Skeleton Based View Invariant Human Action Recognition using Convolutional Neural Networks

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

Abstract: ***The skeletal based human action recognition has its significant applications in the field of human computer interaction and human recognition from surveillance videos. However, the tasks suffers from the major challenges like view variance and noise in the data. These problems are limiting the performance of human action recognition. This paper focuses to solve these problems by adopting sequence based view invariant transform to effectively represent the spatio-temporal information of the skeletal data. The task of human action recognition in this paper is performed in three stages. Firstly, the raw 3D skeletal joint data obtained from the Microsoft Kinect sensor is transformed to eliminate the problem of view variations on a spatio-temporal data by implementing sequence based view invariant transform. In the second stage, the transformed joint locations of the skeletal data will be converted to RGB images by a color coding technique and forms a transformed joint location maps (TJLMs) . As a third stage, the discriminating features were extracted by the novel CNN architecture to performs the human action recognition task by means of class scores. Noticeable amount of recognition scores are achieved. Extensive experiments in four difficult 3D action datasets constantly show our method's superiority. The performance of the proposed method is compared with the other state-of-the-art methods.***

Keywords* : *Human action recognition, Sequence based view invariant transform, Convolutional neural networks.

I. INTRODUCTION

Human action recognition is a demanding area of research, which finds its applications in intelligent surveillance, gaming control and human-computer interaction. Previous works acknowledge RGB information activities involving complicated illumination and confused backgrounds. The fast developments in capturing depth information technology in real time have shown increasing interest in solving these issues through the use of depth sensor produced data [1] in particular by the economical Microsoft Kinect sensor[2].

Compared to RGB information, structured light sensors generate depth information that is more solid in light modifications because infrared radiation estimates the depth values without linking them to visible light. The deepening process is much more easy to remove the foreground from the cluttered background, as confusing texture and color data are ignored. In addition, RGB-D cameras provide detailed and accurate resolution and information on the structure of objects on the scene, with approximate resolution and precision.

The human body can be intuitively portrayed as a articulated system with hinged joints with rigid bones and human action can also be identified as skeleton movements. In real-time, many [3,4] projects in a skeleton-based action assessment were carried out with the application from Kinect capture models. These works are generally intended from a single point of perspective for analogous action. However, while observing an action sequence, a general and reliable action identification scheme for practical purposes must be robust to distinct points of view. This article creates a view-independent measurement method which aims to remove the effects of variable viewpoints and proposes a compact, yet discriminative, representation of the skeleton sequence.

To achieve view invariant on skeletal joint positional data and orientation a sequence based transformation is applied. As the depth sensor i.e. Kinect sensor is fixed at one position while capture, the orientation of the Kinect sensor can be identified by one transform matrix. All torso joints are considered to form transform matrix, which can able to eliminate the noise effect on the skeletal data. The previously existing methods transformed the skeletal data by the transition matrix generates by its own. The limited skeleton joints present in the data may lead to noisy data transformation. Due to this the spatio-temporal relationships between the skeletal joints are greatly effected and leads to misclassification. Compared to other existing transformation methods, the sequence based skeletal transformation greatly reduces the effect of noise on skeletal data and retains the spatio-temporal relation alive for better recognition of actions. The entire process of action recognition from skeletal data has been achieved in the following three steps:

1) Applying the sequence based transformation on raw skeletal data captured via Microsoft Kinect sensor.

2) Creation of color coded maps on the transformed skeletal data.

3) Implementing a sophisticated CNN algorithm on color coded maps to recognize the action.

The entire process of action recognition pipeline proposed in this paper can be visualized in figure 1. The complete action consisting of N number of frames with J joints can be represented as a color coded map of size $N \times 3J$.

Revised Manuscript Received on July 12, 2019.

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

P. V. V. Kishore, Professor, Department of ECE, K L University, KLEF, Guntur, India. E-mail: pvvkishore@kluniversity.in

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

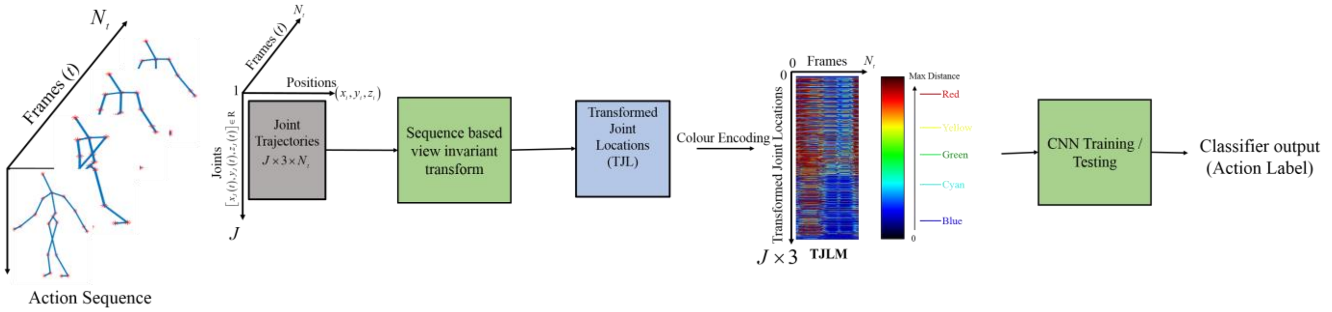

Fig. 1. Pipeline of the proposed method.

The proposed frame work is depicted in figure 1, where the color coded images prepared from the skeletal data are inputted to CNN models for action class identification. The color coding technique used here will highlight the spatial and temporal information and immune to view variations and noise interruption. The rest of the paper is organizes as follows: The next section presents a brief review on related work. The detailed methodology which includes the sequence based transform and CNN architecture is presented in section III. Section IV reports the achieved experimental results and section V concludes the paper by presenting the key investigations done and future of this work.

II. RELATED WORK

The cost-effective depth sensor with real-time algorithms for skeleton estimation can present comparatively accurate joint coordinates. Efficient and effective methods have been created lately, based on these coordinates for recognition of action. In order to be able to recognize sequences oriented actions, temporal dynamics of posture over moment may be modelled as a time-series issue. Skeleton joint co-ordinates can be used as a type of low-level function to portray human positions and their spatial progress. The majority of the current action identification relied on skeletons approach models centered on well-designed local characteristics.

Action recognition task which serves the view invariance problem has many challenges due to 1. the appearance of the skeleton changes in different viewing angles which tends to have different intra variations in the same action itself and 2. difficulty to represent the spatio-temporal relations of the particular actions effectively. Due to these two reasons the action recognition task became more complicated for researchers to solve. In [5], the authors represented the actions from the skeletal information as self-similarity matrix (SSM). Somehow this SSM representation immune to view variations, but it is not firmly view invariant. Traditional techniques design handmade functions to portray spatio-temporal skeleton joints to model temporal evolution through time series simulations. The zenith reference vectors are considered in [6] to implement view invariant human action recognition with little success. Some authors translated the skeletal data in to a new coordinate system and achieved view invariance to absolute human body orientations [7]. The authors in [8], calculated the principal components for the reference points of a human body by selecting the zenith reference point as first principal component. Generally, the zenith reference is always perpendicularly aligned with the torso point with larger dimensions. However all these methods are not effective for the noisy skeletal data and leads to the greater loss in spatio-temporal information of the specific actions. The relative joint differences are calculated in [9] to join the static skeletal information with all joint dynamics of the skeleton. The PCA is applied on the calculated joint differences to get EigenJoint representation, which leads to reduction in noise and redundancy in the spatio-temporal skeletal data. The relative motions among the skeletal joints are modelled by the hierarchical recurrent neural networks [10]. The RNN based methods felt difficulty to process the long sequence temporal information. However, the LSTM can able to learn the patterns from the long skeletal sequences by a gating mechanism. The authors in [11] proposed a spatio-temporal LSTM which can learn both spatial and temporal dynamics of all joints in a skeleton. Moreover, the RNN and LSTM based methods are more aggravate the temporal dynamics. However, the CNN based methods playing a vital role in producing a promising results for human action recognition from an image. In [12], the authors implemented the human action recognition task from the skeletal data by projecting the local coordinates on to orthogonal planes. The 2D trajectories on each plane are turned into a color image and inputted to the CNN. The skeletal joint information in all the frames is concatenated and the each x, y, z is color coded in to R, G, B respectively to form an image which implicitly represent the spatio-temporal evolution of an action [13]. As the CNNs has the power to explore various features of an image automatically, this work choose to represent the spatio-temporal dynamics of human skeleton action sequences. The view invariance is achieved by transforming the skeletal joint coordinate information using a sequence based view invariant transform. The transformed skeletal joint coordinates are encoded in to an RGB image and inputted to a CNN architecture for effectively classifying the human actions.

III. PROPOSED METHODOLOGY

This section discusses the methodologies adopted to solve view invariant problem in recognizing human actions from the skeletal 3D joint locations captured from the Kinect sensor.

A. *Sequence-based view invariant transform*

The skeletal sequences are in general suffers from view variations, which in turn facing a problem in identifying an action. Several traditional methods were proposed by the previous researchers are not up to the mark. In [14] authors transformed the skeletal sequences in order to solve this issue. However, this practice sometimes damaging the relative motion relationship among the raw skeletal joints. At this juncture, the sequence based view invariant transform proved its ability to attain view invariance by transforming all skeletal joints synchronously over the entire skeletal sequence.

A 3D skeletal action sequence A with J joints and T frames is expressed as $\mathrm{A}=\{P_1,P_2,.....,P_T\}\in\mathbb{R}^{J\times3\times T}$. The J joints form a joint set with position vectors $p_i=[p_1,p_2,...p_J]$. The i^{th} joint in a 3D space is a 3D coordinate of $p_i(x_i,y_i,z_i)\in\mathbb{R}^{3\times J}\ \forall\ i=1\ to\ J$. The i^{th} skeleton joint in the t^{th} frame is denoted as $p_i^t=(x_i^t,y_i^t,z_i^t)$.

In general the skeletal coordinates of an action sequence are not immune to view variations. Hence the raw skeletal coordinates (x_i^t,y_i^t,z_i^t) has to be transformed as view invariant values (X_i^t,Y_i^t,Z_i^t). The transformation of 3D coordinates as view invariant is done as follows.

$$\begin{bmatrix}X_i^t\\Y_i^t\\Z_i^t\\1\end{bmatrix}=\Im(\Re_x^\alpha,0)\ \Im(\Re_y^\beta,0)\ \Im(\Re_z^\gamma,0)\begin{bmatrix}X_i^t\\Y_i^t\\Z_i^t\\1\end{bmatrix}\tag{1}$$

Here, $\Im$ is the transformation matrix defined as $\Im(\Re,\tau)=\begin{bmatrix}\Re&\tau\\0&1\end{bmatrix}_{4\times4}$, where the rotation matrix $\Re\in\mathbb{R}^{3\times3}$ and the translation vector $\tau=-\frac{1}{T}\sum_{T=1}^{t}p_n^t,\ \in\mathbb{R}^3$. The translation vector sets the hip joint as origin and bring all the action frame sequences to an unique origin. The rotation has been done around x, y and z axis by an angle of α , β and γ respectively. The coordinate rotated around x axis at an angle of α is $\Re_x^\alpha$ and formulated as

$$\Re_x^\alpha=\begin{bmatrix}\cos\alpha&0&-\sin\alpha\\0&1&0\\\sin\alpha&0&\cos\alpha\end{bmatrix}\tag{2}$$

Similarly, the coordinate rotation around y axis by an angle of β degrees $\Re_y^\beta$ is given as

$$\Re_y^\beta=\begin{bmatrix}1&0&0\\0&\cos\beta&-\sin\beta\\0&\sin\beta&\cos\beta\end{bmatrix}\tag{3}$$

And the coordinate rotation around z axis by an angle of γ degrees $\Re_z^\gamma$ is given as

$$\Re_z^\gamma=\begin{bmatrix}\cos\gamma&\sin\gamma&0\\-\sin\gamma&\cos\gamma&0\\0&0&1\end{bmatrix}\tag{4}$$

The transformed view invariant parameters X_i^t,Y_i^t and Z_i^t together provides the view invariant coordinates of the joint, which can be now color coded to form a transformed joint location map (TJLM) for the recognition of human actions by eliminating the view variance problem.

B. *Color encoding procedure*

Convolutional nets receive images as input for training and testing. Hence, the transformed joint location (TJL) matrix for T frames considering all the joints of size $T\times(J\times3)$ must be coded into an RGB image. The color coding is kept simple, using the 'jet' color map to encode the TJLM's with the following standard mapping procedure [15], as

$$R=\begin{cases}\dfrac{d_{\measuredangle ij}^T+1}{n} & d_{\measuredangle ij}^T\le n-1\\ 1 & d_{\measuredangle ij}^T>n-1\end{cases}$$

$$G=\begin{cases}0 & d_{\measuredangle ij}^T<n-1\\ \dfrac{d_{\measuredangle ij}^T+1-n}{n} & n-1<d_{\measuredangle ij}^T\le 2n-1\\ 1 & d_{\measuredangle ij}^T>2n-1\end{cases}$$

$$B=\begin{cases}0 & d_{\measuredangle ij}^T\le 2n-1\\ \dfrac{d_{\measuredangle ij}^T+1-2n}{m-2n} & d_{\measuredangle ij}^T>2n-1\end{cases}\tag{5}$$

Where, m is number of colors in jet map and $n=round\left(f\left(\frac{3}{8}m\right)\right)$ is scale rounded to nearest integer towards zero. Concatenating the 3-color planes into one, creates a RGB image of joint angular displacement as intensity values. The variable dimensionally in human subjects is handled using normalization at the skeletal stage or at the image level scaling. The figure 2 shows some of the transformed joint location maps (TJLMs) created.

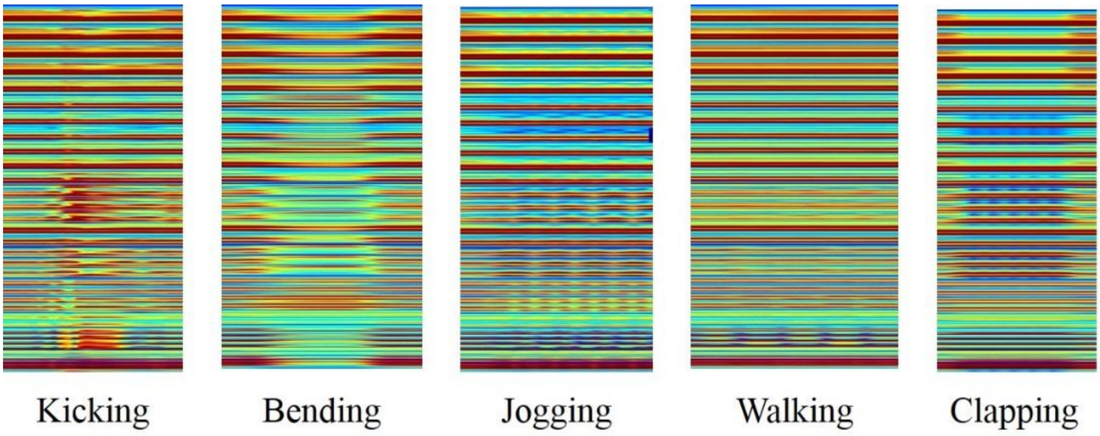

Fig. 2. Visualization of TJLMs for sample actions.

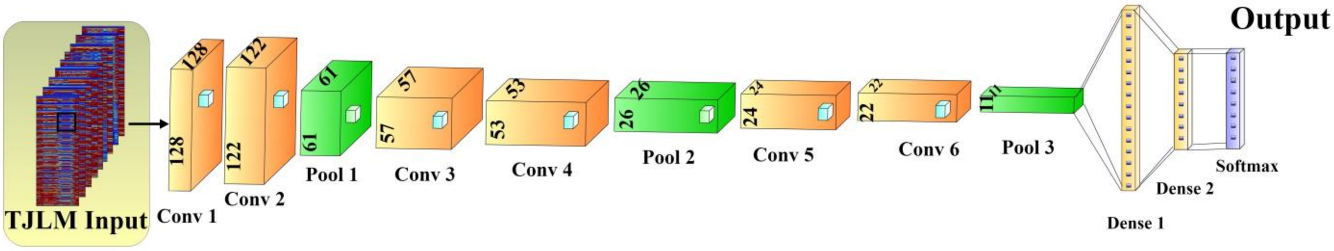

Fig. 3. The proposed CNN architecture.

C. *Proposed CNN architecture*

The proposed CNN is inspired by VGG network architecture introduced by Simonyan and Zisserman [16] is a very deep CNN model which achieved the state-of-the-art accuracy on ILSVRC (Large Scale Visual Recognition Challenge, 2014) classification and localization tasks. VGG net is deep layered CNN with 16 to 19 weight layers with small window sizes of 3×3 followed throughout the convolutional layers. The architecture is no different from that of the originally proposed CNN architectures by Ciresan et al. [17] and Jeffrey Dean et al. [18]. The proposed 3D sign nets architecture is inspired by VGG net. However, the depth is limited to 6 weight layers and 2 fully connected layers. The architecture is being built using python with the help of Keras and tensorflow libraries. Our proposed CNN architecture is exhilarated from VGG, but with 8 layers. Six convolutional layers followed by 2 fully connected layers is what is arrived after multiple testing using different network models in VGG, AlexNet, ResNet and Inception. All are developed from scratch using Keras and Tensorflow in Python 3.6. The proposed CNN architecture is shown in figure 3.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed method of human action recognition task is tested on various publicly available skeletal datasets namely, NTU RGB-D [19], MSR Action 3D [20], UT-Kinect Action 3D [21] and G3D [22]. The created TJLMs are inputted to a VGG based CNN architecture, which trained and tested on various datasets individually. The achieved recognition rates can be observed from the table I. To know the novelty of the method in solving the view variance problem, it is tested against the cross view data. The results shows that the method provides view invariance and provides better recognition rates i.e. an average of 87% on cross view data. Similarly, an average of 86% if recognition rate is achieved on cross subject data.

Table- I: Recognition rates achieved through the proposed method.

Datasets	Recognition Rates (%)			
	Raw Skeletal Data		Transform Skeletal Data	
	Cross View	Cross Subject	Cross View	Cross Subject
NTU RGB-D	64.87	68.49	80.57	82.37
MSR Action 3D	67.32	70.17	82.07	85.97
UT-Kinect Action 3D	73.27	76.84	86.27	90.65
G3D	70.94	72.98	84.37	88.37

A. *Evaluation on NTU RGB-D dataset*

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. The NTU RGB-D data is trained and tested using the proposed method and a reasonable classification rates were achieved.

To further know the robustness of the proposed method, the results were compared with the other state-of-the-art algorithms and tabulated in table II.

Table- II: Recognition rates achieved on NTU RGB-D data through the proposed and other state-of-the-art algorithms.

Datasets	Methods	Recognition Rates (%)			
		Raw Skeletal Data		Transform Skeletal Data	
		Cross View	Cross Subject	Cross View	Cross Subject
NTU RGB-D	Adaptive kernel matching (AKM) [23]	59.79	62.46	74.89	79.57
	Dynamic time wrapping (DTW) [24]	58.27	60.27	71.27	77.24
	Histograms [25]	55.78	59.37	69.84	74.37
	Recurrent neural network (RNN) [26]	61.75	63.87	77.28	81.21
	Long short-term memory (LSTM) [27]	62.15	65.94	78.14	81.37
	Proposed method	64.87	68.49	80.57	82.37

The precision and recall values are also calculated on NTU RGB-D data using the proposed and other state-of-the-art methods. Figure 4 plots the precision recall values. From the figure 4, it can be seen that the proposed algorithm is best performing in solving view variance problem on NTU RGB-D data.

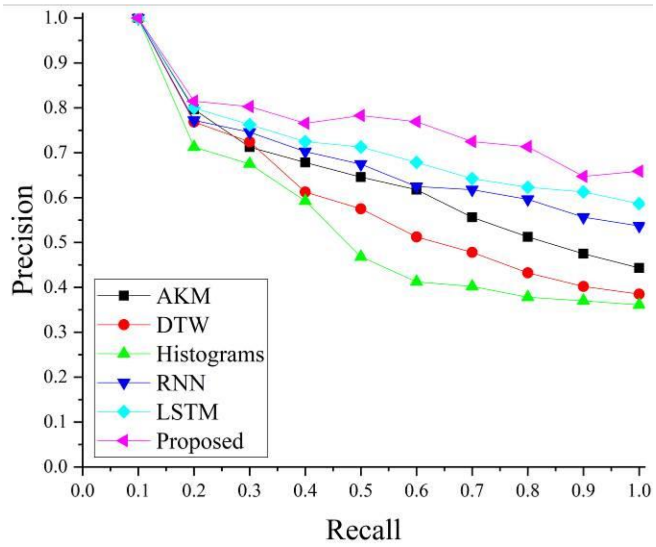

Fig. 4. Precision-Recall plots of NTU RGB-D action classification using various algorithms.

B. *Evaluation on MSR Action 3D dataset*

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. The organized MSR Action 3D data is trained and tested on the proposed CNN architecture.

The cross view recognition is observed as 82.07 % on a transformed skeletal data, i.e. the recognition rate is improved by 15 % when compared to the recognition rates obtained on raw skeletal data. Similarly, the cross subject testing is performed and better recognition rates were noted. To further know the novelty of the method, it is considered to test the TJLMs on other state-of-the-art algorithms, the resulted recognition rates can be studied from table III. The performance of the method is further studied by drawing the precision-recall plots for various methods and visualized in figure 5.

Table- III: Recognition rates achieved on MSR Action 3D data through the proposed and other state-of-the-art algorithms.

Datasets	Methods	Recognition Rates (%)			
		Raw Skeletal Data		Transform Skeletal Data	
		Cross View	Cross Subject	Cross View	Cross Subject
MSR Action 3D	Adaptive kernel matching (AKM) [23]	62.89	65.84	77.67	81.27
	Dynamic time wrapping (DTW) [24]	61.82	63.27	75.29	78.69
	Histograms [25]	60.74	61.97	73.54	75.67
	Recurrent neural network (RNN) [26]	64.28	66.57	81.07	82.14
	Long short-term memory (LSTM) [27]	65.97	67.37	80.97	82.91
	Proposed method	67.32	70.17	82.07	85.97

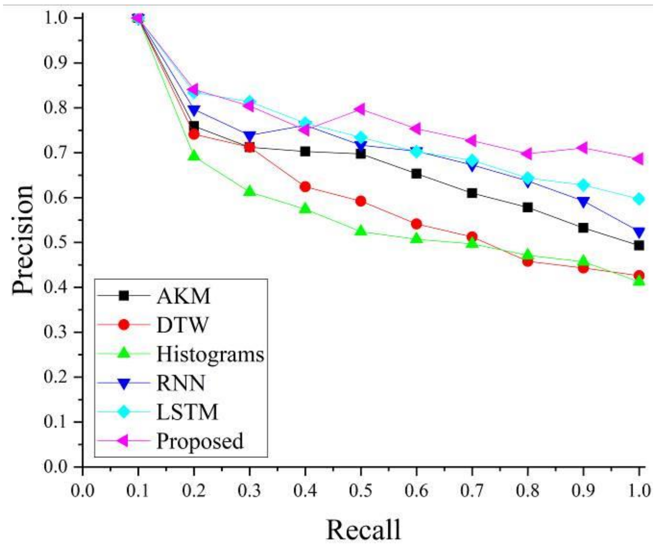

Fig. 5. Precision-Recall plots of MSR Action 3D action classification using various algorithms.

C. *Evaluation on UTKinect-Action3D dataset*

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. From the table IV, it can be observed that the proposed method is also working with acceptable recognition rates i.e. around 86.27 % on cross view data and 90.65 % on cross subject data.

Figure 6 shows the precision-recall plots, which compares the performance of the proposed method with other state-of-the-art classification algorithms.

Table- IV: Recognition rates achieved on UT-Kinect Action 3D data through the proposed and other state-of-the-art algorithms.

Datasets	Methods	Recognition Rates (%)			
		Raw Skeletal Data		Transform Skeletal Data	
		Cross View	Cross Subject	Cross View	Cross Subject
UT-Kinect Action 3D	Adaptive kernel matching (AKM) [23]	68.93	71.95	81.37	86.97
	Dynamic time wrapping (DTW) [24]	67.28	70.15	80.12	83.94
	Histograms [25]	64.29	67.94	78.91	81.21
	Recurrent neural network (RNN) [26]	70.48	74.01	84.94	88.14
	Long short-term memory (LSTM) [27]	71.94	74.29	85.11	88.37
	Proposed method	73.27	76.84	86.27	90.65

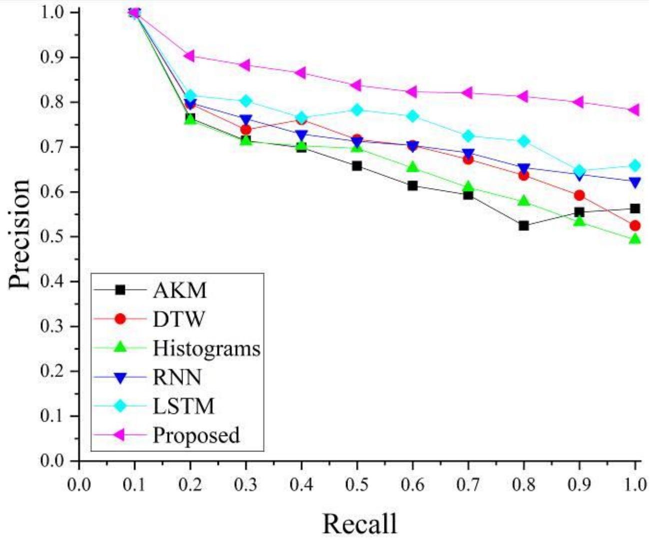

Fig. 6. Precision-Recall plots of UT-Kinect Action 3D action classification using various algorithms.

D. *Evaluation on G3D gaming action dataset*

The further experimentation is done by considering the most popular and less complexity gaming dataset named G3D. 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 V shows the recognition rates achieved on G3D data using the proposed and other methods. An average of 84.37 % of recognition rate is noticed on a transformed skeletal data training with cross view testing. A better improvement is achieved by implementing sequence based view invariant transformation on a raw skeletal data and TJLMs with CNN training.

Table- V: Recognition rates achieved on G3D (Gaming action 3D) data through the proposed and other state-of-the-art algorithms.

Datasets	Methods	Recognition Rates (%)			
		Raw Skeletal Data		Transform Skeletal Data	
		Cross View	Cross Subject	Cross View	Cross Subject
G3D	Adaptive kernel matching (AKM) [23]	65.97	68.29	79.94	81.97
	Dynamic time wrapping (DTW) [24]	64.22	67.14	77.15	79.97
	Histograms [25]	60.19	62.18	73.18	75.18
	Recurrent neural network (RNN) [26]	68.08	69.78	81.08	84.17
	Long short-term memory (LSTM) [27]	67.91	70.37	81.84	85.74
	Proposed method	70.94	72.98	84.37	88.37

The performance of the proposed TJLMs+CNN on G3D data is further validated by plotting their precision-recall plots. Figure 7 depicts the performance of the proposed with other methods in terms of precision-recall. The proposed TJLMs+CNN outperformed among all other methods with an average recognition rate of 84.37 %, which is noted as 15 % increment compared to the observed recognition rates on raw skeletal data.

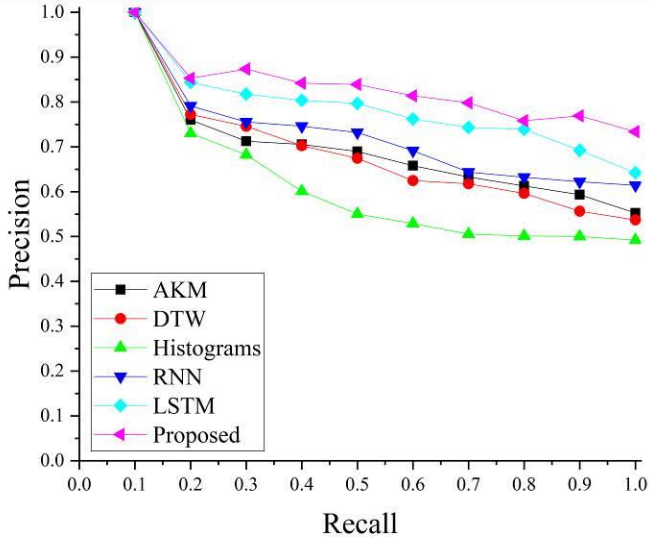

Fig. 7. Precision-Recall plots of G3D action classification using various algorithms.

E. *Evaluation of the proposed method with different input features*

The robustness of the proposed CNN architecture is tested on different features such as RGB, Depth, skeletal, RGB+Depth, RGB+Depth+Skeleton and transformed skeleton. As shown in table VI, the proposed CNN has shown better performance when compared to other state-of-the-art algorithms.

Table- VI: Comparison of recognition rates of different methods on different input features.

Type of features	Methods					
	AKM	**DTW**	**Histograms**	**RNN**	**LSTM**	**CNN**
RGB	64.23	62.19	58.76	67.99	67.49	69.37
Depth	65.29	63.89	60.82	67.23	68.29	71.92
Skeleton	67.13	65.27	62.86	68.55	69.49	72.12
RGB+Depth	70.37	68.23	65.04	72.18	73.87	75.78
RGB+Depth+Skeleton	75.37	73.18	69.97	77.99	78.07	80.96
TJLMs	**80.27**	**78.08**	**74.24**	**82.18**	**83.57**	**84.84**

V. CONCLUSION

The human action recognition with view invariance problem is attempted to solve in this paper. The problem is considered to solve by making the 3D human skeleton joint location immune to view variations with the help of sequence based view invariant transform. The transformed joint locations (TJLs) of a skeleton are encoded in to color coded maps (TJLMs) using a standardized color encoding technique. A vigorous VGG based CNN architecture is designed for the view invariant human action classification task on a color coded TJLMs. A better improvement in recognizing the actions with acceptable classification rate was achieved on a cross view data. An average recognition rate of 84 % is observed using the proposed method on four different datasets, namely NTU RGB-D, MSR Action 3D, UT-Kinect Action 3D and G3D.

REFERENCES

1. C. Chen , M. Liu , B. Zhang , J. Han , J. Jiang , H. Liu , 3D action recognition using multi-temporal depth motion maps and fisher vector, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2016, pp. 3331–3337 .
2. Z. Zhang , Microsoft kinect sensor and its effect, IEEE Multimed. 19 (2) (2012) 4–10 .
3. X. Yang , Y. Tian , Eigenjoints-based action recognition using Naive–Bayes-near- est-neighbor, in: Proceedings of the Conference on Computer Vision and Pat- tern Recognition Workshops, 2012, pp. 14–19 .
4. M. Ding , G. Fan , Multilayer joint gait-pose manifolds for human gait motion modeling, IEEE Trans. Cybern. 45 (11) (2015) 2413–2424 .
5. I.N. Junejo , E. Dexter , I. Laptev , P. Perez , View-independent action recognition from temporal self-similarities, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 172–185 .
6. L. Xia , C.-C. Chen , J. Aggarwal , View invariant human action recognition using histograms of 3D joints, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27 .
7. M. Jiang , J. Kong , G. Bebis , H. Huo , Informative joints based human action recognition using skeleton contexts, Signal Process. Image Commun. 33 (2015) 29–40 .
8. M. Raptis , D. Kirovski , H. Hoppe , Real-time classification of dance gestures from skeleton animation, in: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2011, pp. 147–156 .
9. X. Yang , Y. Tian , Effective 3D action recognition using eigenjoints, J. Vis. Com- mun. Image Represent. 25 (1) (2014) 2–11 .
10. Y. Du , W. Wang , L. Wang , Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the Conference on Computer Vi- sion and Pattern Recognition, 2015, pp. 1110–1118 .
11. J. Liu , A. Shahroudy , D. Xu , G. Wang , Spatio-temporal LSTM with trust gates for 3D human action recognition, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 816–833 .
12. P. Wang , Z. Li , Y. Hou , W. Li , Action recognition based on joint trajectory maps using convolutional neural networks, in: Proceedings of the ACM International Conference on Multimedia, 2016, pp. 102–106 .
13. Y. Du , Y. Fu , L. Wang ,Skeleton based action recognition with convolutional neural network, in: Proceedings of the Asian Conference on Pattern Recogni- tion, 2015, pp. 579–583 .
14. M. Raptis , D. Kirovski , H. Hoppe , Real-time classification of dance gestures from skeleton animation, in: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2011, pp. 147–156 .
15. P.Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, “ConvNetsbased action recognition from depth maps through virtual cameras and pseudocoloring,” in Proceedings of the 23rd ACM International Conference on Multimedia. ACM Press, 2015.
16. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014.
17. D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, “Flexible, high performance convolutional neural networks for image classification,” in IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, no. 1. Barcelona, Spain, 2011, p. 1237.
18. J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng, “Large scale distributed deep networks,” in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, ser. NIPS’12. USA: Curran Associates Inc., 2012, pp. 1223–1231.
19. 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.
20. 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.
21. 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.
22. 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.
23. P. V. V. Kishore, D. A. Kumar, A. S. C. S. Sastry, and E. K. Kumar, “Motionlets matching with adaptive kernels for 3-d Indian sign language recognition,” IEEE Sensors J., vol. 18, no. 8, pp. 3327–3337, Apr. 2018.
24. D. Leightley, B. Li, J. S. McPhee, M. H. Yap, and J. Darby, “Exemplarbased human action recognition with template matching from a stream of motion capture,” in Proc. Comput. Sci., 2014, pp. 12–20.
25. M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, “Ongoing human action recognition with motion capture,” Pattern Recognit., vol. 47, no. 1, pp. 238–247, Jan. 2014.
26. H. Wang and L. Wang, “Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4382–4394, Sep. 2018.
27. S. Zhang et al., “Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks,” IEEE Trans. Multimedia, vol. 20, no. 1, pp. 2330–2343, Sep. 2018.

AUTHORS PROFILE

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

Dr. P. V. V. Kishore is a professor of Image & Video Processing with the department of Electronics and Communications Engineering, where he manages the Image, Speech and Signal processing Research Group. He went on to study M.Tech at Cochin University of science and technology and Ph.D. from Andhra University College of engineering in 2013. He is the chair of the Biomechics and vision computing research center. His works focus on mechine learing, biomechanics, artificial intelligence, human motion analysis and sign language machine translation. His research explores how motion capture data models can effectively model low end video objects in real time for better recogntion and analysis. He is particularly intersted in developing new innovations in the areas of computer vision and mechine learing. He has authored several publications in these fields.

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