Dense ResNet Based Human Action Recognition using Novel Trajectory Maps on 3D Skeletal Data

K. Rajendra Prasad, P. Srinivasa Rao

Abstract*: *The machine learning research community is presently working on human action/activity recognition issue in real-time videos, and facing several hundreds of confronts. In this scenario, deep convolutional neural networks have initiated their powerful role in strengthen the numerous vision-based HAR systems. In recent years there has been impressive performance and great potential for imaging tasks with introducing residual connections along with a traditional CNN model in a single architecture known as the Residual Network (ResNet). In this paper we propose to use skeletal trajectory maps for the detection of human actions. A new ResNet based algorithm named dense ResNet has been proposed to perform the classification task. The trajectories of 3D joint locations are converted into color coded RGB images. These trajectory plotted images are able to capture the spatio-temporal evolutions of 3D motions from skeleton sequences and can be efficiently learned by deep learning algorithms. We then train the proposed dense ResNet to learn the features from these color coded RGB trajectory information of the human body 3D joint locations. The novelty of the proposed method is evaluated on MSR Action 3D, UTKinect-Action3D, G3D and NTU RGB-D datasets. Experimental results shows that the proposed architecture attains good recognition rates with less computation resource.

Index Terms*: *Human action recognition, Residual networks, Trajectory maps, 3D skeletal data.

I. INTRODUCTION

Human Action Recognition (HAR) is a crucial computer field, and is an significant part of several smart systems including video surveillance, interaction between people and the machines, self-driving cars, robot-vision and so on. The foremost objective is to determine what people do in unknown videos and to recognize them. Although significant advances have been made in recent years, the accurate identification of video action remains a difficult task because of a large number of obstacles, including viewing, occlusion or lighting conditions. Traditional HAR studies mainly concentrate on the utilization of hand-made local features such as Cuboids or HOG / HOF. Typically these methods distinguish human actions founded on the appearance and movement of human body components in a RGB videos. A second method to the generation of space-time descriptors of motions is the use of genetic programming (GP). However, the absence of a 3D scene structure is the greatest confines of 2D based data. Single modality of RGB sequence action recognition is therefore not enough to solve the challenges in HAR, particularly in truthful videos. Newly, the hasty evolution of the technology of depth sensing camera time has helped to tackle problems that traditional cameras regard as complex. Thorough information on the 3D model of human movements can be provided by the depth cameras such as Microsoft Kinect or ASUS Xtion. Many approaches to the recognition of RGB, depth, or combination of these data categories (RGB-D) have therefore been proposed, provided by depth sensors. Developmental sensors have been proposed. Moreover, in order to describe the activities in a more precise and effective way, they can also provide real-time algorithms. The advantages of skeleton-based representations are less dimensional than RGB / RGB-D. This advantage simplifies and accelerates action recognition systems. Thus, it is promising to exploit the 3D skeleton data provided by HAR depth sensors. Many approaches have been proposed to skeleton-based action recognition. Tactics constructed on CNNs have been successful during numerous image classification tasks in recent years. A new research direction for higher-performing CNN architecture research has been opened following the success of the AlexNet competition. This indicates that CNNs can significantly enhance their learning performance by improving their depth. A number of studies in the HAR literature showed that CNNs were more capable than handcrafted approaches to learning more complex motion features. However, maximum researchers have only recently fixated on the practice of comparatively simple and small CNNs like AlexNet and have still not exploited fully the potential of the most modern and highly deep-rooted CNN architectures. Moreover, RGB, depth or RGB-D sequences are used for most current CNN based approaches to provide the insert into knowledge models. Even though RGB-D images are informative, they increase quickly when the dimensions of the inputs are large, however, the computational complexity of these models is significant. This makes approaches to be more complicated, sluggish and less convenient to resolve large-scale problems and applications in real-time. In the present document, we intended to take advantage of 3D skeleton-based depictions and to figure an end-to-end HAR learning outline from skeletal data in order to learn highly

Revised Manuscript Received on July 06, 2019.

K. Rajendra Prasad, Research Scholar, Department of Computer Science and Systems Engineering, College of Engineering, Andhra University, Visakhapatnam, India.

P. Srinivasa Rao, Professor, Department of Computer Science and Systems Engineering, College of Engineering, Andhra University, Visakhapatnam, India.

hierarchical image functionality of Deep Convolutional Neural Networks (D-CNNs). All 3D skeleton coordinates in the body recorded by Kinect sensor shall be characterized by 3D arrays. The trajectories of these 3D locations were stored with a simple encoding method as RGB images. The key aim of this dispensation step is to guarantee that color coded images efficiently represent the human action's spatial-temporal structure in skeleton sequences and are compatible as D-CNNs with the deep learning networks. The Residual Networks (ResNets) are proposed to use [1] to get higher levels of performance, to learn picture features and identify their labels. We suggest that we develop a new deep architecture based on the original ResNets that is easier to optimize and better avoid overfitting. The method proposed in four skeleton benchmark datasets is assessed and we have acquired the state of the art recognition accuracy on all of these datasets namely MSR Action 3D[2], UTKinect-Action3D [3], G3D [4] and NTU RGB-D [5]. Moreover, we also emphasize the effectiveness of our computational complexity learning framework, the capability to prevent overfitting and reduce the impact of deprivation on very deep training networks. First of all, we propose a complete framework on which to acquire the spatial-time changes made in images of RGB that are encoded from 3D skeletal sequences for human action recognition, based on ResNets. Secondly, the new building unit ResNet to build deep ResNets is presented. Our experiments on actions are proof of better characteristics than the original ResNet model for the proposed architecture [1]. This architecture is general and can be used not merely to recognize human action for various image recognition problems. Finally, by attaining state-of-the-art performance on four standard datasets. We demonstrate that our learning outline is effective in terms of action recognition tasks. The rest of the paper is organized as follows. Section 2 discusses literature review. The details of our proposed method are given in Section 3. Section 4 describes data sets, training and testing data organization and implementation details. Section 5 shows experimental results and discuss the accuracy of classification, overfits, degradation phenomena and computational efficiency of the deep learning networks proposed. Section 6 draws the conclusions of this work.

II. LITERATURE REVIEW

Our study covers two main themes: the identification of skeleton-based action and the design of dense ResNet architectures for visual classification work. Some major lessons on these issues are presented here. We talk first about previous work on the recognition of skeletons. Then we present an outline of the progress and potential of Deep CNNs for HAR. With regard to HAR, we are referring interested users in RGB / RGB-D, among the most successful approaches to RGB representation from RGB data, together with bag of words (BoWs), dynamic image networks and D-CNNs. Depth sensors have extensively used 3D skeleton data for HAR. Current action recognition approaches based on skeletons can be categorized as two principle clusters. The first cluster combines hand-made skeleton characteristics and graphical representations to identify actions. Spatio-temporal representing skeleton sequences are frequently modelled on several common probabilistic graphic models, such as the Hidden Markov Model (HMM), Latent Dirichlet Allocation (LDA). In addition, it was employed to capture and then predict temporal dynamics of action using Fourier Temporary pyramid (FTP). In particular, the authors defined an action as a sequence of skeletal forms and analysed them with a statistical form analysis tool like Kendall's geometry. The classification was subsequently made use of typical graduation devices, e.g. K-Nearest-Neighbor (KNN) or Support Vector Machine (SVM). However, despite promising results, most of these works require a large amount of feature design. For example, the sequences of skeletons must often be segmented and aligned for approaches based on HMM and CRF. In the meantime, the time sequences of actions cannot be recorded globally by FTP based approaches. Recurrent neural networks with a long short term memory network (RNN-LSTMs) [6] constitute the second group of methods. The RNN-LSTM network architecture enables the long-range context information in a time sequence to be stored and accessed. As a time series problem can be seen in the recognition of a human skeleton action [7], skeletal data can be used to study human movements. Many researchers have therefore explored 3D HAR RNN-LSTMs from skeleton sequences. Some authors used the CNN as a visually impressive extractor in a unified framework for human movement models to better capture the space-time dynamics in skeleton. Although good performance was reported with RNN-LSTM-based approaches. However, it is difficult to overcome certain limitations, i.e. RNN use can lead to problems overfitting if the number of input functions for the training network is insufficient. Meanwhile, when the input characteristics grow, computer time can become a serious problem. Deep CNNs have led to a series of breakthroughs in image recognition and associated tasks for visual identification of CNNs. There has been increasing evidence recently that D-CNN models can enhance image recognition performance. However, it is very difficult to train deep networks. The problem of decline in gradients or degradation phenomena are two main reasons impeding the convergence of deeper networks. The problem with fade gradients occurs when the network is sufficiently deep, with the error signal on its way back to the input layer fully diminished. The standardization of this barrier [8], particularly through batch standardization, has been solved. A degrading phenomenon occurs when the profound networks begin to converge. Adding more layers to a deep network can result in greater training and/or testing failures. This is not as simple as a problem of overfitting. He et al. [1] introduced Residential Networks with shortcut links parallel to their traditional convolutional layers in order to reduce the effects of the disappearance of gradient problems and the degradation phenomenon. This idea helps ResNets increase layer-by-layer information flow. Two well-known datasets including CIFAR-10 and ImageNet have been confirmed by expert results that ResNets can improve recognition performance and reduce the phenomenon of degradation. Many authors have used CNN's skeletal data feature education skills. These studies focus on the search for good skeletal images and learning features with simple CNN architectures. In contrast, this paper focuses on using the power of dense ResNets to detect action using a simple skeleton. In order to

gain action in skeleton sequences and classify them in classes, we are investigating and developing a new deep learning framework based on ResNet. Furthermore, our solution is general and can be applied to various input data types. For example, it can be used with inertial sensor movement capture (MoCap) data.

III. METHODOLOGY

The encoding of skeletal data into trajectory maps and the proposed modified deep residual network are presented in this section. The spatio-temporal skeletal information is encoded as trajectory images. These trajectory maps are inputted to the proposed ResNet architecture to familiarize and recognize the human actions from the skeletal information.

A. Trajectory Map Preparation

An action sequence $A=[F_1,F_2,F_3,\ldots\ldots\ldots F_{N-1},F_N]$ represents one complete action. Where the action sequence formed with N skeletal frames with m joints in each frame represented with their (x,y,z) coordinates in 3D space. The more detailed representation of an action sequence is given in equation 1.

$$A=\underset{frames}{\downarrow}\begin{bmatrix} J_{11} & J_{12} & J_{13} & \cdots & J_{1(m-1)} & J_{1m} \\ J_{21} & J_{22} & J_{23} & \cdots & J_{2(m-1)} & J_{2m} \\ J_{31} & J_{32} & J_{33} & \cdots & J_{3(m-1)} & J_{3m} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ J_{(N-1)1} & J_{(N-1)2} & J_{(N-1)3} & \cdots & J_{(N-1)(m-1)} & J_{(N-1)m} \\ J_{N1} & J_{N2} & J_{N3} & \cdots & J_{N(m-1)} & J_{Nm} \end{bmatrix} \quad (1)$$

Where J represents the joint. Each joint in the A is represented with its corresponding x,y,z joint coordinate values. The joint $J_{Nm}=(x_m,y_{m,}z_m) \in \mathbb{R}^{3D}$ of N^{th} frame.

The difficulty in implementing CNN based techniques for recognizing the human action from skeletal data is how the temporal information can be represented effectively, while inputting to CNN to familiarize the key features. In general, the CNNs showed their superiority in classifying the actions on still images. Hence, transforming the spatio-temporal 3D joint information into 2D image such as trajectory map will work effectively in classifying the human actions. In this work, we considered the trajectory maps in four views via $x-y$ view, $x-z$ view, $y-z$ view and $x-y-z$ view. The trajectory maps created for sample actions from NTU RGB-D data is shown in figure 1.

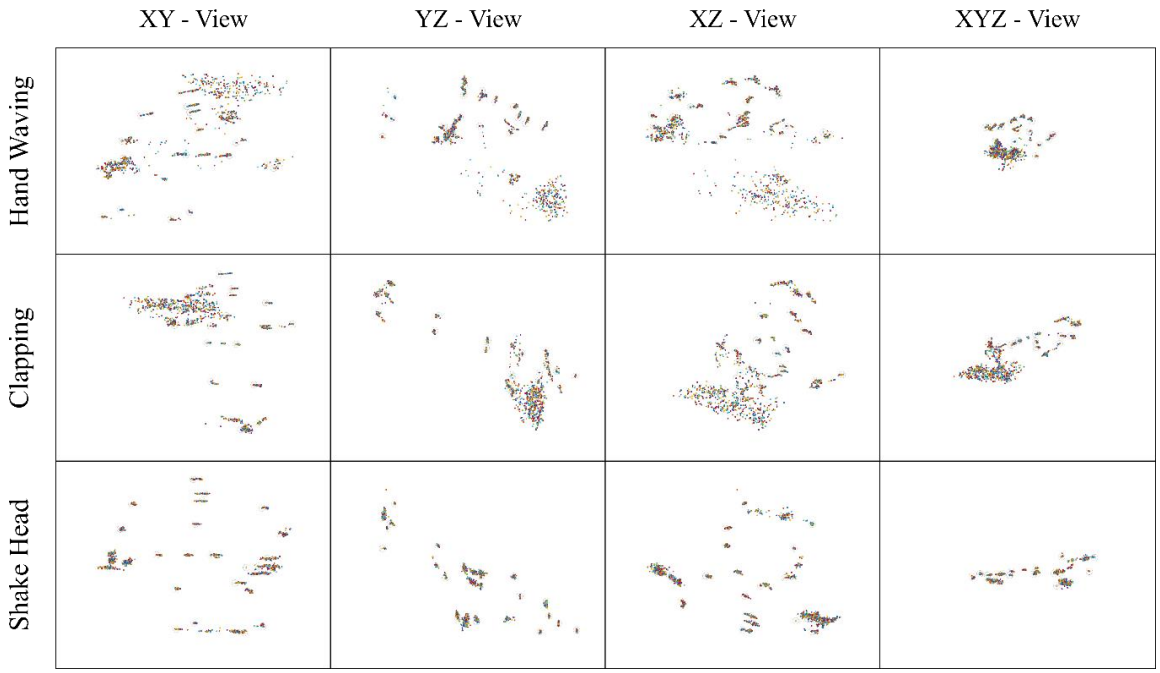

Figure 1. Visualization of sample trajectory maps proposed in this work on NTU RGB-D dataset actions.

In the human action recognition scenario, the human body is divided into five major parts, consisting of two arms, two legs and one thorax or trunk. Simple actions involve the movements of two arms while complex actions involve arm and leg movements along with movements in thorax. Recognition of complex actions is the challenging task for a machine.

B. Proposed Deep ResNet Architecture

The ResNet is constructed from multiple residual building blocks and each block has a shortcut connection, which presents the input at the output of the residual block to provide feedback. The ResNet provides the path for gradients to backpropagate to early layers in the feature learning process (training). The traditional CNNs will not provide any feedback. The nonlinear transformation of the input

of a CNN layer $f(\bullet)$ is achieved by Batch Normalization, Rectified Linear Unit (ReLU) and a convolutional layer operation. The output of a traditional CNN at i^{th} layer $l_{i+1} = f(l_i)$ is defined as follows:

$$l_{i+1} = f(l_i) = \text{ReLU}\left(f(l_i, w_i) + I(l_i)\right) \quad (2)$$

Where l_i and l_{i+1} are the input and output features of the i^{th} residual block. w_i represents the weights associated with the i^{th} Residual block. $I(l_i)$ is the identity function. The nonlinear transformation of input features was obtained by the series of layers covolution – Batch Normalization – ReLU – Convolution – Batch Normalization. The ReLU activation function is performed after each addition. One such ResNet unit proposed in this paper is shown in figure 2.

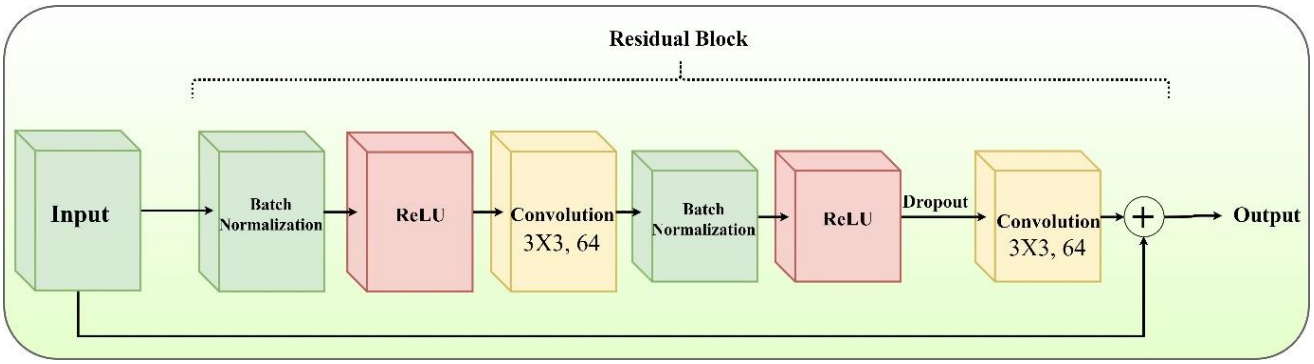

Figure 2. ResNet Unit used in the proposed architecture.

In original ResNet architecture [1] the information within the residual block is propagated through a direct short connection. However, as the ReLU activation behind each element wise addition, the feature propagation cannot occur directly from one residual unit to other. This can be resolved by the architecture proposed in this paper, where features from one element to other element are propagated directly in both forward and backward directions throughout the network. This is achieved by the ReLU activation function adopted after each element wise addition using identity mappings. The feature propagation among the ResNet blocks can be expressed as:

$$l_{i+1} = I(l_i) = f(l_i, w_i) + l_i \quad (3)$$

The above equation evince that the feature l_k of any intermediate layer k can be derived from the features of i^{th} layer l_i as follows:

$$l_k = l_i + \sum_{j=l}^{k-1} f(l_j, w_j) \quad (4)$$

The output features of the k^{th} block can also be drawn from the input features l_{input} using the equation 5.

$$l_k = l_{input} + \sum_{j=1}^{k-1} f(l_j, w_j) \quad (5)$$

From the equation 5, it is seen that the proposed architecture provides the direct short connection which allows the features to propagate through each residual unit of the network. During the supervised training of the network the loss function Ψ required to be optimized. Using the chain rule [9] and equation 4, the backpropagation information in the layers can be expressed as:

$$\frac{\partial \Psi}{\partial l_i} = \frac{\partial \Psi}{\partial l_k}\frac{\partial l_k}{\partial l_i} = \frac{\partial \Psi}{\partial l_k}\frac{\partial\left(l_i + \sum_{j=l}^{k-1} f(l_j, w_j)\right)}{\partial l_i} \quad (6)$$

The above equation can be simplified as:

$$\frac{\partial \Psi}{\partial l_i} = \frac{\partial \Psi}{\partial l_k}\frac{\partial l_k}{\partial l_i} = \frac{\partial \Psi}{\partial l_k}\left(1 + \frac{\partial}{\partial l_i}\sum_{j=l}^{k-1} f(l_j, w_j)\right) \quad (7)$$

The backpropagation information (gradient) $\frac{\partial \Psi}{\partial l_i}$ is controlled by two elements. The term $\frac{\partial \Psi}{\partial l_k}$ is independent of other weight layers, which allows the feature information to be propagated throughout the network among any layers. From the above discussion it can be observed that the replacement of ReLu after each addition by identity mappings provides the direct short connection to the gradient from the loss function and to the input features. Hence the propose architectures eliminates the ReLU activation after each element wise addition and introduces Batch Normalization before to the convolution layer. The ReLU activation is placed after each Batch Normalization layer. To prevent the overfitting problem during the training, 40% dropout was introduced in the network. The detailed architecture of the proposed dense ResNet is shown in figure 3.

IV. DATASETS AND ITS MANIFESTATION

The proposed methodology is tested on well-known datasets, such as MSR Action 3D [2], UTKinect-Action3D [3], G3D [4] and NTU-RGB+D [5]. This section gives brief discussion of these datasets. The MSR Action 3D is a skeletal based action dataset captured by Kinect sensors. This dataset consists of 20 human action classes performed by 10 different subjects in three instances. The skeletal information of this dataset was built with 20 human joints and each joint is represented with its 3D coordinate location values. Figure 4 shows the skeletal representation of MSR Action 3D dataset with 20 joints.

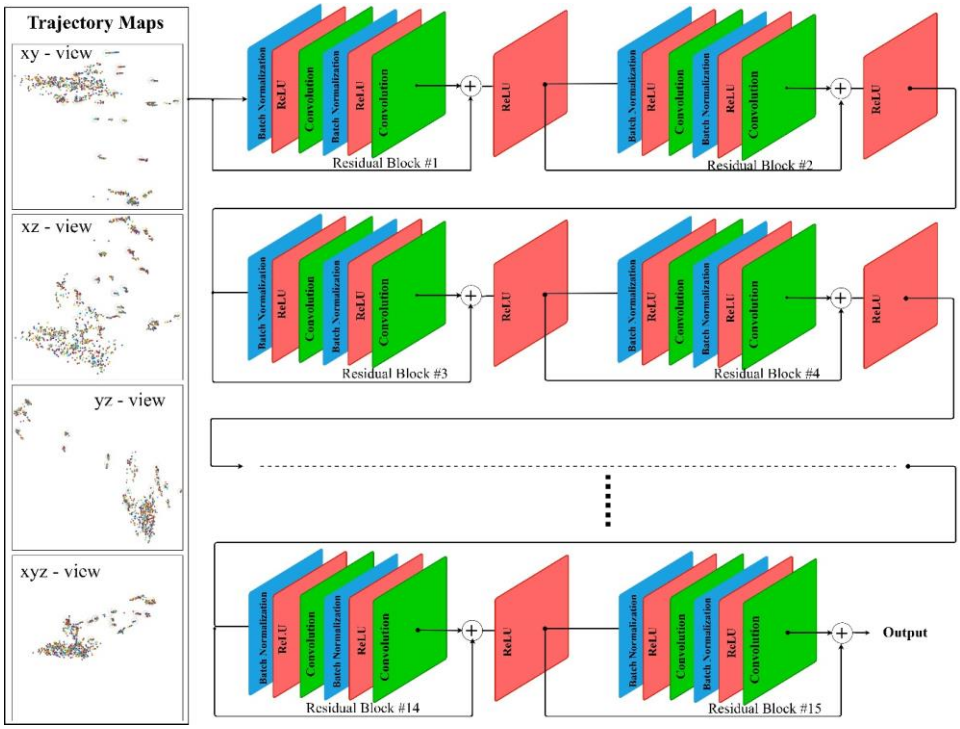

Figure 3: Proposed ResNet architecture with 15 Residual blocks.

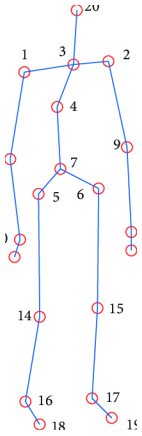

Figure 4: Representation of skeleton used for creating MSR action 3D dataset.

The data introducers divided the captured actions in to three action sub sets. Table 1 shows the actions that are considered for this experiment.

Table 1: Showing the MSR Action labels considered for the experiment.

MSR Action 3D Action Set			
High Arm Wave	High Throw	Two Hand Wave	Jogging
Horizontal Arm Wave	Draw X	Side-Boxing	Tennis Swing
Hammer	Draw Tick	Bend	Tennis Serve
Hand Catch	Draw Circle	Forward Kick	Golf Swing
Forward Punch	Hand Clap	Side Kick	Pickup & Throw

UTKinect-Action3D Action Set is built with 10 actions performed by 10 subjects, of which 9 subjects are male subjects and one is female subjects which include one left handed subject. Each subject performs actions in various views and the length of videos vary from 5 to 120 frames, resulting in significant variation among the recordings. Table 2 shows the action labels considered for the experiment.

Table 2: Showing the List of action labels from UTKinect-Action3D Action Set considered for the experiment.

UTKinect-Action3D Action Set			
Walk	Pick Up	Push	Wave Hands
Sit Down	Carry	Pull	Clap Hands
Stand Up	Throw		

G3D is a gaming 3D action set constructed to recognize the real time gaming actions. It consists of 20 gaming actions performed by 10 subjects. In our experiment we initiated the training with seven subjects and carried out the validation on one subject. The remaining two subjects were used for testing the proposed method. Table 3 shows the list of real time gaming actions considered for the experimentation.

Table 3: Showing the List of action labels from gaming 3D (G3D) actions considered for the experiment.

G3D Action Set			
Punch Right	Golf Swing	Aim And Fire Gun	Crouch
Punch Left	Tennis Swing Forehand	Walk	Steer A Car
Kick Right	Tennis Swing Backhand	Run	Wave
Kick Left	Tennis Serve	Jump	Flap
Defend	Throw Bowling Ball	Climb	Clap

NTU RGB-D is created in a large scale including a

greater number of actions. This is the largest currently available human action dataset. This dataset consists 60 action classes of 40 different subjects providing RGB, Depth and skeletal data. The skeleton of this dataset used 25 major body joints for representation. Out of available 60 action classes, we have considered 20 popular action classes for our experiment and they were listed below in Table 4.

Table 4: Showing the List of action labels from NTU RGB-D dataset, which are considered for the experiment.

NTU RGB-D Action Set			
Drink Water	Throw	Wear Jacket	Jump Up
Brushing Teeth	Clapping	Put on a Hat	Salute
Brushing Hair	Reading	Hand Waving	Staggering
Drop	Writing	Kicking	Slapping
Pickup	Tear Up Paper	Hopping	Handshaking

A. Training and Testing data organization

The training is initiated on the various action datasets discussed above by splitting the data for both training and validation. In MSR action 3D, UTKinect-Action3D, G3D datasets, 6 subject's data is used for training and the remaining 4 subjects data for testing. Whereas for the NTU RGB-D dataset, 30 subjects were used to train the network and the testing was carried out on 10 subjects. ResNets requires a large data to train, the trajectory maps were created on four different viewing angles.

B. Implementation details

In the proposed dense ResNet, is residual block contains two convolutional layers and two non-linear activation functions ReLUs. Local Response Normalization (LRN), pooling and ReLU are considered and the dropout is set to 40%. The weight of the network is derived from mini-batch stochastic gradient descent with a dynamic value of 0.9 and a weight reduction of 0.00005. The initial learning rate was set to 0.001 and the maximum training cycle is considered to be 200. A mini-batch of 50 samples will be built during each cycle by randomly sampling 50 samples from each dataset. The implementation is performed on a NVIDIA GeForce GTX 1080 card and 6 G RAM in Python 3.6 platform with TensorFlow front end and Theano as back end.

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section discuss the experimental results in detail. The proposed dense ResNet is inputted with the trajectory maps of various skeleton based human action datasets. The proposed method is evaluated on MSR Action 3D, UTKinect-Action3D, G3D and NTU-RGB+D. There are serious differences in view and noisy joints in the first three skeletal datasets. The NTU RGB+D is currently the biggest skeleton-based action identification dataset, containing several difficulties to identify the action accurately. In this work, by introducing the novel dense ResNet will accurately recognize the actions from these datasets even though there may be some challenges like noise and viewing orientations.

We designed dense ResNet with 16, 32, 64 and 128 convolutional layers. The experimental results are presented for these four proposed architecture and a comparison of recognition rates were presented.

A. Comparing with state-of-the-art classifiers

Different state of the art classifier algorithms were considered to know the novelty of the proposed dense ResNet. Traditional classifiers like ANN, deep ANNs were considered for classifying the proposed trajectory maps of the various action datasets. A reasonable recognition rates were observed which falls in between 81 to 88 %. As the data to be classify is very huge, the traditional classifiers exhibit their difficulty in classification. Hence the popular machine learning based architectures were considered for classification. LeNets showed an average recognition rate of 87%. The CNNs with 18 and 34 layers exhibited 86.79% and 87.7% respectively. Somehow the dense CNN provided 90.93% of recognition rate. However the proposed dense ResNet stood first with 93% compared to all other classifiers. The table 5 shows the recognition rates obtained on four datasets considered for the experimentation with various classifiers, trajectory maps as input.

Table 5. Comparison of proposed architecture with the state-of-the-art classifiers on different skeletal datasets.

Dataset	% of recognition						
	ANN	Deep ANN	LeNet	CNN-18 [11]	CNN-34 [11]	Dense CNN [12]	Proposed Dense ResNet
MSR Action 3D	82.04	83.97	77.95	84.32	84.56	88.27	89.53
UTKinect-Action3D	81.96	84.69	79.61	86.91	88.94	91.19	92.42
G3D	82.95	83.74	78.62	85.75	86.75	90.83	92.06
NTU RGB-D	85.67	87.25	80.28	90.21	90.57	93.45	94.68

B. Comparing with state-of-the-art ResNet architectures

This section of the paper compares the results achieved in different ResNet architectures. Table 6 shows the recognition rates achieved on different datasets with different architectures. From the table 6 it can be observed that the results obtained using the proposed dense ResNet 64 provides better recognition rates than the original

ResNet 101. The number 64 and 101 denotes the number convolutional layers present in the architecture. In general as the number of layers increase the depth of the architecture increase and results in larger complexity in computations and longer times for training. Our proposed dense ResNets with 64 layers and 128 layers showed almost similar average recognition rates 92.17%, 92.58% respectively and tends us to choose 64 layer architecture over 128 layer architecture, which greatly reduces the computational complexity.

Table 6. Comparison of proposed dense ResNet architectures with original ResNet architectures on different skeletal datasets.

Dataset	% of recognition						
	Original ResNet-20 [1]	Original ResNet-56 [1]	Original ResNet-101 [1]	Dense ResNet-16	Dense ResNet-32	Dense ResNet-64	Dense ResNet-128
MSR Action 3D	82.04	83.97	77.95	84.32	84.56	89.53	89.87
UTKinect-Action3D	81.96	84.69	79.61	86.91	88.94	92.42	93.02
G3D	82.95	83.74	78.62	85.75	86.75	92.06	92.45
NTU RGB-D	85.67	87.25	80.28	90.21	90.57	94.68	95.01

C. Comparing with state-of-the-art features

This section compares the original ResNet-101 and the proposed dense ResNet-64 in terms of recognition rates by inputting different features. We experimented these two networks by giving only RGB, Only Depth, RGB+Depth and proposed skeletal trajectory maps as input. Noticeable amount of recognition rates were observed by inputting the trajectory maps as features for both the architectures. Table 7 compares the two architectures with different input features of different 3D action datasets.

Table 7. Comparison of proposed dense ResNet architectures with original ResNet architectures on various features and proposed features.

Input Features	3D Human Action Datasets							
	MSR Action 3D		UTKinect-Action3D		G3D		NTU RGB-D	
	Original ResNet-101	Dense ResNet-64	Original ResNet-101	Dense ResNet-64	Original ResNet-101	Dense ResNet-64	Original ResNet-101	Dense ResNet-64
RGB	71.46	74.08	71.66	74.96	72.61	75.98	76.74	79.16
Depth	68.97	70.67	70.68	73.12	70.19	75.15	74.63	77.54
RGB+Depth	81.99	83.42	82.04	84.91	80.43	83.67	83.17	86.19
Skeletal Trajectory Maps	87.35	89.53	90.18	92.42	89.48	92.06	91.83	94.68

D. Cross data validation

To further test the robustness of the proposed architecture, the cross data validation has been carried out on four datasets. In the cross data validation, training was done with one dataset and testing was carried out by considering other dataset. To perform cross data validation, the common actions from the four datasets were selected. A good amount of classification rates were recorded with the proposed architecture. Table 8 summarizes the recognition rates obtained in cross data validation.

Table 8. Recognition rates achieved in cross data validation among different 3D human action skeletal datasets.

Training Data	Testing Data			
	MSR Action 3D	UTKinect-Action3D	G3D	NTU RGB-D
MSR Action 3D	89.53	83.49	85.61	82.98
UTKinect-Action3D	82.16	92.42	86.78	80.14
G3D	80.42	84.81	92.06	83.33
NTU RGB-D	92.22	93.41	92.06	94.68

Further, the recognition accuracy was improved by increasing the training to the ResNet. The trajectory maps created in four directions namely $x-y$, $y-z$, $z-x$ and xyz. The training dataset is increased by considering actions from multi subjects in multi views. Figure 5(a), 5(b) visualizes the training and validation accuracies for MSR Action 3D dataset and NTU RGB-D dataset respectively. From these accuracy plots it can be observed that the validation accuracy is almost reaching towards training accuracy, which in turns provides better recognition rates with less number of overfitting.

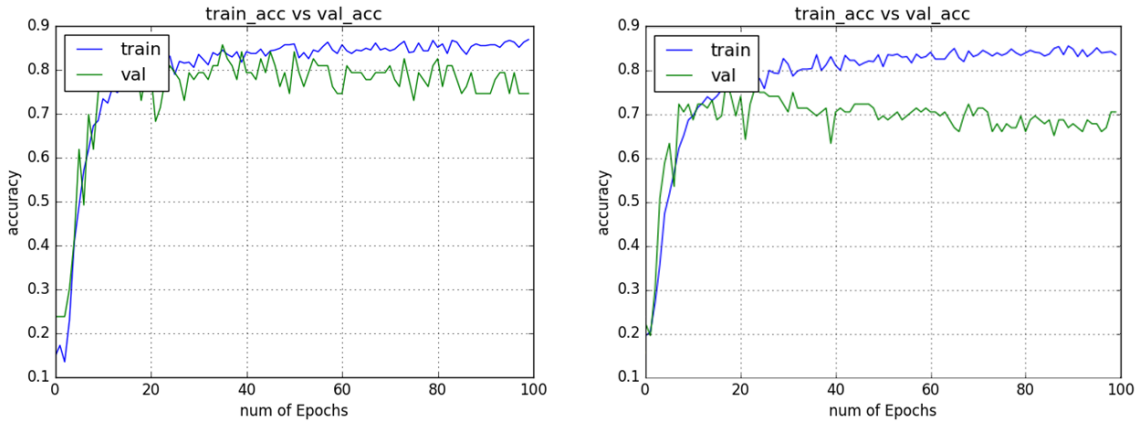

Figure 5. Training and validation accuracy plots for, a) MSR Action 3D data, b) NTU RGB-D data.

Similarly, figure 6 visualizes the loss plots during the proposed network training and validation. Figure 6(a) shows the training and validation loss plot for MSR Action 3D data for 100 epochs and figure 6(b) for NTU RGB-D dataset.

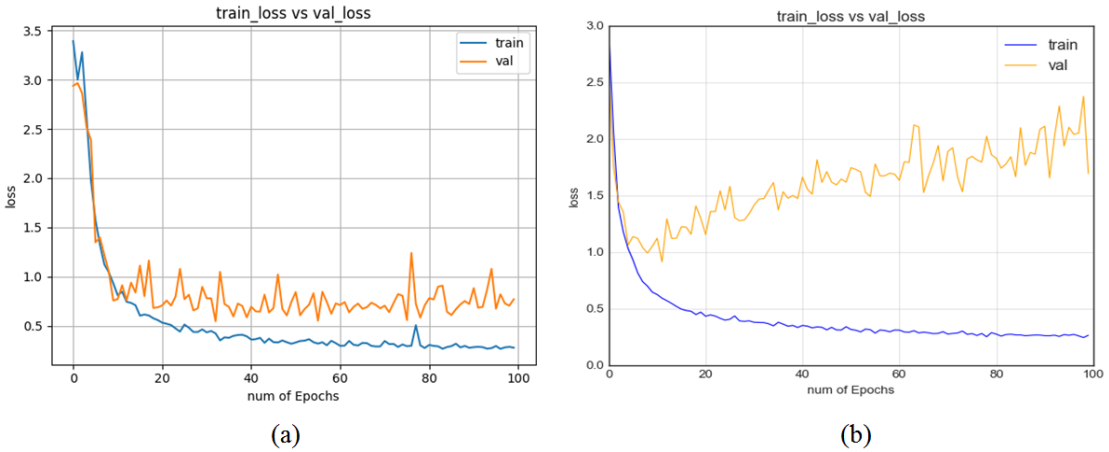

Figure 6. Training and validation loss plots for, a) MSR Action 3D data, b) NTU RGB-D data.

The recognition tasks performed on different datasets. The recognition rates achieved on individual actions of various datasets were presented as confusion matrices. Figure 7 shows the confusion matrix obtained in recognizing individual actions of MSR Action 3D dataset using the proposed dense ResNet-64. For visualization clarity the matrices were shown for limited number of actions from each dataset. Figure 8 and 9 shows the confusion matrices for UTKinect-Action3D and G3D dataset respectively. Finally, the figure 10 visualizes the recognition rates of some sample individual actions from the world's largest human action dataset NTU RGB-D.

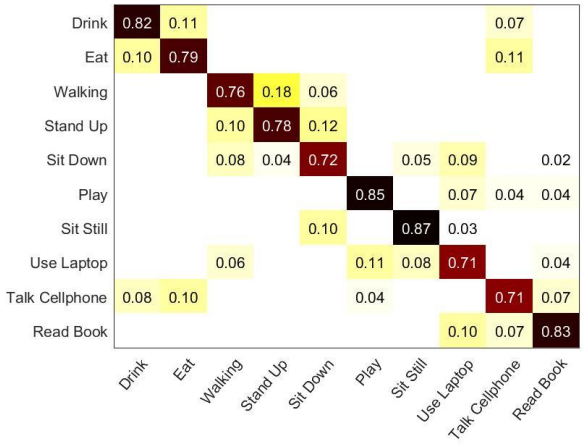

Figure 7. Confusion matrix showing the recognition rates achieved using Dense ResNet on MSR Action 3D dataset.

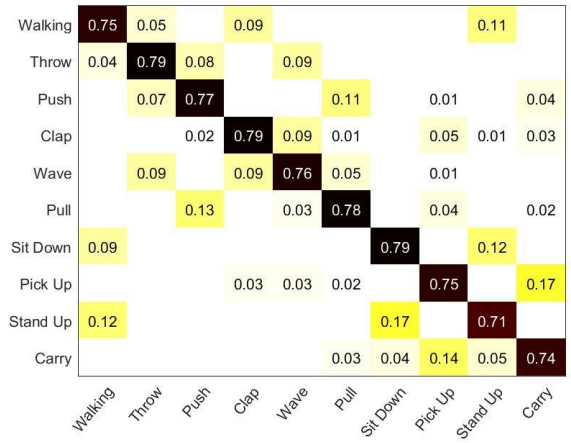

Figure 8. Confusion matrix showing the recognition rates achieved using Dense ResNet on UTKinect-Action3D dataset.

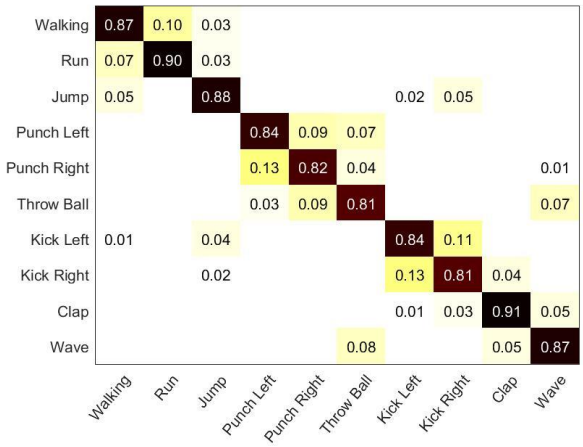

Figure 9. Confusion matrix showing the recognition rates achieved using Dense ResNet on G3D dataset.

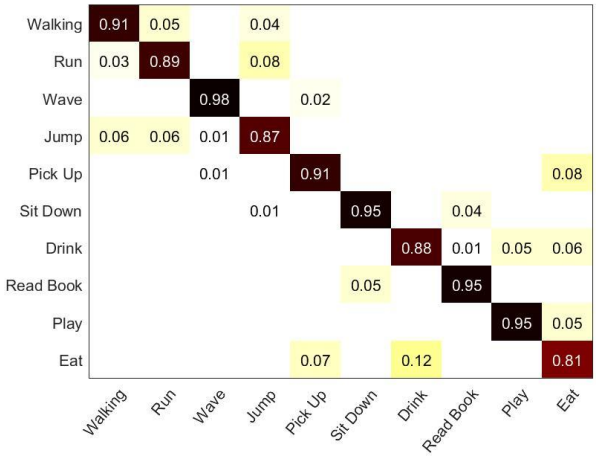

Figure 10. Confusion matrix showing the recognition rates achieved using Dense ResNet on NTU RGB-D dataset.

Due to the large amount of data with different variations in subjects and views, the NTU RGB-D datasets achieved good recognition rates using dense ResNet-64.

VI. CONCLUSION

This paper implemented human action recognition on 3D human action skeletal data by preparing trajectory maps. The 3D skeletal joint trajectories were color coded and represents as an image in four different views. A novel dense ResNet architecture was proposed to effectively classify the actions. The proposed features and architecture were implemented on four publicly available 3D human action datasets. A better recognition rates were noted using the proposed method. The novelty of the proposed features and architecture was highlighted by comparing with other networks with various input features. The promising cross data validation results make the system to be implemented for real time action recognition task. The system developed in this paper will reduce the amount of drop out needed for efficient implementation. The method achieved nearly an average of 93% recognition rate in classifying the human actions.

REFERENCES

1. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
2. Li, Wanqing, Zhengyou Zhang, and Zicheng Liu. "Action recognition based on a bag of 3d points." In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 9-14. IEEE, 2010.
3. Xia, Lu, Chia-Chih Chen, and Jake 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, pp. 20-27. IEEE, 2012.
4. Bloom, Victoria, Dimitrios Makris, and Vasileios Argyriou. "G3D: A gaming action dataset and real time action recognition evaluation framework." In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 7-12. IEEE, 2012.
5. Shahroudy, Amir, Jun Liu, Tian-Tsong Ng, and Gang Wang. "NTU RGB+ D: A large scale dataset for 3D human activity analysis." In

Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1010-1019. 2016.

6. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780.
7. Gong, Dian, Gerard Medioni, and Xuemei Zhao. "Structured time series analysis for human action segmentation and recognition." IEEE transactions on pattern analysis and machine intelligence 36, no. 7 (2013): 1414-1427.
8. Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).
9. LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. "Backpropagation applied to handwritten zip code recognition." Neural computation 1, no. 4 (1989): 541-551.
10. Huang, Zhiwu, Chengde Wan, Thomas Probst, and Luc Van Gool. "Deep learning on lie groups for skeleton-based action recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6099-6108. 2017.
11. Hou, Yonghong, Zhaoyang Li, Pichao Wang, and Wanqing Li. "Skeleton optical spectra-based action recognition using convolutional neural networks." IEEE Transactions on Circuits and Systems for Video Technology 28, no. 3 (2016): 807-811.
12. Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.