Multi Modal RGB D Action Recognition with CNN LSTM Ensemble Deep Network

D. Srihari[1] , P. V. V. Kishore[2]
Department of ECE, Koneru Lakshmaiah Education Foundation,
Guntur (DT), Andhra Pradesh, India

***Abstract*—Human action recognition has transformed from a video processing problem into multi modal machine learning problem. The objective of this work is to perform multi modal human action recognition on an ensemble hybrid network of CNN and LSTM layers. The proposed CNN - LSTM ensemble network is a 2 - stream framework with one ensemble stream learning RGB sequences and the other depth. This proposed framework can learn both temporal and spatial dynamics in both RGB and depth modal action data. The hybrid network is found to be receptive towards both spatial and temporal fields because of the hierarchical structure of CNNs and LSTMs. Finally, to test our proposed model, we used our own BVCAction3D and three RGB D benchmark action datasets. The experiments were conducted on all the datasets using the proposed framework and was found to be effective when compared to similar deep learning architectures.**

***Keywords*—Human action recogniiton; RGB D video data; convolutional neural networks; long short-term memory**

I. Introduction

Human action recognition is basically considered as a computer vision problem where a set of video processing algorithms were proposed to extract features that became input to a classification algorithm. However, these video processing algorithms depended heavily on the orientations of pixels in the video frames which affected the performance of the classifier as whole. Despite their instabilities in generalizing the classifiers performance the human action recognition are applied in surveillance networks, industrial automation, medical and sports analysis to name a few. In contrast to RGB video sensors, we now have low cost multi modal sensors such as Microsoft Kinect, that can enhance RGB sequences with depth and skeletal information. On the other hand the progress of deep learning algorithms like Convolutional Neural Networks (CNNs) and Recurrent models (RNNs) has been instrumental in enhancing the performance of multi modal action recognition systems.

In the recent years deep learning architectures have been shown to learn and complement the unique features in RGB, depth and skeletal data for performance improvements in action recognition tasks [1], [2]. Specifically, the work in [3] shows the effectiveness of using auxiliary datasets in the form of skeleton and depth has enhanced the accuracy of action recognition system using RGB videos. Multiple Kernel based learning framework was applied effectively on RGB D action data for extracting multi modal features and further fusing them, which improved their accuracy positively [4]. Further, a few of these models explored the sparse modelling of dense RGB and depth features that were translated into weighted bag of words (BOW) [5] representation for classification. Most of the works experimented with full action sequences ignoring the temporal information accompanying the action.

Initially, the idea was to extract motion information from RGB and depth sequences using optical flow, Kalman tracking and sometimes packing motion into a single image called as motion history images (MHI) [6]. Even though these methods offered an improvement in performance of the classifiers, they showed difficulty in learning spatio temporal features for generalizing an action. The fundamental difficulty in multi modal sequences is the formation of a multi-dimensional tensor indexing modalities, their spatial and temporal knowledge in one field. Subsequently, learning and temporal pooling operations on this multi-dimensional tensor is a challenging task. Moreover, time varying modalities will always induce constraints due to variable length effects. Despite the above gaps in data acquisition and processing, the time varying multi modal features can enhance the performance of the learning algorithms. However, the question posed at this instance is how to teach a classifier the spatio temporal modalities for RGB D action recognition tasks.

Previously, we approached the above problem by dividing multiple modalities to fixed length action sequences which are then arranged as a multi layered multi modal tensor. These multi-dimensional tensors are processed through deep convolutional neural networks (CNN) for learning spatial representations thereby completely ignoring the temporal structures [7].

In this paper, we propose to develop a hybrid recurrent CNN based deep learning framework for multi modal action recognition from RGB and depth data. Our proposed CNN LSTM network has been an inception of recurrent CNNs for action recognition in [8]. However, it is different from models proposed previously on multi modal action sequences [9], [10], [11], [12] in two aspects. One, the multi modelled data used in our work is RGB and depth sequences and two, our proposed CNN-LSTM Action Network (CLANet) is an ensemble of streams of layers.

The CLANet extracts the spatial features from RGB and depth sequences using CNN and infuses the extracted features into the LSTM network which is bidirectional in structure. The LSTM streams learn the temporal patterns in both the RGB and depth sequences during training. The last layer of the CLANet is a dense layer with SoftMax activation the outputs of which are score fused to decide on the input class. We opted for RGB and depth sequences for this experimentation and no skeletal action inputs due to the data dimensionality representation between them. Skeletal data has a higher dimensionality over the RGB and depth which share a common representation.

In order to validate our proposed framework, we have our own BVC3DA RGB D action dataset with 40 actions from 10 different actors with 10 repetitions per action. However, we evaluated the proposed framework on benchmark RGB D datasets, NTU RGB D, MSRAction and UTKINECT to test the learning strategies of the proposed CNN LSTM network.

The rest of the paper is organized as follows. The second section presents the previous works on RGB D multi modal action recognition with an insight into gaps and the achieved breakthroughs. Section three discusses the methods applied in this study to achieve higher performances across datasets for multimodal action recognition. Results are presented in section four and conclusions drawn on the obtained results in section five.

II. Background

Multi modal or RGB D based action recognition models has been studied extensively which led to the development of the proposed CLANet. The previous methods have shown to have used data with RGB frames, depth and skeletal information for recognition of human actions across multiple identification platforms such as machine learning [1] to deep learning [13]. The machine learning models apply segmentation and feature extraction algorithms on RGB or depth or both frames for extracting meaningful representations of actions [14]. On the other hand, deep learning models extract features and segments based on the training algorithms on the RGB D data sequences [9]. The most formidable of these deep learning models are grouped into spatial and temporal domains. In spatial domain the models extract features with respect to the pixel location in image space using models such as Convolutional Neural Networks (CNNs) [2], [7]. For temporal or time series modelling of the RGB D data, Recurrent neural networks (RNNs) and their upgrades such as Long Short-Term Memory (LSTM) nets [15], [16].

However, the spatial and temporal models have their share of advantages and disadvantages. The spatial models cannot effectively learn the time series information which is necessary to represent action sequences that dependent on continuous data variations. Contrastingly, using exclusive time series modelling on video frames will not capture the spatial representations of action movements in image spaces. Hence, a hybrid combination involving both spatial and temporal models in found to be necessary to represent actions in video sequences for recognition [17], [18]. The early models applied optical flow to extract the temporal features on RGB video frames which are further fused with the spatial features during the training of CNNs. A few state-of-the-art models used multiple streams of independent CNNs with inputs from RGB and optical flow based RGB giving satisfactory results [19], [20]. One stream of CNN used RGB spatial features and the other uses motion information during training the networks simultaneously. All these networks are accompanied with feature fusion layer before or after the dense layers for decision making on the inputted action sequence. However, these models require additional computation time in the form of motion vectors which makes them computationally inefficient due to data alignment problems. Moreover, few also tried 4 streams by adding motion information from depth sequences producing better recognition accuracies than the previous 2 stream model [7].

Similar to the above models, properties of the RGB and depth modalities have produced efficient action recognition algorithms such as depth rank pooling with CNNs [21], scene flow based RGB D channels on CNN [22] and sequence based methods with RNNs [23]. However, the most successful are models that combine the advantages of both spatial and temporal networks. These models are named as spatio temporal recurrent convolutional neural networks (rCNNs) [24]. These models operate in twofold: one, the primary network extracts the spatial features using CNNs and the secondary network encodes that spatial features into temporal data using recurrent models. The most frequently applied recurrent model was Long short-term memory (LSTM) for representing temporal information in the action video sequences due to their ability handle long term dependencies by avoiding gradient vanishing problems [25]. Consequently, it was found that the operating the feature pooling model with LSTM can influence the temporal learning capabilities of the hybrid CNN LSTM architectures. Through feature sharing mechanism between the two networks, they were able to produce higher level representations of actions in a video sequence [26]. Moreover, bidirectional LSTM based methods have shown to handle multiple length video sequences when compared to RNNs. Therefore, the hybrid combination of CNN and LSTMs is the most widely applied model for human action recognition because of their abilities to decode spatial and temporal information simultaneously [27].

Literature is filled with CNN LSTM models for action recognition using skeletal actions as inputs [28], [29], [30]. These models use 3D skeletal joints as time series data along with RGB video frames for training and testing. However, depth-based models were rarely used along with these hybrid models [31]. In this work, we try to learn through a hybrid model which uses both RGB and depth data to draw inferences on the input action sequences. Both CNNs and LSTMs allow end to end trainable models that eliminates the need for tracking variations through time series data. The advantages of using RGB inputs along with depth instead of skeletal data are threefold. First, the depth features are more profound in assisting the spatial information in RGB data when compared to skeletal data. Second, the depth data is analogous to RGB data, which allows for complex processing mechanisms in transforming the skeletal data to image data. Finally, the skeletal data at times is found to be noisy with missing joints or overlapping joints making it difficult to process.

Eventually, in this work we describe a hybrid framework by combining LSTMs with CNNs for action recognition called as CLANet to construct an end-to-end trainable architecture that has capabilities in handling visual action recognition and sequence prediction tasks.

III. Methodology

This section provides a detailed description of the proposed CLANet hybrid CNN - LSTM architecture for action recognition. First, we design a deep CNN model to extract RGB and depth features of multiple frames to generate spatial features in the considered RGB and depth modals respectively. Then we will build an end-to-end pipeline architecture by combining multi modal CNNs with bi-directional LSTMs, followed by a

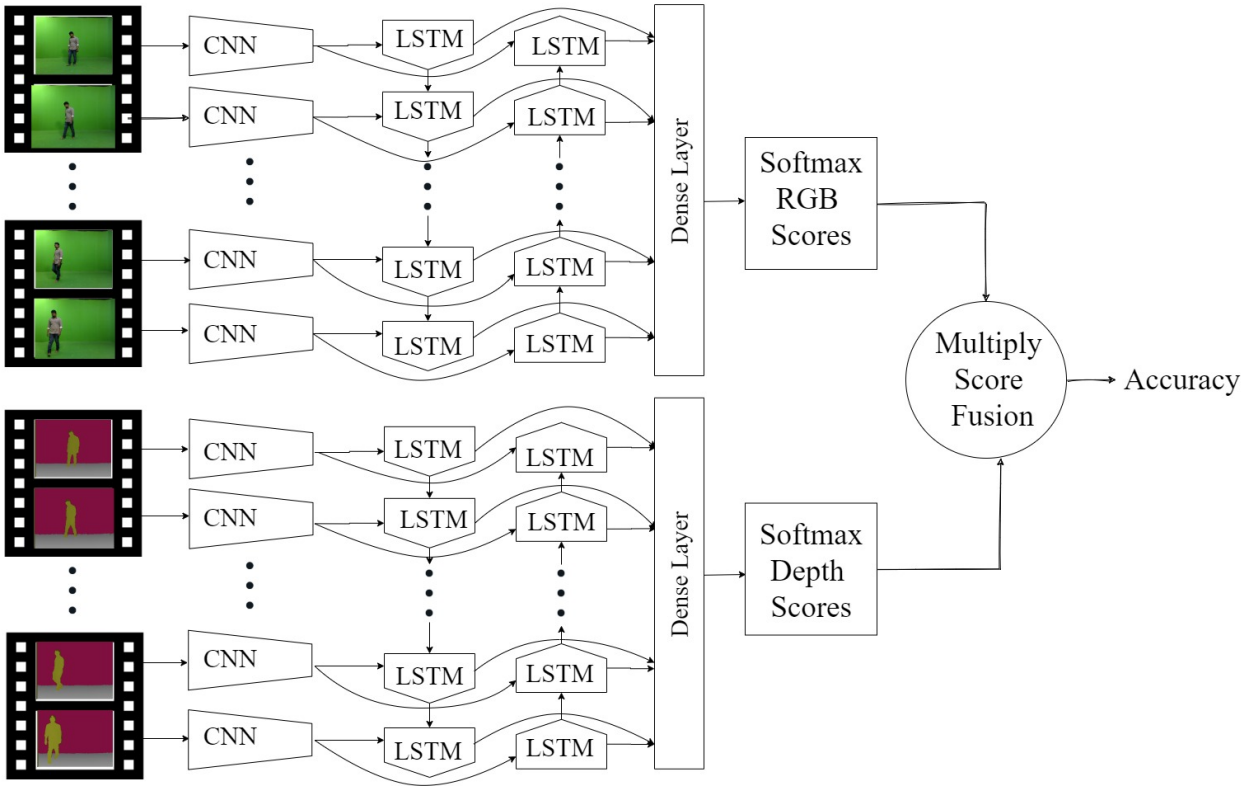

Fig. 1. Proposed CLANet Architecture for RGB D action recognition.

multiply score fusion to estimate the actions. The proposed architecture is shown in Fig. 1.

A. The Spatial CNN Network

This subsection describes in detail the architecture for extracting spatial information from RGB and depth video frames. To accomplish this, we employed convolutional neural networks in multiple streams that take input as RGB and depth video frames. Based on the GPU memory, we found that the maximum number of streams that can be applied in a batch is 16. Hence, the first hyperparameter selected was batch size which is set to 16. Hence each ensemble of CNNs will feed into 16 frames of RGB and depth frames. Lets name the two ensembles are CRGBe and CDe. The CRGBe and CDe are multi stream ensembles of CNNs for RGB inputs and depth inputs, respectively. Fig. 2 shows the CNN architecture developed for extracting spatial features from RGB video frames. Consequently, we have a similar network, CDe for processing depth frames.

Given an RGB action video frame $V_{rgb}\left(v_r, v_g, v_b\right)$ with a pixel position of (x, y), the output of 2D convolutional kernels are feature maps. Eventually, the j^{th} feature map from the i^{th} convolutional layer is extracted using the expression

$$F_{ij}\left(x,y\right) = f\left(\sum_{p}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\left(W_{ijp}^{nm} * V_{(i-1)p}\left(x+n, y+m\right)\right) + b_{ij}\right) \quad (1)$$

Where, $N \times M$ is the size of the video frame $V$and f is the activation function. W_{ijp}^{nm} is the weight vector at position (n, m) associated with p^{th} feature map in the $(i-1)^{th}$ layer of the CNN network. The parameter b_{ij} is the bias associated with each of the neurons. Eq'n(1) depicts the convolutional operation between the video frames and the weight matrix, which is updated sequentially during training of the network. There are 16 streams in CRGBe ensemble network to extract spatial features of 16 consecutive frames per action video. To maintain uniformity, we divided each action class video into 128 frames. That is there will 8 batches of RGB video frames per class for training on the CRGBe network. Subsequently, the depth network, CDe will also have the same configuration as CRGBe. The CDe extracts the depth spatial features from the depth sequences of actions.

As shown in Fig. 2, the architecture for RGB spatial feature extraction module with RGB input video frames. The CRGBe is an ensemble of 16 streams with a depth of 10 layers each. The 10 layers depth across each stream consists of 6 convolutional plus ReLu layers, 3 max pooling layers and a flatten layer. The filter kernels are selected as 7×7, 5×5 and 3×3 framework. This kernel selection framework has ensured a hierarchical feature extraction model that has ensured maximal spatial preservation of pixels towards the end of the network. Similar functionality is achieved on depth frames using CDe network. The spatial maps from CRGBe and CDe are now used for modelling temporal information in the features by passing them through LSTM module. The LSTM module is presented in the following section.

B. The LSTM Temporal Coding

The extracted spatial features from the two ensemble nets, CRGBe and CDe, are then temporally coded for recognition

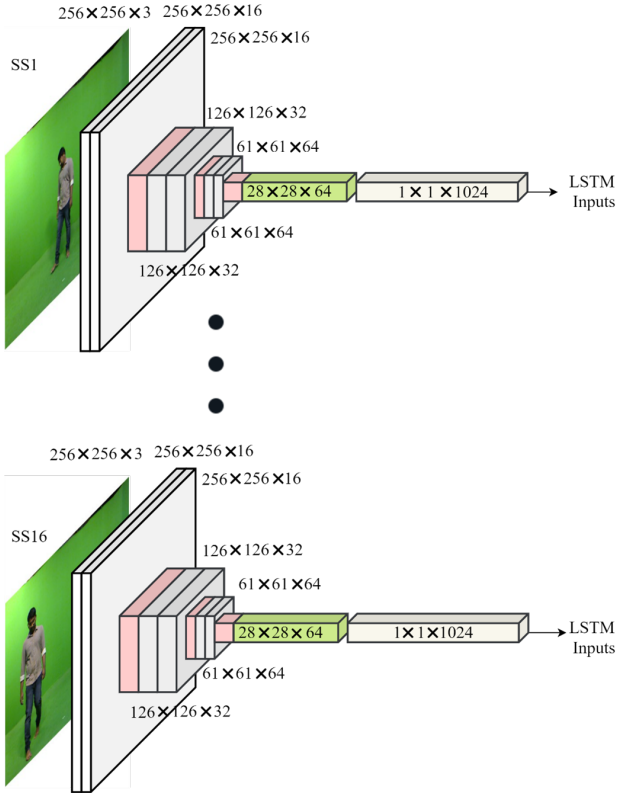

Fig. 2. The CRGBe ensemble for extracting spatial features from RGB action video frames.

of actions at the highest level. LSTM blocks provide temporal dynamics for the extracted spatial features across both the input modalities. Fig. 3 illustrates the single LSTM block used in this work. The following expressions are implemented during the operation of an LSTM block. It consists of an input unit I_t, forget gate F_t, output gate O_t, momentum factor G and the LSTM cell outputs (C_t, h_t).

$$I_t = \sigma\left(\left(x_t + h_{t-1}\right) W_I + b_I\right) \tag{2}$$

$$F_t = \sigma\left(\left(x_t + h_{t-1}\right) W_F + b_F\right) \tag{3}$$

$$O_t = \sigma\left(\left(x_t + h_{t-1}\right) W_O + b_O\right) \tag{4}$$

$$G = \tanh\left(\left(x_t + h_{t-1}\right) W_G + b_G\right) \tag{5}$$

$$C_t = C_{t-1} * F_t + G * I_t \tag{6}$$

$$h_t = O_t * \tanh\left(C_t\right) \tag{7}$$

Where, W and b are weights and bias. x_t are the feature inputs extracted using the spatial CNN network.C_t andh_t are LSTM's cell state at time step t. The sigmoid σ acts as control gates for transfer of inputs to the outputs. The forget gate initiates the progress of inputs to the next LSTM block. Based on the state of forget gate, the LSTM cell either forgets or memorizes the features in a sequence. However, the flow is unidirectional in a single stream LSTM model. In general, a sequence labelling problem such as video-based action recognition we need access to the past and future inputs at a single time step during the training sequence. This is found to be achievable in the past using bidirectional LSTM networks as shown in Fig. 1. This is performed by two LSTM streams with one moving past data forward and the other moving the future data backwards for a specific time step. This biLSTM network is also trained using the same backpropagation through time algorithm. In our work, we performed the backward and forward passes for each action sequence. Subsequently, the hidden states of LSTMs were reset after each action class. Our work uses a bidirectional LSTM architecture from [25]. The following subsection describes the complete multi modal action recognition framework with bidirectional LSTM network on top of CNN networks.

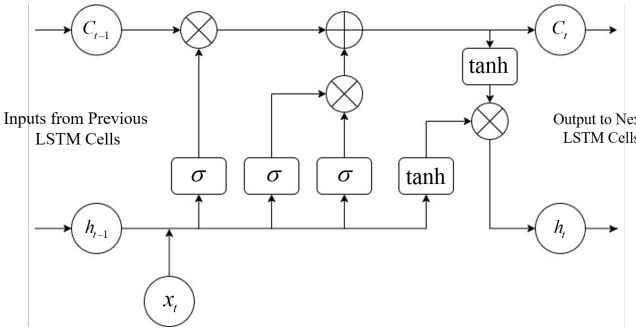

Fig. 3. A Single LSTM Cell Architecture

C. The Hybrid CLANet Training

The hybrid CLANet is designed by stacking bidirectional LSTM cells on top of spatial CNNs to create an end-to-end trainable model. The CNNs are capable of extracting global and highly discriminating spatial features from the RGB and depth video frames. On the other, LSTM capture the local and time representations in the extracted features. Finally, the outputs of LSTM network is passed through a dense layers and a SoftMax layer to compute the probabilistic distribution of the class labels as

$$Y_{class} = SoftMax\left(h_t\right) \tag{8}$$

In the proposed bidirectional LSTM, the hidden states from forward pass and backward pass are combined in the output dense layer. We used 2 dense layers of sizes 1024 each along with a SoftMax to compute the recognition scores. The validation losses are calculated after the first dense layer to update the weights and biases through backpropagation. The validation data is 15% of the total training data and cross entropy loss is used for error calculation. The hyper parameters such as weights and biases are selected randomly with zero mean random gaussian generator. Stochastic gradient descent algorithm is used calculating the losses during training with an initial learning rate of 0.001 across all datasets. However, the learning rate is readjusted, whenever the loss became constant during training. The entire CLANet is end-to-end trainable.

D. The CLANet Testing

The action datasets are divided into 65% training, 15% validation and 20% testing. The outputs of the network give a probability distribution across classes for a particular test input sample. We used multiple machines for training and testing at different frame rates to understand the characteristics of the CLANet in processing multi model spatio temporal data.

Finally, we perform multiple test mechanisms on our own BVRCAction3D dataset and other benchmark datasets such as MSRDailyActivity3D, UTKinect and NTU RGB D.

IV. Results and Analysis

This section presents results of experimentation with analysis of various components that were instrumental in generating the results on various datasets. We start by describing the datasets for training, validation and testing. Next, we initiate the training and testing of the proposed CLANet across different actions in our dataset. Subsequently, we apply benchmark datasets on CLANet for inspecting its rationale against our dataset. Finally, we compare our CLANet with other state-of-the-art multi stream CNN LSTM models for cross data action recognition.

A. Datasets and Performance Measures

The NTU RGB D [32] is the largest dataset with 60 action classes in 80 views recorded with 40 subjects with a total sample size of 56880 videos of skeleton, depth and RGB. We selected 60 action classes with 40 subjects for training and testing the proposed CLANet. The NTU RGB D dataset used in our work has 2400 video samples with 40 subjects in 60 action classes. MSRDailyActivity3D [33] is another standard benchmark dataset using Microsoft Kinect with 16 activity types. It consists of 320 video samples in both RGB and depth modes with actions performed in both sitting and standing positions. The other most widely used RGB D action dataset for benchmarking is UTKinect [34] which has 10 actions from 10 subjects each performing the action twice. It has 10 classes with $10 \times 10 \times 2 = 200$ videos of both RGB and depth data. Inspired from the above benchmark datasets, we collected our own BVRCAction3D action dataset with 40 single human and 10 two human actions using 5 subjects. The complete list of actions is available at [7]. Fig. 4 presents some action sequences in RGB and depth from our BVRCAction3D dataset.

Fig. 4. BVRCAction3D Dataset. Sample RGB and Depth Video Frames of (a)-(b): Clapping, (c)-(d): Mopping the Floor and (e)-(f): Eating.

Our BVRCAction3D dataset consists of $50 \times 5 \times 2 = 500$ video sequences with 50 classes from 5 subjects each performing the action twice. We used Kinect 1.0 for capturing the actions. Each action was recorded for 60 seconds at 30fps. Consequently, each action video has 1800 frames with a resolution of 640×480 for RGB and 320×240 for depth. In order to maintain uniformity across datasets, we resized the frame sizes to 256×256 in both RGB and depth modal videos. Moreover, we found the number of frames in each video clip to have a high degree of similarity among themselves. To increase the redundancy in the action videos, we selected 120 Key frames per action by applying correlation based key frame extractor [35].

The performance of the proposed deep network is measured using two standard parameters: mean Recognition Accuracy (nRA) and mean f1 score (mf1). Apart from the two, we also obtained confusion matrices and region of convergence (ROC) plots across all datasets. In the following subsection, we apply various datasets to our proposed CLANet and evaluate its performance.

B. CLANet Performance

The proposed multi modal CLANet is trained with RGB D action sequences from our BVRCAction3D and other benchmark datasets. The training parameters were kept constant across dataset to understand the implications of data on the network. The hyperparameters of the network were selected as discussed in the previous section. Fig. 5 shows the confusion matrices on the datasets used in this work. The performance of CLANet on our dataset is high when compared to other datasets due to less noisy backgrounds in BVRCAction3D as shown in the Fig. 4. The scores from CLANet are found to be better than our previous work in [7], where we used multi stream CNN with motion information. The reason for higher accuracies is because of the LSTM network which models the time series information in a more accurately. The testing in this case in preformed with 10 test samples only.

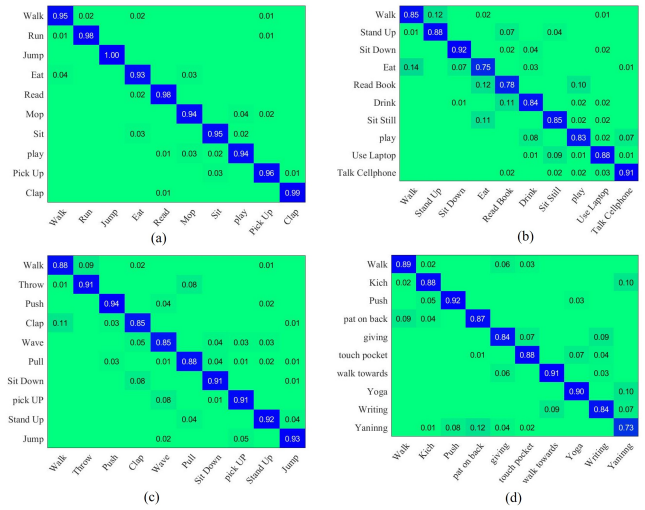

Fig. 5. Confusion Matrices for CLANet on (a) BVRCAction3D, (b) MSRDailyActivity3D, (c) UTKinect and (d) NTU RGB D Datasets for 10 Test Samples.

Eventually, we tested the trained CLANet with the entire testing dataset from each dataset and projected the results

in Table I. The results in Table I indicate two performance parameters mRA and mF1 for the proposed CLANet across multiple datasets used in this work. The testing is conducted in cross subject mode, means that the network is shown samples with subjects that are previously unseen by the network during training. The average recognition rate achieved is around 93.32% on our BVRCAction3D dataset, which is found to be better than our previous work in [7].

The above comparison with our work in [7] is important in the context of understanding the need for time series modelling against motion modelling using optical flow. Contrastingly, optical flow-based motion estimation and processing it with regular spatial network has limitations in characterizing the changes across multiple frames. Additionally, the flow-based models fail to capture the long-term dependencies in the action video sequences. Interestingly, hybrid CNN LSTM networks have performed exceptionally well by modelling spatio temporal contents in the action video sequences. Meanwhile, the depth data has come in to assist this process by increasing the performance of the network.

However, it is not possible to generate depth data in real time and hence, we conducted a RGB only test on our proposed CLANet to understand its usability as a real time application. We supplied zero matrices in place of depth data during testing for the depth stream. This test resulted in a mean accuracy of 84.76% and a mean f1 score of 0.862 for BVRCAction3D dataset. The second right half of table I shows the results on all datasets. In spite of depth data absence during testing, the proposed CLANet has performed better on our BVRCAction3D dataset when compared to other benchmark datasets. Consequently, the performance of the network has to be gauged by comparing its performance against state-of-the-art networks as presented in the next subsection.

C. Comparison with Recurrent Hybrid Networks

This subsection gives the comparison of hybrid CNN LSTM networks with RGB and depth inputs as training data. Surprisingly, there are very few works which used both RGB and depth data with hybrids networks for action recognition applications. However, there are a large contingent of networks for skeleton based action recognition using CNN LSTM architectures. Table II presents the comparison of our proposed CLANet with the previously proposed methods for action recognition using the benchmark datasets. We implemented all these networks on the datasets and the mean average recognition is calculated across the training data. The results show that the proposed network outperforms the existing models. All the hyper parameters of the networks were incepted from the proposed CLANet. This is because of the spatio temporal characteristics that are learned effectively by the network in two modalities simultaneously. However, it would be interesting to check the network performance against different action recognition models. Hence, in the next subsection, we compare our method with other state-of-the-art RGB D based action recognition models.

D. Comparison with RGB D Action Recognition Deep Models

The parameters used for training and testing were as described in Section III. In all experiments, the video resolutions were fixed at $256 \times 256 \times 3$ for both training and testing for both

TABLE I. PERFORMANCE OF CLANET ACROSS DATASETS AND COMPARISON WITH WORK FROM [7]

Dataset	RGB and Depth Testing				RGB only testing			
	mRA	mf1	mRA [7]	mf1 [7]	mRA	Mf1	mRA [7]	Mf1 [7]
BVRCAction3D	93.32	0.965	92.05	0.942	84.76	0.862	72.26	0.687
MSRDailyActivity3D	88.42	0.924	84.86	0.876	74.89	0.784	62.24	0.638
UTKinect	90.25	0.937	87.63	0.894	75.33	0.795	63.85	0.643
NTU RGB D	91.42	0.948	90.27	0.912	78.96	0.812	66.11	0.672

TABLE II. COMPARISON OF AVERAGE RECOGNITION (MRA) OF HYBRID RECURRENT CNN MODELS WITH RGB AND DEPTH INPUTS.

Method	Modality	BVRCAction3D	MSRDailyActtivity3D	UTKinect	NTU RGB D
Deep Bilinear CNN [1]	RGB+Depth	82.36	74.8	76.24	79.2
Auxiliary Dataset Model [3]	RGB+Depth+Skeletal maps	92.22	88.12	88.36	89.36
Optical Flow CNN [7]	RGB+Depth	92.05	84.86	87.63	90.27
2 Stream CNN [4]	RGB+Depth	80.23	94.53	95.23	96.32
CLANet Proposed	RGB+Depth	93.32	88.42	90.25	91.42

TABLE III. COMPARISON OF VARIOUS DEEP LEARNING MODELS FOR RGB D ACTION RECOGNITION

Modality	Methods	mRA			
		BVRCAction3D	MSRDailyActtivity3D	UTKinect	NTU RGB D
RGB	Two Stream CNN [36]	69.22	68.32	69.03	69.96
	CNN LSTM	73.34	70.82	71.98	73.89
	Spatio Temporal CNN [19]	71.89	69.33	71.12	72.02
	3D CNN LSTM [29]	78.96	74.02	75.64	78.91
Depth	CNN [21]	71.84	68.96	69.98	70.52
	CNN [22]	72.36	70.50	71.22	72.36
	RNN [23]	74.71	72.44	73.58	74.55
	CNN RNN [24]	78.83	74.99	76.12	77.18
Skeletal	Hierarchical RNN [8]	84.36	79.03	81.52	82.92
	CNN LSTM [10]	87.11	82.37	85.05	86.03
	Temporal Sliding LSTMs [15]	87.94	81.55	85.22	86.07
	Visual Attention [18]	89.36	84.05	86.85	88.96
RGB + Depth	CNN LSTM	93.32	88.42	90.25	91.42

RGB and depth data across all subjects. The aim of this section is to investigate the suitability of RGB and depth information for action classification through deep learning networks. Given that, we compare the mRA from multiple architectures on three multi modal action data. Table III presents the results of our investigation. The networks were borrowed from previous methods and were trained from scratch on the datasets used in this work. All the networks are trained and tested only once. From Table III, we were able to generate two insights regarding the performance of the action recognition models. One is based on the use of input data and the other is on the deep networks. We see that RGB based methods performed poorly when compared to the other two modalities, depth and skeletal. This is because of the background noise that exists in the RGB video frame that are difficult to learn during training of spatial networks. Contrastingly, this background noise is relatively less in-depth frames, and it is completely absent in skeletal data. Hence, skeletal action recognition is the popular choice for producing higher accuracies with deep networks. Despite their success the skeletal action data becomes noisy when there is a joint overlap during the action sequence producing ambiguous results.

Simultaneously, skeletal action data is represented as time series data which is perfectly characterized and discriminated using RNNs and LSTMs together. These networks have produced the highest recognition accuracies across all datasets. However, modelling RGB and Depth as time series data by extracting features and inputting those features to recurrent networks has shown to improve performance. However, the most obvious choice of combination is the skeletal data with either depth or RGB. The fusion with skeletal data has improved the discriminating confidence of the networks. The most suitable network architecture is the hybrid CNN LSTM which can extract spatial and temporal dynamics of the action data. Contrasting to the regular phenomenon, we applied RGB and depth modalities to CNN LSTM architecture to generate a highly discriminating feature vector for action recognition. Table III shows that our proposed method is on par with the existing state of the art models and in fact better than some of the existing models. All the models are tested with cross subject data. Finally, the last subsection evaluates the networks for cross data validation.

E. Cross Data Validation

This section shows the experimental evaluation of CLANet across datasets. We found some of the common actions across datasets and evaluated the performance of CLANet with separate training and testing data from multiple datasets. Incidentally, we trained the CLANet with our BVRCAction3D dataset and tested with same actions from another dataset. We used seven common actions across datasets. The results of this experiment were presented as mean recognition accuracy across these seven actions used for training and testing in Table IV. Here, the network has to fine tuned multiple times and the recognition rates obtained are averaged across multiple runs of the algorithm. Table IV shows that the proposed network has capabilities in evaluating cross data action recognition. Interestingly, we found that training with less noisy data could result in good recognition accuracies when compared to a noisy data training. The average recognition was around 65% across datasets with the proposed CLANet with RGB and depth input data.

Despite better performance by the hybrid CNN LSTM architecture across RGB D action datasets for recognition tasks, there are many challenges such as view invariance, cross data and occlusions that need attention. We found that it is difficult to achieve high degree of robustness for some complex actions from the existing deep learning frameworks. Moreover, deep networks are data intensive models and require a wide variety to provide actionable intelligence across action recognition platforms. Finally, more hybrid models with multiple levels of abstraction are required for designing deployable action recognition models.

V. Conclusion

In this paper, we have proposed a novel approach for recognizing RGB D action data. Specifically, our method involves training of a hybrid CNN LSTM multi stream network on multi modal data, RGB and depth videos. The CNN network is designed to extract spatial features from both RGB and depth action frames. Subsequently, bidirectional LSTM network is used to model the sequential information in the extracted multi modal features at the output of the CNN. The hybrid CLANet is trained and tested using our generated BVRCAction3D dataset and other benchmark datasets for recognition. The results conclude that the proposed network is capable of achieving higher average recognition rates of around 93.32% on our dataset and an average of 90.24% across all benchmark datasets.

References

[1] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for RGB-d action recognition. In *Computer Vision – ECCV 2018*, pages 346–362. Springer International Publishing, 2018.

[2] Sunitha Ravi, Maloji Suman, P.V.V. Kishore, Kiran Kumar E, Teja Kiran Kumar M, and Anil Kumar D. Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–d based sign language gesture recognition. *Journal of Computer Languages*, 52:88–102, jun 2019.

[3] Yen-Yu Lin, Ju-Hsuan Hua, Nick C. Tang, Min-Hung Chen, and Hong-Yuan Mark Liao. Depth and skeleton associated action recognition without online accessible RGB-d cameras. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014.

[4] Qiuxia Wu, Zhiyong Wang, Feiqi Deng, Zheru Chi, and David Dagan Feng. Realistic human action recognition with multimodal feature selection and fusion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(4):875–885, jul 2013.

[5] Z. Gao, S.H. Li, Y.J. Zhu, C. Wang, and H. Zhang. Collaborative sparse representation leaning model for RGBD action recognition. *Journal of Visual Communication and Image Representation*, 48:442–452, oct 2017.

[6] Jing Zhang, Wanqing Li, Philip O. Ogunbona, Pichao Wang, and Chang Tang. RGB-d-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, dec 2016.

[7] D. Srihari, P. V. V. Kishore, E. Kiran Kumar, D. Anil Kumar, M. Teja Kiran Kumar, M. V. D. Prasad, and Ch. Raghava Prasad. A four-stream ConvNet based on spatial and depth flow for human action classification using RGB-d data. *Multimedia Tools and Applications*, 79(17-18):11723–11746, jan 2020.

[8] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[9] E. Kiran Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. Teja Kiran Kumar, and D. Anil Kumar. Training CNNs for 3-d sign language recognition with color texture coded joint angular displacement maps. *IEEE Signal Processing Letters*, 25(5):645–649, may 2018.

[10] Juan C. Núñez, Raúl Cabido, Juan J. Pantrigo, Antonio S. Montemayor, and José F. Vélez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, apr 2018.

[11] P. V. V. Kishore, D. Anil Kumar, A. S. Chandra Sekhara Sastry, and E. Kiran Kumar. Motionlets matching with adaptive kernels for 3-d indian sign language recognition. *IEEE Sensors Journal*, 18(8):3327–3337, apr 2018.

[12] Eepuri Kiran Kumar, P. V. V. Kishore, Maddala Teja Kiran Kumar, Dande Anil Kumar, and A. S. C. S. Sastry. Three-dimensional sign language recognition with angular velocity maps and connived feature ResNet. *IEEE Signal Processing Letters*, 25(12):1860–1864, dec 2018.

[13] Earnest Paul Ijjina and Krishna Mohan Chalavadi. Human action recognition in RGB-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72:504–516, dec 2017.

[14] Anne Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, and Rahul S Patwardhan. Sensor tracked points and hmm based classifier for human action recognition. *Computer Science and Emerging Research Journal*, 5:4–8, 2016.

[15] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[16] Danilo Avola, Marco Cascio, Luigi Cinque, Gian Luca Foresti, Cristiano Massaroni, and Emanuele Rodola. 2-d skeleton-based action recognition via two-branch stacked LSTM-RNNs. *IEEE Transactions on Multimedia*, 22(10):2481–2496, oct 2020.

[17] Jiajia Luo, Wei Wang, and Hairong Qi. Spatio-temporal feature extraction and representation for RGB-d human action recognition. *Pattern Recognition Letters*, 50:139–148, dec 2014.

[18] Zhengyuan Yang, Yuncheng Li, Jianchao Yang, and Jiebo Luo. Action recognition with spatio–temporal visual attention on skeleton image sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2405–2415, aug 2019.

[19] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector CNNs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[20] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Computer Vision – ECCV 2018*, pages 392–408. Springer International Publishing, 2018.

[21] Pichao Wang, Wanqing Li, Zhimin Gao, Chang Tang, and Philip O. Ogunbona. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Transactions on Multimedia*, 20(5):1051–1061, may 2018.

[22] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-d segmentation. In *Computer Vision – ECCV 2018*, pages 144–161. Springer International Publishing, 2018.

[23] Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based RNNs for action recognition from depth sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[24] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with rCNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

[25] Chao Li, Zhongtian Bao, Linhao Li, and Ziping Zhao. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Information Processing & Management*, 57(3):102185, may 2020.

[26] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia - MM -15*. ACM Press, 2015.

[27] Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, and Jiaji Wu. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE Access*, 6:17913–17922, 2018.

[28] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using LSTM and CNN. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, jul 2017.

[29] Xuanhan Wang, Lianli Gao, Jingkuan Song, and Hengtao Shen. Beyond frame-level CNN: Saliency-aware 3-d CNN with LSTM for video action recognition. *IEEE Signal Processing Letters*, 24(4):510–514, apr 2017.

[30] Wen-Nung Lie, Anh Tu Le, and Guan-Han Lin. Human fall-down event detection based on 2d skeletons and deep learning approach. In *2018 International Workshop on Advanced Image Technology (IWAIT)*. IEEE, jan 2018.

[31] Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based RNNs for action recognition from depth sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[32] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[33] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012.

[34] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2012.

TABLE IV. TABLE SHOWS MRA OF MULTI-SOURCE TRAINING AND TESTING OF THE PROPOSED CLANET

Training Dataset	Testing Dataset	Recognition Rates on Different Actions							
		Walking	Jumping	Jogging	Kicking	Hand Waving	Clapping	Running	Average Recognition
MSRDailyAction3D	BVRCAction3D	67.26	69.23	69.51	69.93	67.54	69.23	68.73	68.77
	MSRDailyAction3D	86.98	88.88	87.54	89.79	88.30	88.73	87.55	88.25
	UT Kinect	66.90	65.85	67.26	67.37	67.35	66.43	65.60	66.68
	NTU RGB D	69.52	61.00	69.36	60.67	60.25	61.52	69.00	65.18
BVRCAction3D	BVRCAction3D	98.85	97.55	97.26	96.81	97.51	96.97	97.50	97.49
	MSRDailyAction3D	68.94	68.50	66.35	66.90	66.53	67.89	67.33	67.49
	UT Kinect	65.89	64.73	65.86	65.61	65.13	65.25	65.73	65.45
	NTU RGB D	65.93	65.81	64.89	65.94	63.79	64.78	64.74	65.12
UT Kinect	BVRCAction3D	52.59	53.98	53.70	55.51	53.90	55.66	53.68	54.14
	MSRDailyAction3D	59.90	57.75	55.97	66.83	56.29	54.38	54.94	56.58
	UT Kinect	89.90	87.75	85.97	86.83	86.29	84.38	84.94	86.58
	NTU RGB D	58.97	58.75	59.90	57.95	50.46	50.75	50.96	55.67
NTU RGB D	BVRCAction3D	66.64	64.93	66.56	64.89	66.91	65.92	65.45	65.9
	MSRDailyAction3D	62.75	63.97	63.80	63.93	65.41	64.65	63.94	64.06
	UT Kinect	66.96	66.44	66.82	66.62	65.74	66.43	65.43	66.34
	NTU RGB D	94.80	94.90	94.94	94.80	93.53	93.65	93.58	94.31

[35] Li Liu, Ling Shao, and Peter Rockett. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*, 46(7):1810–1818, jul 2013.

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.