Summer Internship Report

On

HUMAN ACTIVITY RECOGNITION USING DEEP NETWORKS

Delhi Technological University

By

Paras Agarwal

16115055

Computer Science and Engineering

National Institute of Technology, Raipur

Under the guidance of

Dr. Dinesh Kumar Vishwakarma

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

June 2019

CONTENTS

DECLARATION

I, PARAS AGARWAL, student of B.tech 6th Semester, studying at NIT Raipur, Raipur hereby declare that the summer internship report on “Human Activity Recognition Using Deep Networks” submitted to Delhi Technological University, New Delhi in the Department of Information technology has not previously submitted to any other University for award of any Degree,Diploma or other similar title or recognition.

Place: New Delhi

Date:

Paras Agarwal
Department of Computer Science and Engineering
National Institute of Technology, Raipur

INTERNSHIP CERTIFICATE

This is to certify that Mr. **PARAS AGARWAL** *Student of* **NATIONAL INSTITUTE of TECHNOLOGY, RAIPUR** *Roll No.* **16115055** *Branch* **COMPUTER SCIENCE AND ENGINEERING** *has completed successfully Summer Research Internship for the period from* **31st May 2019** *to* **30th June 2019** *Duration* **4(Four)** *weeks.*

Topic of Internship was **"HUMAN ACTIVITY RECOGNITION USING DEEP NETWORKS".** *During the internship period his conduct at* **DELHI TECHNOLOGICAL UNIVERSITY, DELHI** *was good.*

Place: New Delhi

Date:

Dr. Dinesh K. Vishwakarma
Department of Information Technology

ACKNOWLEDGEMENT

The realization and absolute conclusion of this summer internship involved a lot of guidance and assistance from numerous people and I am really fortunate to have got this all along the internship.Whatever I have prepared is because of this guidance and assistance and it is worth mentioning my thanks to them.

My grateful thanks is extended to Delhi Technological University for providing me the platform to go in for this work.I would also like to acknowledge the Information Technology department for providing me with useful and constructive resources.

I express my sincere gratitude to Dr. Dinesh Kumar Vishwakarma (Associate Professor) for providing me an opportunity to undergo summer internship at Delhi Technological University, New Delhi.

I am very thankful to Ms. Chhavi Dhiman for her support,cooperation and motivation provided to me during the internship for constant inspiration,presence and blessings.

Paras Agarwal

Department of Computer Science and Engineering

National Institute of Technology, Raipur

Chapter 1
INTRODUCTION

In recent years human activity recognition has been widely used in various fields such as human-machine interaction, intelligent surveillance, robotics, health care, gaming, security purpose, video analysis. However, human action recognition is a challenging research topic due to non-rigid nature of human body and complexity of body movements. Visual-based action recognition can be implemented in either single-modal (color, depth, skeletal) or multi-modal schemes. With the popularity of low-cost depth sensors such as Microsoft Kinect and Asus Xtion, skeletal data can now be generated from depth data by reliable pose estimation techniques. Skeletal data extracted from depth images are more discriminative in representing actions and require much less storage than traditional color image data. Therefore, skeleton-based action recognition has become an attractive research topic in recent years [1] [2]. Current approaches for skeleton-based action recognition can be roughly divided into two main categories. The first category uses hand-crafted features while the second category investigates deep learning methods to automate feature extraction process. Deep learning based techniques usually require large datasets and high performance computing hardware. Among hand-crafted descriptors for action representation, Cov3DJ with covariance matrix of 3D joint positions proves its effectiveness and computational efficiency [3] . Lots of advanced approaches have been proposed [4–6], especially deep learning methods like Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM). These RNN-based methods tend to overstress the temporal information [7]. However, for a given skeleton sequence, there are two important factors to recognize action classes: one is the description of the spatial structure of skeleton joints, and the other is to extract temporal information among multiple frames of the sequence. Hence, the combination of spatial and temporal information is the most effective representation.So, in my work I apply 2D CNN on the sample image representing both spatial and temporal information of a particular action.

Chapter 2
RELATED WORK

Single Handcrafted feature representation based action recognition as in [9][10][11] and combined hand-crafted features based as in [12].Feature designing after keenly analysing the datasets.Feature designing is time taking but action recognition using these features is fast process comparatively.In [12] From a joints data spatial features and temporal features were extracted and then the extracted features were combined using metric learning method.In [13] a method was proposed in which they divided skeleton into five parts(right arm,left arm,right leg,left leg and spine).Each part containing four joints. For each body part Covariance matrix were computed and then to create a feature vector they combined covariance matrices. Using LSTM (Long Short Term Memory), the long term and short term dependencies among the frame sequences can be learnt that's why many researches choose LSTM for feature learning.In[14] the relative motion between skeleton joints encoded using end-to-end hierarchical RNN. The local features were extracted from each part of skeleton by passing through independent subnet. RNN-based methods focus more on the temporal information as proposed in [15]. In [16] they first use (LSTM-AE) LSTM Auto-encoder network in order to store both spatial and temporal information in skeleton frame sequences and then it had been integrated with RNN based models after that loss function(Regularised cross entropy) of LSTM-AE were intoduced.In another approach deep learning based techniques were used to perform on large datasets to yield better results. In [17] 3D CNN were used to learn information stored in spatial - temporal well encoded representations of skeleton sequences.In it two parallel streams of 3D CNN were used which will compensate the spatial-temporal information mutually.

Chapter 3
PROPOSED WORK

In my proposed work to understand the videos and classify the actions, a deep frame work is proposed for human action recognition using skeletons.The framework is broadly divided in four steps which are as follows:

1. Preprocessing
2. Feature Extraction
3. Feature Learning
4. Classification

The block diagram of the proposed framework is shown in Fig.1

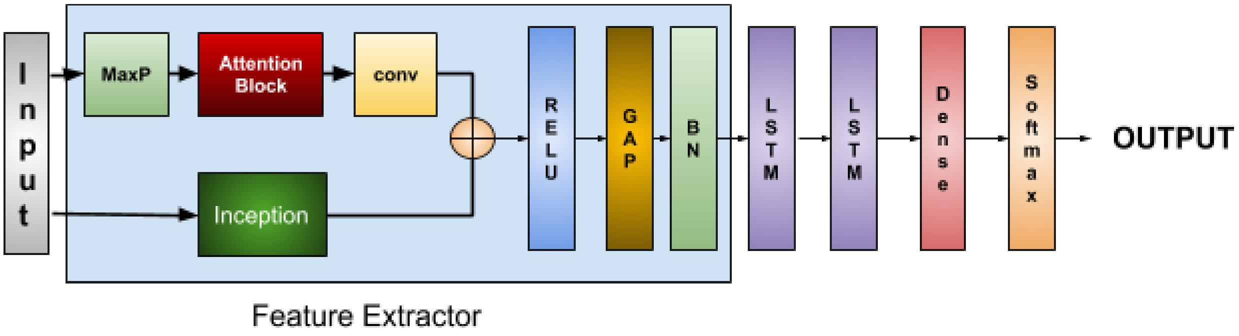

Fig. 1. ***RIALNet Architecture***

1. PREPROCESSING

To utilise the 3D coordinates of human skeleton joints, they are initially projected in 3D space to view the spatial representation of skeletons.The experimentation is divided in two phases using i) Complete Skeletons and ii) Partwise Skeleton. For this purpose the complete skeletons with 20/15 joints (provided in the dataset) of the entire action sequence are stacked in one frame which preserves both spatial and temporal information of the action. For partwise skeleton based analysis the complete skeleton is divided into five parts i) Head Spinal (HS) ii) Left Hand (LH) iii) Right Hand (RH) iv) Left Leg (LL) v) Right Leg (RL) . Each part is individually processed and weighted late fusion is applied to recognise the action in the video.

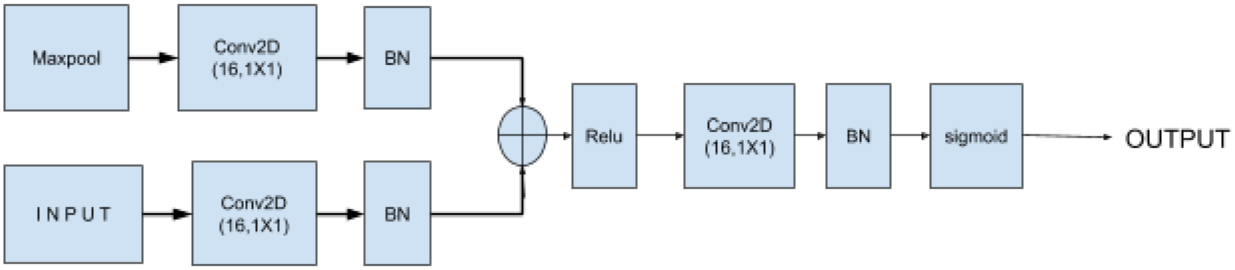

Fig. 2. Attention Block

2. FEATURE EXTRACTION

In order to extract the features from the processed skeleton sequences of actions a novel feature extractor is defined as shown in Fig. 1. Residual Nets are easier to train as opposed to other CNN architectures e.g. VGGnet. For example, a 152-layer ResNet which is 8 times deeper than VGGnet, is still less complex and trains faster. Very deep networks are known to cause overfitting and saturation in accuracy. However, residual learning and the identity mappings (shortcut connections) [16] in ResNets have been shown to overcome these problems. This enables ResNets to achieve outstanding results in image detection, localization and segmentation tasks [17]. In this paper, we utilise the residual concept while extracting discriminating features of an image. In ***RialNet Architecture ,*** we first find the attentive regions in an image, using attention block, Fig.2, which hold more attention than other regions in an image and use this attention ROI instead of complete image as a residue, which is added with image features extracted by using the power of convolution filters. Different combination of convolutions filters of different sizes (1X1), (3X3), (5X5), as shown in Fig. 3, help to extract different kinds of features from the same sample image tensor, called Inception block.

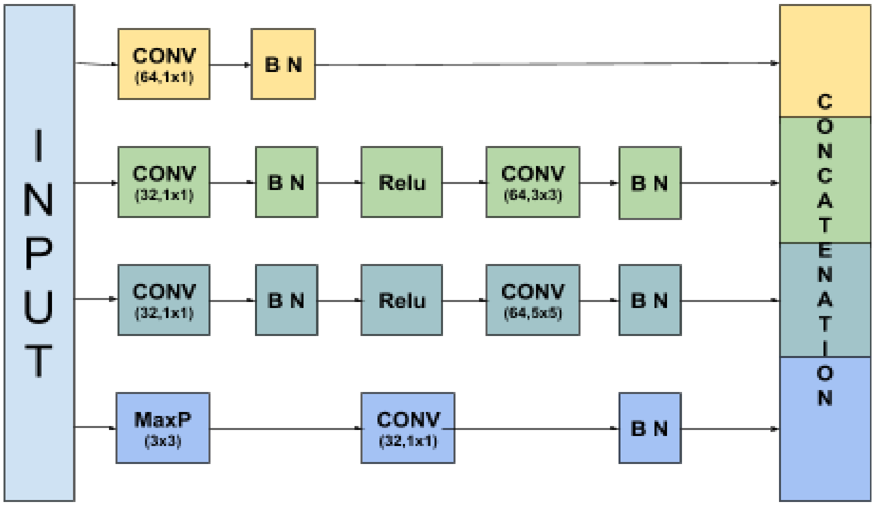

Fig. 3. Inception based Feature Extractor

3. FEATURE LEARNING

The Processed features so obtained are then learned for each sample using LSTM (Long Short Term Memory). It is mainly used for capturing the weights assigned to the neurons of deep layer and to learn the pattern in the sequences by holding, forgetting some of the learned information. In ***RialNet Architecture,*** two LSTM layers are used at the end for learning the features extracted after applying global average pooling (GAP) on the output obtained after applying the residual concept on attention specific image region and inception based extracted image features.

4. CLASSIFICATION

For classification the learned features are dense to (1xn), n number of classes layer and softmax function is applied on it.

Chapter 4
DATASETS

To validate the performance of the proposed frame work, experiments are performed on two publicly available datasets - UT Kinect and Florence 3D actions Datasets.

i) UT Kinect Action-3D Dataset -

This dataset was collected as part of research work on action recognition from depth sequences. The research is described in detail in CVPRW 2012 paper View Invariant Human Action Recognition Using Histograms of 3D Joints.The videos was captured using a single stationary Kinect with Kinect for Windows SDK Beta Version. There are 10 action types: *walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, clap hands.* There are 10 subjects, Each subject performs each actions twice. Three channels were recorded: RGB, depth and skeleton joint locations. The three channel are synchronized.

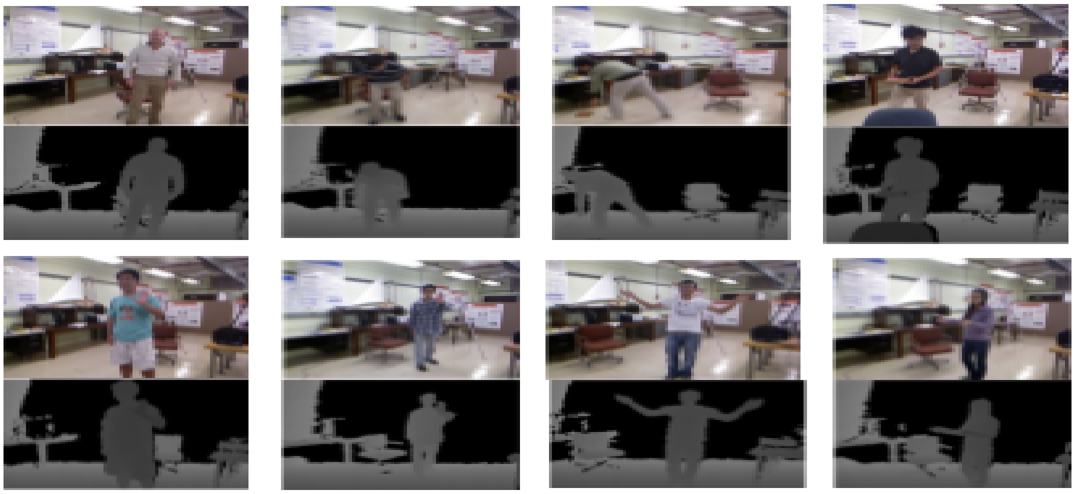

Fig. 4 Sample images of UT Kinect Action 3D Dataset

The framerate is 30f/s. The sample frames of the dataset are shown in Fig. 4.

ii) Florence 3D actions Dataset-

The dataset collected at the University of Florence during 2012 [8], has been captured using a Kinect camera. It includes 9 activities: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, bow. During acquisition, 10 subjects were asked to perform the above actions for 2/3 times. This resulted in a total of 215 activity samples. As an example, frames in Fig. 5 are extracted from a read watch sequence used for test (upper line), and from read watch training sequences (lower line) of this dataset.

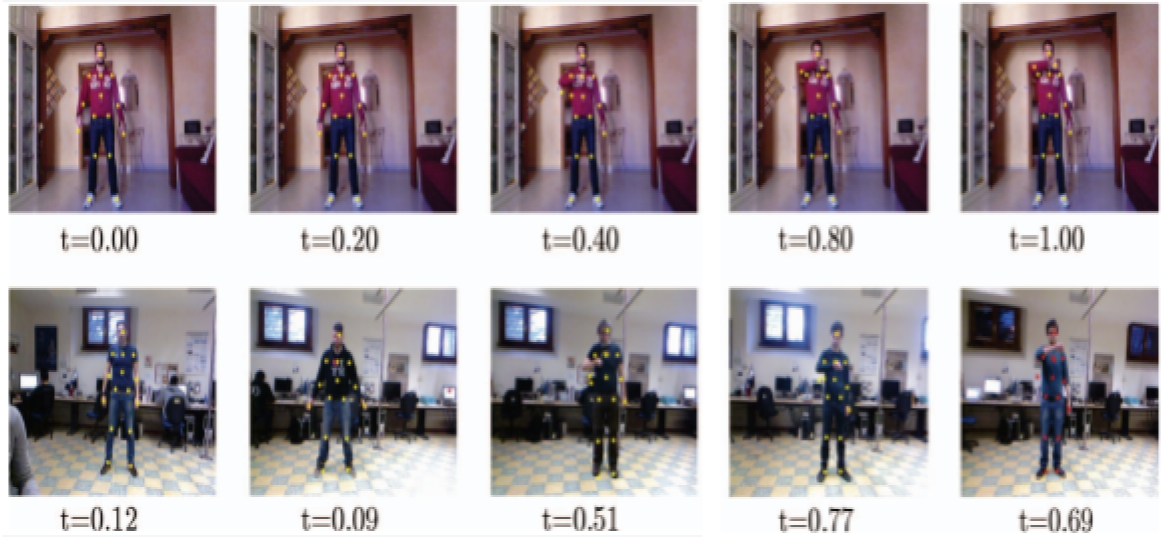

Fig. 5 Sample images of UT Kinect Action 3D Dataset

Chapter 5
EXPERIMENTAL RESULTS

The performance is measured in terms of accuracy and the response is observed using top1, top3, top5 matrices. The results for UT Kinect and Florence dataset are shown in Table 1 and Table 2. It is observed from both table1 and table2 that the partwise analysis of spatial and temporal information of skeleton sequences is more efficient than complete skeletons. It supports the fact that local features of the shape are more robust than global features. Top3 and Top5 based accuracy analysis help to understand the preciseness of the proposed framework to recognise the action. And from the results it is observed that 100% weighted Top5 accuracy and 98.63% weighted Top5 accuracy for UT Kinect and Florence 3D Action Datasets respectively is observed. A comparison of the proposed framework for skeleton based human action

Table 1: Performance of the proposed framework for UT Kinect dataset

UT Kinect	**Training loss**	**Training Accuracy**			**Validation Accuracy**		
		top1	**top3**	**top5**	**top1**	**top3**	**top5**
Full Skeleton	0.2836	0.8903	0.9940	0.9996	0.5018	0.8768	**0.9771**
Head Spinal	0.4480	0.8388	0.8978	0.9294	0.5578	0.8442	**0.9749**
Left Leg	0.4697	0.8205	0.8944	0.9330	0.6000	0.8450	**0.9400**
Right Leg	0.3209	0.8760	0.9361	0.9765	0.5736	0.8426	**0.9645**
Left Hand	0.3164	0.8655	0.9308	0.9713	0.6142	0.8528	**0.9594**
Right Hand	0.3801	0.8498	0.9146	0.9592	0.6281	0.8442	**0.9648**
Weighted fusion	W_hs,w_ll,w_rl,w_lh,w_rh **(2,3,4,4,5)**				**0.94**	**1.00**	**1.00**

recognition in videos is outlined in table 3 and 4 for UT Kinect Action 3D Dataset and Florence 3D Dataset.

Table 2: Performance of the proposed framework for Florence 3D Action Dataset

Florence	Training loss	Training Accuracy			Validation Accuracy		
		top1	top3	top5	top1	top3	top5
Full Skeleton	0.1221	0.9578	1.0000	1.0000	0.3699	0.8014	**0.9589**
Head Spinal	0.3700	0.8537	1.0000	1.0000	0.5959	0.9315	**1.0000**
Left Leg	0.0279	0.9867	1.0000	1.0000	0.3699	0.7808	**0.9247**
Right Leg	0.0630	0.9829	0.9969	0.9969	0.4658	0.8151	**0.9589**
Left Hand	0.6868	0.8240	0.9630	0.9665	0.5253	0.8065	**0.9263**
Right Hand	0.0127	0.9941	1.0000	1.0000	0.4022	0.7663	**0.9185**
Weighted fusion	W_hs,w_ll,w_rl,w_lh,w_rh **(3,4,2,3,2)**				**0.8698**	**0.9452**	**0.9863**

Table 3.Comparison of the proposed framework with other state of the arts for UT Kinect 3D Action Dataset

Methods	**Protocol**	**Accuracy(%)**
Feature combination [16]	LOOCV	90.88
St-LSTM+trust gate [18]	LOOCV	95.23
Grassmann Manifold [19]	LOOCV	87.04
JLd+RNN [20]	cross validation	92.25
TS-LSTM [21]	Cross validation	94.39
Lie groups [17]	Cross validation	95.37
Kernel Linearization [22]	Cross validation	98.2
LRCNLG [26]	LOCCV	98.5
Proposed	**Cross validation**	**94%(Top1)**
		100%(Top3)
		100%(Top5)

Table 4.Comparison of the proposed framework with other state of the arts for florence 3D Action Dataset

Methods	**Protocol**	**Accuracy(%)**
Lie groups [17]	Cross validation	90.88
Kernel Linearization [23]	Cross validation	95.23
Shape analysis [24]	LOOCV	87.04
Mining key pose [25]	LOCCV	92.25
Feature combination [16]	LOOCV	94.39
LRCNLG [26]	LOOCV	95.37
Proposed	**Cross validation**	**86.98(Top1)**
		94.52(Top3)
		98.63(Top5)

Chapter 6
CONCLUSION

This report is presenting a novel Rial Network model for action recognition based on skeleton sequences. The proposed Attention and Inception stream can learn more motion details of local and global by individual stream’s mutual enhancement. Meanwhile, the simple yet effective RialNet architecture overcomes the overfitting problem. Experimental results show that our method outperforms most of state-of-the-art RNN-based approaches and 3D CNN models and also verify the effectiveness of using 2D CNN learn the processed skeleton data. And the multitemporal version do increase the ability of RialNet model to capture multi-scale information. In the future, in order to train Rial more effectively, we will focus on different ways of encoding skeleton data. We are also planning to pass skeleton frame sequences data with its riemannian manifold which will surely enhances its performance.

References

[1] A. Yao, J. Gall, G. Fanelli, and L. Van Gool, "Does human action recognition benefit from pose estimation?"," in Proceedings of the 22nd British machine vision conference-BMVC 2011, 2011.

[2] L. L. Presti and M. L. Cascia, "3d skeleton-based human action classification: A survey," Pattern Recognition, vol. 53, pp. 130 – 147, 2016.

[3] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in Proceedings of the TwentyThird International Joint Conference on Artificial Intelligence, ser. IJCAI '13. AAAI Press, 2013, pp. 2466–2472.

[4] J. Aggarwal and X. Lu, "Human activity recognition from 3D data: A review," PRL, vol. 48, pp. 70–80, 2014.

[5] L. Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," PR, vol. 53, pp. 130–147, 2016.

[6] J. Zhang, W. Li, P. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey," PR, vol. 60, pp. 86–105, 2016.

[7]Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in Proc. ACM on Multimedia Conference (ACMMM), 2016, pp. 102–106.

[8] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," in 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, June 2012, pp. 8–13.

[9] L. Tao and R. Vidal, "Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition," in 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Dec 2015, pp. 303–311.

[10] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 1290– 1297.

[11] D. C. Luvizon, H. Tabia, and D. Picard, "Learning features combination for human action recognition from skeleton sequences," Pattern Recognition Letters, vol. 99, pp. 13 – 20, 2017, user Profiling and Behavior Adaptation for Human-Robot Interaction.

[12] H. A. El-Ghaish, A. Shoukry, and M. E. Hussein, "Covp3dj: Skeletonparts-based-covariance descriptor for human action recognition," in Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP, Funchal, Madeira, Portugal, January 27-29, 2018., 2018, pp. 343–350.

[13] Yong Du, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1110–1118.

[14] J. Tu, H. Liu, F. Meng, M. Liu and R. Ding, "Spatial-Temporal Data Augmentation Based on LSTM Autoencoder Network for Skeleton-Based Human Action Recognition," *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, 2018, pp. 3478-3482

[15] J. Tu, M. Liu and H. Liu, "Skeleton-Based Human Action Recognition Using Spatial Temporal 3D Convolutional Neural Networks," *2018 IEEE International Conference on Multimedia and Expo (ICME)*, San Diego, CA, 2018, pp. 1-6.

[16] D. C. Luvizon, H. Tabia, and D. Picard, “Learning features combination for human action recognition from skeleton sequences,” Pattern Recognition Letters, 2017.

[17] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2014, pp. 588–595.

[18] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in IEEE conference on computer vision and pattern recognition (CVPR), 2015, pp. 1110–1118.

[19] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, “Accurate 3d action recognition using learning on the grassmann manifold,” Pattern Recognition, vol. 48, no. 2, pp. 556 – 567, 2015.

[20] S. Zhang, X. Liu, and J. Xiao, “On geometric features for skeletonbased action recognition using multilayer lstm networks,” in IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017, pp. 148–157

[21] I. Lee, D. Kim, S. Kang, and S. Lee, “Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks,” in 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017, pp. 1012–1020.

[22] P. Koniusz, A. Cherian, and F. Porikli, “Tensor representations via kernel linearization for action recognition from 3d skeletons,” in European Conference on Computer Vision. Springer, 2016, pp. 37–53

[23] P. Koniusz, A. Cherian, and F. Porikli, “Tensor representations via kernel linearization for action recognition from 3d skeletons,” in European Conference on Computer Vision. Springer, 2016, pp. 37–53.

[24] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. D. Bimbo, “3-d human action recognition by shape analysis of motion trajectories on riemannian manifold,” IEEE Trans. Cybernetics, vol. 45, no. 7, pp. 1340–1352, 2015.

[25] C. Wang, Y. Wang, and A. L. Yuille, “Mining 3d key-pose-motifs for action recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2639–2647

[26] M. Rhif, H. Wannous, I. R. Farah, “Action Recognition from 3D Skeleton Sequences using Deep Networks on Lie Group Features”, 24th International Conference on Pattern Recognition (ICPR), 2018