# Motion Trajectory for Human Action Recognition Using Fourier Temporal Features of Skeleton Joints

Naresh Kumar and Nagarajan Sukavanam
Department of Mathematics, Indian Institute of Technology Roorkee-247667, India
Email: {atrindma, nsukvfma}@iitr.ac.in

***Abstract*—Spatial and temporal dynamics of human being create a rich set of information to process and analyze very important human activities that can attract the attention of various discipline of real life applications. Finer view of data modality for human body can be characterized by skeleton, contour, silhouette and articulated geometrical shapes. All modalities of a video are be affected by challenging vision problems like view invariance, occlusion and camera calibration at varying scale. In this work, we focused skeleton based human activity recognition and proposed motion trajectory computation scheme using Fourier temporal features from the interpolation of skeleton joints of human body. This is accomplished by considering human motion as trajectory of skeleton joints. Experimental observations ensures that this approach outperforms many of state of the arts. The proposed algorithm is tested on MSRAction3D benchmark dataset. For this we have experimented on three action sets AS1, AS2 and AS3 categorized from the dataset. After different training and testing samples this gives overall accuracy 95.32% for human action recognition.**

***Index Terms*—human action recognition, Histogram of Gradient (HoG), Microsoft Kinect sensor, motion trajectory, human skeleton**

## I. Introduction

In any video analytics problem, the focus of attention depends on the majority vote of perceptual vision. This is measured in terms of action recognition in the video sequences. The process is similar to parse the video sequences for particular labels. One class of videos can be categorized in RGB and depth motion maps. Video sequence pertaining depth information is simply meant that its pixels have normal 2-D information along the depth of image plane. In market, several sensor available that can provide depth as well as skeletal information of human body. From the literature this is observed that depth information are easier to compute human actions with higher scale [1]-[3]. Thus motivates to work with depth information of all the joints of human body to achieve better recognition accuracy. Human action recognition is the toughest phase of video analytics problems. Since, human action is generally varies with multiplicity of other activities that's makes the actions unstructured. On the other hand, features of each activity varies with person to person. Generally, this work is considered to recognize the human actions like brushing teeth, talking phone and drinking water. Older methods for human action recognition have majority of work on 2-D videos using RFID sensor. These methods are highly affected by many vision problems of occlusion and noise which results the accuracy and image quality are degraded. In this case highest accuracy is limited by 80%. Moreover, fitting a costly and complex

RFID tags at every part of human adds many several incompatibility with motion. We focus to develop a model than can recognize the difference between picking, taking and hearing the phone. The structure of such model is motivated by the fact that the activity in the next video frame is determined by the activity happening in the previous frame [2], [4]. In this case, Markov model can be helpful but Hidden Markov Model (HMM) strictly assumes that the activity in the current frame is determined using the information just from the previous frame. This means the information, apart from the previous frame creates the unnecessary computational overhead. This fact can modified that using improved model like Hidden Conditional Random Fields (HCRF) can provide high discrimination to classify closely related human actions.

Thus, the performance of all such model is degraded due to several environmental issues. This demands some efficient and fast approach to deal with real time challenging issues in human activity recognition research. The sensitivity of Markov model gives failure to work with temporal pattern recognition in noisy environment. This causes the generative model to have less accuracy in comparison new models [4], [5] like Fourier Temporal Pyramid (FTP) and Dynamic Temporal Wrapping (DTP).

### A. Visual Data Descriptors

Video data comprise rich set of information. Variety of features engineering can described for human action recognition [2]. Computing 3-D silhouettes gives a bag of 3-D points to represent the structure of human body [6]. By calculating histogram from these silhouettes, a single person activity can be easily determined. These features use RGB domain. This cause to degrade the accuracy due

Manuscript received July 1, 2018; revised December 11, 2018.


doi: 10.18178/joig.6.2.174-180

to vision problems. By fusing others features, the accuracy can be improved. The vital actions of human body cab be determined by head, torso, hands and legs. This motivates to segment the human body into connected sets of joints called skeletal joints [4], [5], [7]. In market, several software like PrimeSense packages are available to track the joint location in human body using Kinect 3D camera. Further, visual data can be represented by spatial coordinate $x_i$, $y_i$, and $z_i$ any particular time t. After small change in time the location of the point is updated. The joint information of space and time is collected as spatiotemporal features using Spatiotemporal Interest Points (STIP). This fact gives an idea to transform the coordinates into x; y; z; t a 4-D points sets. The space along three normal Nx; Ny; and Nz; can be constructed as a cubic grid. The number of pixels inside each grid are counted and computed local occupancy (LOP) features to compute human action sets [1], [3], [8], [9]. Optical flow computations in 3D visual data represents the flow of pixel intensity variation in image sequences. This has wide applicability in video analytics problems like motion detection, video segmentation, and object detection and action recognition. Scene flow can be calculated by transforming the 2-D optical flow into 3-D optical flow using depth z. Let, focal length of the sensor be f, then $X = (x - x_0)\ Z=f$; $Y = (y - y_0)\ Z=f$ where $x_0$; $y_0$ is the principal point of the sensor. Thus 3D scene flow can be calculated subtracting respective 3D vectors in subsequent frames.

### B. Datasets from Kinect 3D Sensor

Kinect RGBD sensor was developed in 2010. The evolution of depth sensors make rich the literature in depth datasets. The data statistics presented in Table I created from MSRAction3D dataset.

TABLE I. THREE CATEGORY OF DATASETS

| AS1 | AS2 | AS3 |
|---|---|---|
| Horizontal arm wave | High arm wave | High throw |
| Hammer | Hand catch | Forward kick |
| Forward punch | Draw x | Side kick |
| High throw | Draw tick | Jogging |
| Hand clap | Draw circle | Tennis swing |
| Bend | Two hand wave | Tennis serve |
| Tennis serve | Forward kick | Golf swing |
| Pickup & throw | Side boxing | Pickup & throw |

### C. Challenges

Recognizing activities for several dynamical system, this sounds the applications as well as challenges due to the complexity in the domain of various activities. Varying the environmental conditions makes harder to recognize the activity exactly. The robustness of action recognition system is termed a highly challenging problem due to temporal changes in activities. Generally long videos are very hard to clip out for annotation. This demands a robust classifier that can easily be trained by large amount unannotated data. Another challenge can be pointed out as a multiple person and parallel activities with varying degree of measure. Variety of the objects category makes the problem more difficult. This leads developing an unsupervised action recognition classifier.

We arranged rest of our work on human action recognition in 5 sections. Section-2 and section-3 represent literature with recent state of the art and proposed methodology of human action recognition system respectively. In section-4 and section-5, experimental results and conclusion with future directions is presented.

## II. RELATED WORK

The most relevant research for human being is human activity recognition. Since 1990s, this research mainly focused to recognize the human activity in 2D RGB videos [10]. For achieving accuracy, both algorithmic and data modality standards are demanded due to high occlusion and noise. The problem was compromised by the evolution low-cost Microsoft Kinect 3D sensor. Which provide 2D video with depth information [11]. Furthermore, the sensor packages like PrimeSense makes easier the job of tracking joints locations of human body in video. In the literature, many of the techniques work only on depth information, while some techniques works on skeleton of human body for activity recognition [12], [13]. The important noise-free and space-time features called depth maps, gives local or global data for the patterns in video. Depth Map features do not have texture data as the case of color images. Higher sensitivity with occlusion makes the whole data noisy due to slight disturbance. These facts conclude the designing of a robust human action recognition system is a quite challenging research problem. Hence, the motivation is to find semi-local with high gradient features to get better efficiency on such datasets

### A. Volumetric Depth Maps

Depth maps based human action recognition methods are proposed in [13]-[16]. They used Lie group features to represent the human body skeleton. The features extracted from human skeleton are incorporated with several layers including rotation mapping and rotation polling in deep network architecture. The most popular algorithms Stochastic Gradient Decent (SGD) is used to train the developed deep network LieNets. HOG features from DMM, are classified on MSR Action3D dataset. Oreifej and Liu combined shape, motion cues to recognize human activity from depth video.

### B. Human Skeletal Features

The central focus for skeletal features is kept at the joints characteristic of the bones. This feature is qualitatively important for temporal modeling of human body. Many of the literature have been found that joints location, angles and Euclidean group as a Lie group structure. The most basic techniques of skeletal data collection are active motion capture (MoCap) and multiple views based color images. Being these sensor not very economical, low cost sensor like Leap motion and Kinect are used to collect skeletal joints features of human body.

### C. Sequential Approaches from Skeletal Data

Our proposed approach is much closed to Vemulapalli *et al.* [17] research in which rolling 3D rotations for several parts of body are considered to compute the action curve in Lie group. Spatial Occupancy (SO) features points are computed from the coordinates of joints skeleton features. The semantics of environmental issues with human activities are described using TUM dataset and Cornell activity datasets [18]. Yang *et al.* [19] proposed an approach by accumulating the activity in whole video from the orthogonal projections of 3D maps, called Depth Motion Map (DMM).

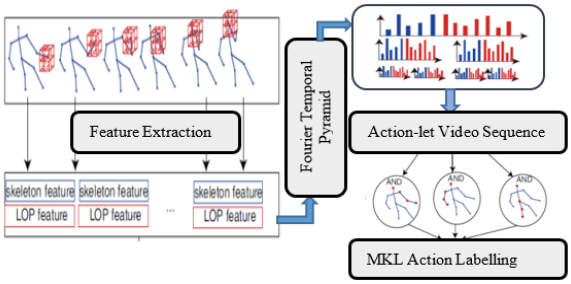


Figure 1. Action-let overview of human body

They created 4D projections of histogram of oriented normal (HON4D). The descriptor gives 88.89% accuracy on MSR Action3D, MSR DailyActivity3D datasets [19], [20]. Campbell and Bobick used mechanics of body movement by tracking Cartesian information and presented phase of actions from spatial curve. Which further used to learn new moments and recognizing the unsegmented actions. Joint location of skeletons is used to represent the action in the video sequence. Generally phase-based methods are view and space variant. Xia *et al.* [21], [22] proposed a novel method to compute feature called Histogram of 3d Joint Locations (HOJ3D). It includes the spatial occupancy pattern relative to the center point of the human body, i.e. torso. Based on the torso of the skeleton a modified spherical space is defined for view invariant features. In this process, the increase dimensionality is maintained by Linear Discriminant Analysis (LDA). The results obtained using MSRAction3D dataset. Zhu *et al.* proposed human tracking based articulated motion. Various classes of sensor are used to capture real-time motion and dynamic orientation is achieved by Kalman fusion-based algorithm. Focusing the problem of noise and occlusion, Wang *et al.* presented a novel sampling scheme of large space to extract semi-local features. These features are termed as Random Occupancy Features (ROP). Finally, SVM model was used to detect the actions in the video. Motivated by several 3D skeleton based schemes and depth sensors for estimating human body, Vemulapalli *et al.* proposed a novel idea by using rolling maps for human action recognition. 3D rotations in human skeleton are represented by a special orthogonal group. The Riemannian manifold is used to represent human action curves in Lie algebra. A temporal classifier Dynamic Time Warping (DTW) is applied on the nominal curve features for each class. Using FTP, they achieved better accuracy by multiclass SVM on MSRAction3D dataset. The architectural overview is presented in Fig. 1.

Xia *et al.* [22] presented a rich set of skeleton capturing techniques which gives spatiotemporal features of human body.

## III. Proposed Model Architecture

The basic motivation of our proposed methodology is taken from the research work in which multiple RNNs are represented as a hierarchical RNN tree to determine hierarchy of action category. Primitive action sketching and localization from the input image sequence a recent idea to action recognition [23]. Fusing features from MBH, HOF and HOG can increase the processing overhead to SVM. We compute the relative joint position from the torso of 3D human skeleton captured Kinect sensor. Motion interpolation from the joints is performed using cubic spline. Next step is remove higher frequency signal from FTP signals. Simple classification is performed by linear SVM for which proposed model architecture flow is given in Fig. 2.

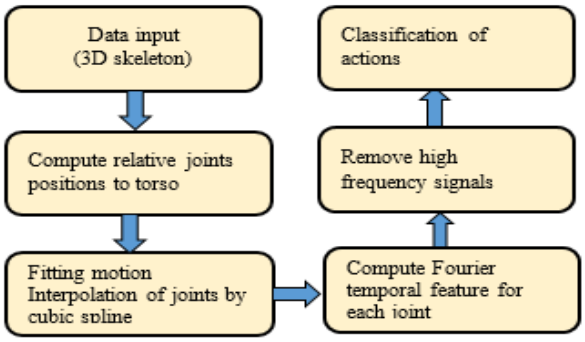


Figure 2. Flow diagram for action recognition system

### A. Cubic Spline Interpolation

Piecewise interpolation using m points represents the cubic spline curve. All these m points are used to control the action curve. By setting second order derivative to zero, we consider a system of tridiagonal system of order (m-2). These equations are referred from (1) to (5).

$$S_i(x) = a_i + b_i(x - x_i) + C_i(x - x_i)^2 + D_i(x - x_i)^3 \quad (1)$$

where (xi; a; b; c; d) is a 5-tuple describing the parameters of Si(x). Given our set of data Y and locations X, we wish to find n polynomials Si(x) for i = 0,. . . ,n - 1 such that

$$S_i(x_i) = y_i = S_{i-1}(x_i), \mathrm{i} = 1,2,\ldots,\mathrm{n}-1 \quad (2)$$

$$S_i(x_i)' = y_i' = S_{i-1}'(x_i), \mathrm{i} = 1,2,\ldots,\mathrm{n}-1 \quad (3)$$

$$S_i(x_i)'' = y_i'' = S_{i-1}''(x_i), \mathrm{i} = 1,2,\ldots,\mathrm{n}-1 \quad (4)$$

$$S_i(x_0)''' = y_0''' = S_{n-1}'''(x_n), \mathrm{i} = 1,2,\ldots,\mathrm{n}-1 \quad (5)$$

### B. Fourier Temporal Pyramid (FTP)

We exploits the temporal relation between joints and compute human action in hieratical manner. For this, we compute Fourier pyramid from the interpolation of the

joints of cubic spline. Fourier coefficients with high frequency, are rejected to maintain the system noise-free. FTP is given high preference than DTW as FTP is less sanative to noise.

## IV. Experiments and Results Discussion

In this phase, the experimental results and analytical comparisons are presented by using standard benchmark of MSRAction3D dataset. The dataset consists of depth maps and ground truth of skeleton joints of human body in various activities. The contents of the dataset is 20 activities like High right arm wave, Horizontal right arm wave etc. as represented by Table II. For the experimental purpose, we created three data samples AS1, AS2, and AS3 from the dataset.

TABLE II. Splitting Set of Actions [17]

| AS1 | AS2 | AS3 |
|---|---|---|
| Horizontal arm wave | High arm wave | High throw |
| Hammer | Hand catch | Forward kick |
| Forward punch | Draw x | Side kick |
| High throw | Draw tick | Jogging |
| Hand clap | Draw circle | Tennis swing |
| Bend | Two hand wave | Tennis serve |
| Tennis serve | Forward kick | Golf swing |
| Pickup & throw | Side boxing | Pickup & throw |

### A. *Test-1*

In this experimental evaluation, we consider 1/3rd of the dataset for training and 2/3rd of the dataset for testing. Test and training samples are generated for experimental work. Under the experiment of Test-1, 1/3rd of datasets is chosen for training samples and rest of data is used for test samples. The diagonal score measure of the confusion matrices generated under test-1 is shown by column 1 of each of Table III, Table IV and Table V. The performance of proposed methods can be observed easily. The comparative results and the outperformance of proposed method can be observed from the corresponding precision, recall and f-measure are presented by Fig. 3, Fig. 4 and Fig. 5 for samples AS-1, AS-2 and AS-3.

TABLE III. Matching Score on Set AS1 under Test1, Test2 and Test3

| AS-1 | Test-1 | Test-2 | Test-3 |
|---|---|---|---|
| Horizontal-Right-Arm-Wave | 100% (17) | 100% (9) | 100% (12) |
| Right-Arm-Hammer | 100% (18) | 100% (9) | 67.87% (8) |
| Right-Fist-Forward-Punch | 88.24% (15) | 88.89% (8) | 81.82% (9) |
| Right-Hand-Height-Throw | 100% (17) | 88.89% (8) | 100% (11) |
| Front-Hand-Clapping | 100% (20) | 100% (10) | 100% (15) |
| Forward-Bend | 88.33% (15) | 100% (9) | 100% (15) |
| Right-Hand-Tennis-Service | 100% (20) | 90.00% (9) | 100% (15) |
| Right-Hand-Pickup-Throw | 76.47% (13) | 100% (9) | 92.31% (12) |

TABLE IV. Comparative Results under Test-1

| Approach | AS1 | AS2 | AS3 | Overall |
|---|---|---|---|---|
| Bag of 3D Points [26] | 89.50% | 89.00% | 96.30% | 91.60% |
| Proposed | **93.75%** | **95.77%** | **99.32%** | **96.28%** |

TABLE V. Matching Score on Set AS2 under Test1, Test2 and Test3

| AS-2 | Test-1 | Test-2 | Test-3 |
|---|---|---|---|
| High-Right-Arm-Wave | 100% (18) | 88.89% (8) | 75.00% (9) |
| Right-Hand-Catch | 93.75% (15) | 87.50% (7) | 75.00% (9) |
| Right-Hand-Draw-Cross | 82.35% (14) | 100% (9) | 83.33% (10) |
| Right-Hand-Draw-Tick | 95.00% (19) | 100% (10) | 73.33% (11) |
| Right-Hand-Draw-Clockwise-Circle | 95.00% (19) | 100% (10) | 86.67% (13) |
| Two-Hand-Up-Wave | 100% (20) | 100% (10) | 100% (15) |
| Right-First-Right-Side-Boxing | 100% (20) | 90.00% (9) | 100% (15) |
| Right-Foot-Forward-Kick | 100% (19) | 100% (10) | 100% (15) |

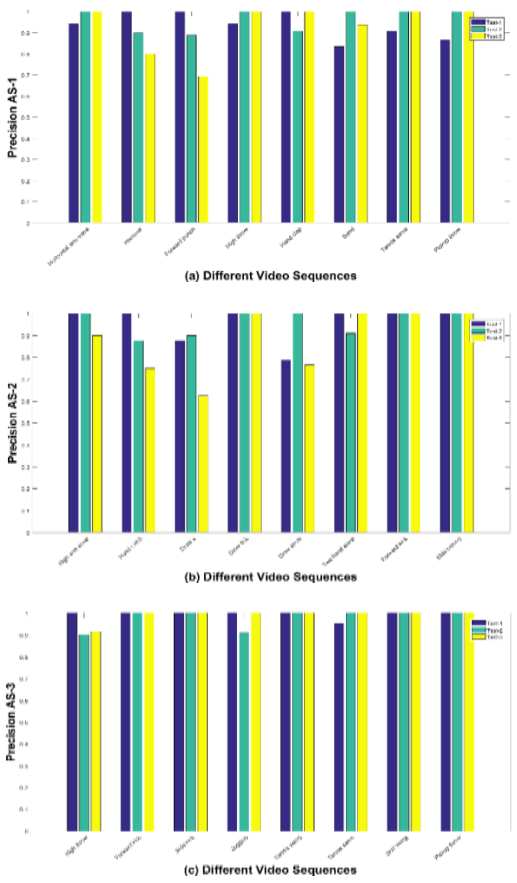


Figure 3. (a), (b) and (c) Precision on action sets AS1, AS2 and AS3

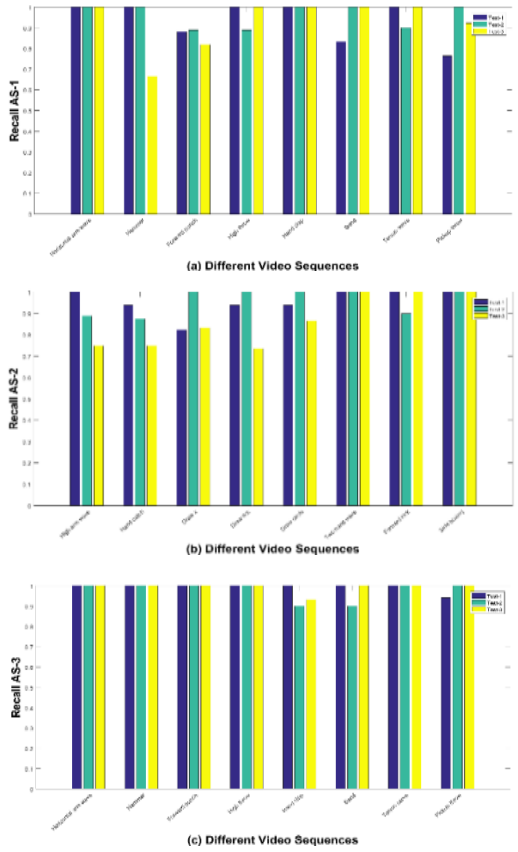


Figure 4. (a), (b) and (c) Recall on action sets AS1, AS2 and AS3

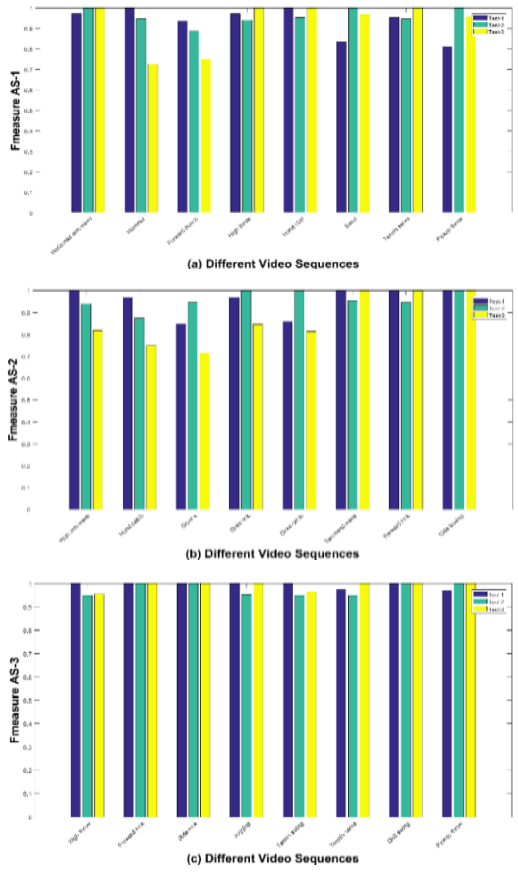


Figure 5. (a), (b) and (c) F-Measure on action sets AS1, AS2 and AS3

### B. *Test-2*

In the experimental phase of Test-2, we select 2/3rd of the dataset presented in Table I for training samples and 1/3rd of the dataset is taken for testing samples. The diagonal values of confusion matrices generated under test-2 are represented by column 2 of each of Table III, Table VI and Table VII. The comparison along the 3-sets is also shown.

TABLE VI. MATCHING SCORE ON SET AS3 UNDER TEST1, TEST2 AND TEST3

| AS-3 | Test-1 | Test-2 | Test-3 |
|---|---|---|---|
| Right-Hand-Right-Throw | 100% (17) | 100% (9) | 100% (11) |
| Right-Foot-Forward-Kick | 100% (19) | 100% (10) | 100% (15) |
| Right-Foot-Side-Kick | 100% (13) | 100% (7) | 100% (12) |
| Jogging | 100% (20) | 100% (10) | 100% (15) |
| Right-Hand-Tennis-Swing | 100% (20) | 90% (9) | 93.33% (14) |
| Right-Hand-Tennis-Serve | 100% (20) | 90% (9) | 100% (15) |
| Golf-Swing | 100% (20) | 100% (10) | 100% (15) |
| Right-Hand-Pickup-Throw | 94.12% (16) | 100% (9) | 100% (13) |

TABLE VII. COMPARATIVE RESULTS TEST-2

| Approach | AS1 | AS2 | AS3 | Overall |
|---|---|---|---|---|
| Bag of 3D Points [26] | 93.40% | 96.90% | 96.30% | 94.20% |
| **Proposed** | **95.95 %** | **96.05 %** | **97.33%** | **96.44%** |

### C. *Cross-Subject Test-3*

In this evaluation of Test-3, we select half of subject as training and rest are taken as testing case. Then all the test cases on sets AS1, AS2 and AS3 are evaluated using SVM classifier shown in Table VIII. Table IX presents over all comparative results of Cross Test (Test-3). In Test-3, half of the subjects are trained and rest are taken as test subjects [24], [25]. The proposed method gives a much improved accuracy of 92.95% as compared with the state of the art.

TABLE VIII. COMPARATIVE RESULTS TEST-3

| Approach | AS1 | AS2 | AS3 | Overall |
|---|---|---|---|---|
| Bag of 3D Points [26] | 72.90% | 71.90% | 79.20% | 74.70% |
| **Proposed** | **93.27** | **87.39** | **99.12** | **93.26%** |

TABLE IX. OVERALL COMPARATIVE RESULTS

| Approach | Overall Accuracy |
|---|---|
| Bag of 3D Points [26] | 74.7% |
| Histograms of 3D joints [27] | 78.97% |
| Random forests [8] | 90.90% |
| **Proposed** | **95.32%** |

## V. CONCLUSION AND FUTURE DIRECTIONS

This research work focuses on human action recognition in videos of common daily life activities, captured with the help Kinect sensor. In the work, the motion of human body is represented as a trajectory of the joints of human body. By applying cubic spline interpolation, we achieved nominal temporal features. The objective of choosing skeleton features helps us getting in better state of the art. Observation shows that our proposed methodology achieved significant improvement as compared with Li *et al.* approach. Experiential values presented in Table IV, Table VII, VIII and Table IX reflect comparative performance of the proposed technique. The values of matching score from the confusion matrices are represented in Table III, Table V and Table VI for all testes along action sets AS1, AS2 and AS3 respectively. Under Test-1, precision, recall and f-measure are shown by Fig. 3(a), Fig. 4(a) and Fig. 5(a) respectively. In the similar way, for Test-2 and Test-3, all these performance measures are shown by corresponding figures. In the future work, our algorithm is not for multiple people and also actions are disjointed. There is a scope of building approach to determine the online actions in which all the actions are chained and multiple people involved. Also, we can use object interaction to get the more accurate results. Our algorithm works on the MSRAction3d dataset, which is very complex dataset but it is a very clean dataset. Processing complex and unconstraint video for multiple activities using deep network libraries is supposed the next objective of this research. [26].

## REFERENCES


[1] B. Ni, G. Wang, and H. M. P. Rgbd, "A colour-depth video database for human daily activity recognition," in *Proc. IEEE International Conference on Computer Vision Workshops*, November 2011, pp. 6-13.

[2] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5344-5352.

[3] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *Proc. ECCV Workshop Consum. Depth Cameras Comput. Vis.*, 2012, pp. 52-61.

[4] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358-386, 2005.

[5] W. Li, L. Wen, M. C. Chang, S. N. Lim, and S. Lyu, "Adaptive RNN tree for large-scale human action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1444-1452.

[6] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 14-19.

[7] L. Xia and J. K. Agrawal, "Spati-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2834-2841.

[8] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2010, pp. 9-14.

[9] S. Vedula, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 475-480, 2005.

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

[11] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen dataset of everyday manipulation activities for motion tracking and action recognition," in *IEEE 12th International Conference on Computer Vision Workshops*, 2009.

[12] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Proc. European Conference on Computer Vision*, 2012, pp. 872-885.

[13] L. W. Campbell and A. F. Bobick, "Recognition of human body motion using phase space constraints," in *Proc. Fifth International Conference on Computer Vision*, June 1995, pp. 624-630.

[14] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *IJRR*, vol. 32, no. 8, 2013.

[15] Y. Guo, D. Tao, W. Liu, and J. Cheng, "Multiview cauchy estimator feature embedding for depth and inertial sensor-based human action recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 617-627, 2017.

[16] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Computer Vision and Image Understanding*, vol. 158, pp. 85-105, 2017.

[17] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 588-595.

[18] Cornell activity datasets CAD-60. [Online]. Available: http://pr.cs.cornell.edu/humanactivities/data.php

[19] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. 20th ACM international conference on Multimedia*, October 2012, pp. 1057-1060.

[20] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716-723.

[21] Z. Huang, C. Wan, T. Probst, and L. V. Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6099-6108.

[22] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 20-27.

[23] R. Zhu and Z. Zhou, "A real-time articulated human motion tracking using tri-axis inertial/magnetic sensors package," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 12, no. 2, pp. 295-302, 2004.

[24] Y. Zheng, H. Yao, X. Sun, S. Zhao, and F. Porikli, "Distinctive action sketch for human action recognition," *Signal Processing*, 2017.

[25] C. Wolf, *et al.*, "Evaluation of video activity localizations integrating quality and quantity measurements," *Computer Vision and Image Understanding*, vol. 127, pp. 14-30, 2014.

[26] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

**Mr. Naresh** is currently a senior PhD student at Mathematics Department in Indian Institute of Technology, Roorkee, India. He received his Master Degree in Computer Science and Applications from I.I.T. Roorkee (U.K.) in 2012.He is in the editorial team of International Journal of Information Engineering and Applied Computing (IEAC). He has been associate reviewer in IEEE Trans in Fuzzy System and The Journal of Experimental and Theoretical Artificial Intelligence including several Springer and IEEE international conferences. He is permanent member of CSI. SCRS, IEANG and IEEE. His main research interests include visual media processing for pattern recognition in security analysis using computer vision and deep learning.

**Nagarajan Sukavanam** is currently a Senior Professor & Head of Department of Mathematics, IIT Roorkee, India. He received his Ph.D. degree from Indian Institute of Science (I.I.Sc.) Banglore and did postdoctoral work from Indian Institute of Technology (I.I.T.) Bombay. He was scientist B at NSTL, DRDO Vizag from 1984 to 1986. He has supervised many PhD students. Currently, the students enrolled under his supervision are working in control theory, robotics and robotics vision problems.