cross view	Cross Subject
NTU RGB-D	67.87	72.49	70.28	74.68	71.25	75.31	81.15	89.92	85.74	90.12
MSR Action 3D	71.32	74.84	73.82	76.82	74.11	76.28	82.96	90.91	87.15	91.25
UT-Kinect Action 3D	77.27	80.84	79.37	82.37	80.14	83.67	87.27	95.37	92.77	96.47
G3D	74.94	76.98	76.92	81.59	77.95	82.17	85.37	94.14	90.59	94.79

The results also indicate higher recognition accuracies for cross view data from our JCDIs as against other maps.

D. Cross data evaluation

Table 3 gives an insight into the behaviour of the proposed MCNN towards generalizing an action. Hence a cross data evaluation is being conducted for common actions in all datasets. The recognition rates shown by the proposed framework are found to be of good interest and are better than the minimum average rates.

Table-3
Cross data validation using proposed CNN-LSTM architecture

Training Dataset	Testing Dataset	% of Recognition
NTU RGB-D	MSR Action 3D	83.4
	UT-Kinect Action 3D	85.7
	G3D	84.2
MSR Action 3D	NTU RGB-D	84.2
	UT-Kinect Action 3D	84.9
	G3D	83.4
UT-Kinect Action 3D	NTU RGB-D	85.9
	MSR Action 3D	84.6
	G3D	85.1
G3D	NTU RGB-D	83.4
	MSR Action 3D	84.7
	UT-Kinect Action 3D	85.2

E. Deep model architecture evaluation

Finally, a verity of deep learning frameworks from literature are compared against our proposed MCNN network on various aspects with respect to the considered datasets. Table 4 gives the recognition rates from literature on the considered datasets using CNN and LSTM based architectures. In all, the hybrid model proposed in this work has shown to outperform the state of the arts from literature. The proposed method recorded an average recognition of 84.77% on all challenging datasets, which is better than other deep learning frameworks.

Table-4 Comparison of proposed method against CNN and LSTM based architectures

	NTU RGB-D		MSR Action 3D	UT-Kinect Action 3D	G3D
	Cross view	Cross Subject			
LSTM [25]	76.5	82.3	85.9	91.7	89.9
2 Stream LSTM [26]	78.3	83.7	86.7	92.1	90.4
Multi Stream LSTM [27]	80.2	84.1	87.6	92.4	91.7
CNN [15]	79.7	83.9	86.9	92.1	90.9
2 Stream CNN [28]	81.3	85.1	87.1	92.9	91.4
Multi Stream CNN [29]	82.4	86.7	87.4	93.9	91.7
Proposed (CNN+LSTM)	85.7	90.1	89.2	94.5	92.4

V. CONCLUSION

This paper proposes an effective approach for 3D skeleton based action recognition. The system uses a hybrid deep learning framework with CNN and LSTM score fusion. Consequently, the architecture takes advantage of LSTM and CNN models for recognition. The CNN model decodes spatial information and LSTMs show their might in modeling temporal data. Average score fusion has shown highest recognition accuracies on the hybrid Memory CNN. State of the art results were achieved on skeletal data in challenging datasets such as NTU RGBD, MSR Action 3D, UTKinect and G3D which have verified the effectiveness of the proposed hybrid framework.

REFERENCES

1. Turaga, P., Chellappa, R., Subrahmanian, V.S. and Udrea, O., 2008. Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video technology, 18(11), p.1473.
2. Dawn, D.D. and Shaikh, S.H., 2016. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. The Visual Computer, 32(3), pp.289-306.
3. Vezzani, R., Baltieri, D. and Cucchiara, R., 2010, August. HMM based action recognition with projection histogram features. In International Conference on Pattern Recognition (pp. 286-293). Springer, Berlin, Heidelberg.
4. Qian, H., Mao, Y., Xiang, W. and Wang, Z., 2010. Recognition of human activities using SVM multi-class classifier. Pattern Recognition Letters, 31(2), pp.100-111.
5. Simon, C., Meessen, J. and De Vleeschouwer, C., 2010. Visual event recognition using decision trees. Multimedia Tools and Applications, 50(1), pp.95-121.
6. Papadopoulos, G.T., Axenopoulos, A. and Daras, P., 2014, January. Real-time skeleton-tracking-based human action recognition using kinect data. In International Conference on Multimedia Modeling (pp. 473-483). Springer, Cham.
7. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z. and Campos, M.F., 2012, September. Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In Iberoamerican congress on pattern recognition (pp. 252-259). Springer, Berlin, Heidelberg.
8. Chen, C., Jafari, R. and Kehtarnavaz, N., 2017. A survey of depth and inertial sensor fusion for human action recognition. Multimedia Tools and Applications, 76(3), pp.4405-4425.
9. Hou, Y., Li, Z., Wang, P. and Li, W., 2016. Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology, 28(3), pp.807-811.
10. Chen, H.S., Chen, H.T., Chen, Y.W. and Lee, S.Y., 2006, October. Human action recognition using star skeleton. In Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks (pp. 171-178). ACM.
11. Hussein, M.E., Torki, M., Gowayyed, M.A. and El-Saban, M., 2013, June. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In Twenty-Third International Joint Conference on Artificial Intelligence.
12. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L. and Xie, X., 2016, March. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In Thirtieth AAAI Conference on Artificial Intelligence.
13. Liu, J., Shahroudy, A., Xu, D. and Wang, G., 2016, October. Spatio-temporal lstm with trust gates for 3d human action recognition. In European Conference on Computer Vision (pp. 816-833). Springer, Cham.
14. Maddala, T.K.K., Kishore, P.V.V., Kumar, K. and Kumar, A., 2019. YogaNet: 3D Yoga Asana Recognition Using Joint Angular Displacement Maps with ConvNets. IEEE Transactions on Multimedia.
15. Kumar, E.K., Kishore, P.V.V., Sastry, A.S.C.S., Kumar, M.T.K. and Kumar, D.A., 2018. Training CNNs for 3-d sign language recognition with color texture coded joint angular displacement maps. IEEE Signal Processing Letters, 25(5), pp.645-649.
16. Du, Y., Fu, Y. and Wang, L., 2015, November. Skeleton based action recognition with convolutional neural network. In 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) (pp. 579-583). IEEE.
17. Kumar, D.A., Sastry, A.S.C.S., Kishore, P.V.V., Kumar, E.K. and Kumar, M.T.K., 2018. S3DRGF: Spatial 3-D Relational Geometric Features for 3-D Sign Language Representation and Recognition. IEEE Signal Processing Letters, 26(1), pp.169-173.
18. Lee, I., Kim, D., Kang, S. and Lee, S., 2017. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1012-1020).
19. Nunez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S. and Velez, J.F., 2018. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition, 76, pp.80-94.
20. Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou and Wanqing Li, "Skeleton-based action recognition using LSTM and CNN," 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, 2017, pp. 585-590.
21. Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, “NTU RGB+ D: A large scale dataset for 3D human activity analysis,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019.
22. Mining Actionlet Ensemble for Action Recognition with Depth Cameras, Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan, IEEE Conference on Computer Vision and Pattern Recognition (CVPR2012), Providence, Rhode Island, June 16-21, 2012.
23. Li. Xia, C. Chen and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, 2012, pp. 20-27.
24. V.Bloom, D. Makris and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, 2012, pp. 7-12.
25. Liu, J., Shahroudy, A., Xu, D., Kot, A.C. and Wang, G., 2017. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. IEEE transactions on pattern analysis and machine intelligence, 40(12), pp.3007-3021.
26. Liu, J., Wang, G., Hu, P., Duan, L.Y. and Kot, A.C., 2017. Global context-aware attention LSTM networks for 3D action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1647-1656).
27. Zhang, S., Yang, Y., Xiao, J., Liu, X., Yang, Y., Xie, D. and Zhuang, Y., 2018. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. IEEE Transactions on Multimedia, 20(9), pp.2330-2343.
28. Liu, H., Tu, J. and Liu, M., 2017. Two-stream 3D convolutional neural network for skeleton-based action recognition. arXiv preprint arXiv:1705.08106.
29. Ding, Z., Wang, P., Ogunbona, P.O. and Li, W., 2017, July. Investigation of different skeleton features for CNN-based 3D action recognition. In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 617-622). IEEE.

AUTHORS PROFILE

Mr. Kommani Vijaya Prasad received the B.E degree in Electronics and Communication Engineering from Osmania Univeristy, Hyderabad, India, M.Tech degree from JNTU Ananthapuram, India, specializing in evolving optimized object segmentation and recognition, and pursuing the Ph.D. degree from JNT University Kakinada, Kakinada, Andhra Pradesh, India. His research interests are human machine interaction, Machine learning and computer vision and its applications.

Dr. P. V. V. Kishore is a professor of Image & Video Processing with the department of Electronics and Communications Engineering, where he manages the Image, Speech and Signal processing Research Group. He went on to study M.Tech at Cochin University of science and technology and Ph.D. from Andhra University College of engineering in 2013. He is the chair of the Biomechics and vision computing research center.

His works focus on mechine learing, biomechanics, artificial intelligence, human motion analysis and sign language machine translation. His research explores how motion capture data models can effectively model low end video objects in real time for better recogntion and analysis. He is particularly intersted in developing new innovations in the areas of computer vision and mechine learing. He has authored several publications in these fields.

Dr. O. Srinivasa Rao did B.Tech, M.Tech and obtained Ph.D in CSE from JNTUK, KAKINADA. His Ph.D specialization is cryptography and Network security. He presented more than 60 research papers in various International journals and two research papers in National conferences, one research paper in international conference. He had more than 20 years of teaching experience and he was former Head of CSE at University College of Engineering vizianagaram, JNTUK and currently working as Professor of CSE at University College of Engineering, JNTUK, Kakinada. He guided one Ph.D and more than 90 M.Tech and MCA students' projects. Currently he is guiding 4 Ph.D and 8 M.Tech, 2 MCA students' projects. His fields of interest are Cryptography, Network security, Image Processing and Data Mining.