

Fusing Spatio-Temporal Joint Features for Adequate Skeleton Based Action Recognition using Global Alignment Kernel

A.S.C.S. Sastry, S. Geetesh, A. Sandeep, V.S.V.A. Vitru Varenya, P.V.V. Kishore, D.Anil Kumar, E. Kiran Kumar, M. Teja Kiran Kumar

Abstract: Many researchers in computer vision community have been solving challenging problems in human action recognition. Most of the algorithms fail on the limited set of features. The proposed framework for skeleton based action recognition (AR) from the sequences of 3D joint locations. Following the proposed framework, we introduced fusing three different features namely distance, angle and velocity to improve the recognition accuracy. The kernel-based methods are remarkably detecting the RGB and 3D actions. This work explores the potential of the global alignment kernels in skeleton based human action recognition from Microsoft Kinect sensor skeleton data. Accordingly, the distance, angle and velocity features were encoded into global alignment kernels. The recognition is carried out based on the similarity between the query and database features. The framework has been tested on our own 53 class, 5 subject action data named as KLU3D Action, captured using Microsoft Kinect v2 sensor and three other publicly available action datasets NTU RGBD, G3D and UTD MHAD. The performance of our algorithm outperforms when compared to other previous algorithms on the above datasets.

Index Terms: Spatio-Temporal Features, Kinect sensor, Action Recognition, Global Alignment Kernel.

I. INTRODUCTION

For the past two decades, the AR is most challenging

Manuscript published on 30 April 2019.

* Correspondence Author (s)

A.S.C.S. Sastry*, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

S. Geetesh, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

A. Sandeep, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

V.S.V.A. Vitru Varenya, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

P.V.V. Kishore, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

D.Anil Kumar, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

E. Kiran Kumar, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

M. Teja Kiran Kumar, Department of Electronics and Communications Engineering, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

unsolved problem in computer vision applications in home service robot, health care and video surveillance, etc. Over the past several years, there has been a large amount of research was carried out on 2D action recognition and proposed few techniques includes graph based methods, trajectory based methods, 3D volume based methods, kernel based methods, depth maps based deep learning methods, and appearance based methods. Even though the AR from RGB videos is still a not solved problem. Recently, the 3D human action recognition has become popular with the development in 3D sensor technology. The 3D data can be obtained in two ways, one is by using cost effective depth sensors like MS Kinect, and the second is to use marker-based motion capture (Mocap) system.

In this paper, we are focusing on how to increase skeleton based spatio-temporal human action recognition by solving the challenging problems in capturing the human actions in complex viewpoints. The challenges are:

- 1) the location of the sensor with the subject may differ in many causes,
- 2) change in the signer's position/orientation while capturing the data, which may influence the performance of recognition methods.
- 3) variational dimensions of feature representation due to the variation in frame rates of the actions.

To overcome the above problems, we proposed to use Global Alignment Kernel (GAK), which fuse all joint relational features (distances, angles and velocities) for action recognition. In this work, a 3D skeleton coordinate is used to implement our proposed idea. The three features such as joint relative distance features (JRDF), joint relative angle features (JRAF) and joint relative velocity features (JRVF) are used to recognize actions. The proposed algorithm is validated on our KLU3D Action dataset and above datasets [1-3], and a fair comparison with other previous algorithms is also presented.

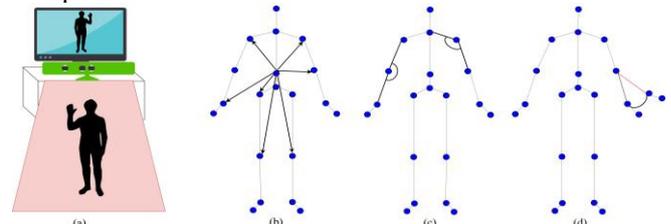


Figure.1. Kinect setup and skeleton joints approaches. (a) Kinect world coordinate system, (b) JRDF, (c) JRAF and (d) JRVF.



The joint relational features are the JRDF, JRAF JRVF, are extracted by using 3D skeleton coordinates. Figure.1 shows the Kinect world coordinate system and joint relational feature representation.

The outline of the paper is presented as follows: related work is in Section II. Detailed methodology is given in Section III. Performance of our algorithm on various 3D skeletal action datasets in Section IV. Section V concludes the findings from our algorithm.

II. RELATED WORK

In recent AR is very popular for AR, the skeleton features give more information to predict actions. Various features based skeleton action recognition works were presented in [4-5]. Wang et al [6] used convolutional neural networks (CNN) to model the spatio-temporal information to joint trajectory maps (JTM) for recognizing actions. Kumar et al [7] introduced a joint angular feature expressing relation between the spatial joints of an action sequence. Motionlet ensemble model [4] is divided the joints into two subsets: motion (MJ) and non-motion joints (NMJ). For example, the action ‘hand waving’ involves right or left-hand joints only. All the body joints were modelled as MJ and NMJ and the joint relative distances and angle features were extracted between the motion and non-motion joints and an adaptive kernel-based classification was initiated for classifying human action.

Form previous work, the recognition accuracy was improved by considering different 3D skeleton features in action recognition [8]. Apart from joint positions [4], joint relative distances [4], joint angles [4], joint velocities [9], joint angular distances [7] and a combination of these features [4], [8] had shown very good accuracy for AR. In [8], the authors proposed the geometric features such as joint surfaces. The surface is formed by 2 consecutive edges on the skeleton. Other geometric features were used by Zhang et al. [5], the authors fused ten different geometric relational features between joints and lines.

In recent for AR are using algorithms like CNNs, LSTMs and RNNs are widely used. Even through their outstanding performance was consistent with huge training sets. Kernel based similarity methods will overcome these problems to some extent. Recently, the kernel-based methods in [4] were performed well and used SVM to classify the kernel. GAKs [10] are being exclusively used for AR. In this work, we propose a new fusion model where checked by using alignment kernel to achieve higher recognition rates compared to other previous algorithms.

Here, the GAKS are used to find the maximum matching between query action and the database actions. This work achieves two major objectives that are currently faced by the skeletal based action recognition algorithms.

1. The process to be change of the subject and the action frame rate.

2. The database has to be transformed into relational features to show the closeness of the query in the database.

Our proposed architecture is giving its best in classifying or recognizing the human actions from the skeletal data. The detailed methodology is given in next session.

III. PROPOSED SYSTEM ARCHITECTURE

The detailed description of our proposed methodology is

given in this section. We start with extracting the joint relative features namely distances, angles, velocities. The global alignment kernels were constructed for each feature type. These GAKs are used to find the similarity between the two action sequences.

A. Feature Extraction

In this work, first we extract the relational geometric features like distances, angles, velocities. In this work, every action is performed by 5 different actors. Microsoft Kinect used to capture human actions with 20 joint full body skeletal model [11].

Joint relative distance features

In 3D space, p_t^j denotes the joint j at a given frame t i.e., $p^j = (x^j, y^j, z^j) \in R^{3 \times J} \forall j=1:J$, where J is the number of joints ($J = 20$). In time series, the joint j at frame t is $p_t^j = \{x_t^j, y_t^j, z_t^j\}_{t=1:T}^{j=1:J}$, where T is the frame number.

The JRDF measures the Euclidian distance between a pair of joints j and $(j+1)$, with a positional vectors p_t^j and p_t^{j+1} in the similar frame t . The $JRDF_t^j$ is given by

$$JRDF_t^D = \|p_t^j - p_t^{j+1}\|_2 \quad \forall j=1 \text{ to } J, t=1 \text{ to } T \quad (1)$$

Where $JRDF_t^D \in \mathbb{R}^{\binom{J(J-1)}{2} \times T}$ is a matrix size of the action.

Joint relative angle features

Let $JRAF$ between joints are calculated by choosing three joints with two projection vectors, where one joint is similar to both the vectors. For three joint marker set $(p^1, p^2, p^3) \subset p^j, \forall j=1:J$ (where $p^j = (x^j, y^j, z^j) \in R^{3 \times J}$ is a 3D space location), which forms two projection vectors $\overline{P^{12}} = d(p^1, p^2) \in R^3$ and $\overline{P^{23}} = d(p^2, p^3) \in R^3$, the angle θ_{p^2} at joint p^2 is given as

$$\theta_{p^2} = \frac{\overline{P^{12}} \cdot \overline{P^{23}}}{\sqrt{\overline{P^{12}} \cdot \overline{P^{12}}} \sqrt{\overline{P^{23}} \cdot \overline{P^{23}}}} \quad (2)$$

The $JRAF_t^A$ matrix size is $\frac{(J-2)(J-1)}{2} \times 1$ for single frame. The total size of the action is $\frac{(J-2)(J-1)}{2} \times T$.

Joint relative velocity features

The joint relative distance, angle features does not extract any temporal features, which is very important for the effective classification of action recognition. Given the 3D position of the joint j at two successive 3D frames p_t^j and p_{t+1}^j , here j is the joint position index and t is time sequence video frame. Joint relative velocity features $JRVF_t^V$ of j^{th} is



$$JRDF_t^V = \|p_t^j - p_{t+1}^j\|_2^2 \quad \forall j=1 \text{ to } J, t=1 \text{ to } T \quad (3)$$

$JRVF_t^V \in R^{J \times (T-1)}$ is action velocity matrix. The joint relative distance, angle features $JRDF_t^D$, $JRAF_t^A$ are in many cases complimentary to the joint relative velocity features $JRVF_t^V$. The joint relative velocity features $JRVF_t^V$ extract the velocity distributions, π_2 of all joints in temporal domain, $JRDF_t^D$ extract the changing information of all joints and $JRAF_t^A$ captures rotational values of all joints in spatial domain.

B. Features Similarity

The global alignment kernel [10] is constructed and used for classification. The GAKs is to find matching between the query and dataset sequences. Figure.2. shows the proposed GAK algorithm.

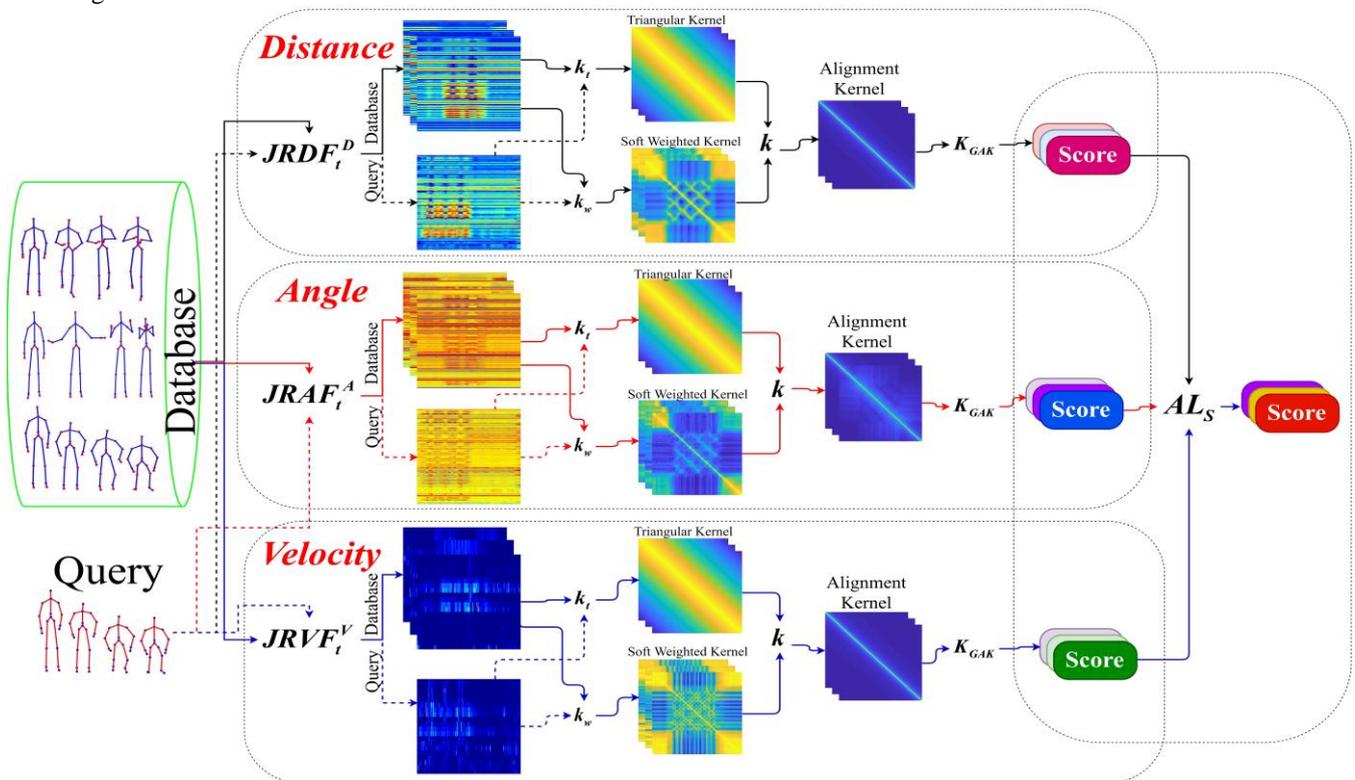


Figure.2. Flow chat of the proposed global alignment kernel (GAK) based action recognition framework.

The first kind of representing our features, we need to define a kernel k which is able to work properly by combining all different features. The Toeplitz kernel is

$$k_t(i, j) = \left(1 - \frac{|i - j|}{k_o}\right)_+ \quad (5)$$

Where k_o is the order of the kernel and $+$ refers to the fact that $k_t(i, j) = 0$ if $|i - j| \geq k_o$. To find similarity between two features f_i^q and f_j^d , we use a Dirac kernel $\delta(f_i^q, f_j^d)$, which is defined as

$$\delta(f_i^q, f_j^d) = \begin{cases} 0 & \text{if } f_i^q \neq f_j^d \\ 1 & \text{if } f_i^q = f_j^d \end{cases} \quad (6)$$

As we can see, the Dirac kernel fails to show how closely the two feature clusters are related. To overcome this

Formally, an alignment between two feature vectors $f^q = [f_1^q, f_2^q, \dots, f_n^q]$ and $f^d = [f_1^d, f_2^d, \dots, f_m^d]$ of length n and m respectively, the GAK K_{GAK} is

$$k_{GAK}(f^q, f^d) = \sum_{\pi \in A(n, m)} \prod_{i=1}^{|\pi|} k(f_{\pi_1(i)}^q, f_{\pi_2(i)}^d) \quad (4)$$

Where (π_1, π_2) is a pair of increased integral vectors of length $l < n + m$, such that $1 = \pi_1(1) \leq \pi_1(2) \leq \dots \leq \pi_1(l) = n$ and $1 = \pi_2(1) \leq \pi_2(2) \leq \dots \leq \pi_2(l) = m$, with no simultaneous repetitions. Let $A(n, m)$ be the set of all possible alignments between the two features of length n and m . It can be shown in that k_{GAK} is a p.d. (positive define) kernel if k and $k/(1+k)$ are positive definitive kernels.

problem, we propose a novel soft weighted version of Dirac kernel of k_w , defined as

$$k_w(f_i^q, f_j^d) = e^{-\phi_\sigma(f_i^q, f_j^d)}, \quad (7)$$

Where:

$$\phi_\sigma(f_i^q, f_j^d) = \frac{1}{2\sigma^2} d(f_i^q, f_j^d) + \log \left(2 - e^{-\frac{d(f_i^q, f_j^d)}{2\sigma^2}} \right) \quad (8)$$

Here $d(f_i^q, f_j^d)$ is distance between two feature vectors. σ is the variance. In our approach the kernel k combines the triangular kernel k_t and soft weighted kernel k_w :

$$k(f_i^q, f_j^d, i, j) = \frac{k_i(i, j) \cdot k_w(f_i^q, f_j^d)}{2 - k_i(i, j) \cdot k_w(f_i^q, f_j^d)} \quad (9)$$

Where f_i^q and f_j^d are two feature vectors, i and j are the position of the features for alignment of the kernel. The kernel defined by equation (4) is finally normalized so as to obtain a similarity value K_{GAK} in the interval $[0,1]$:

$$K_{GAK} = \frac{k_{GAK}(f^q, f^d)}{\sqrt{k_{GAK}(f^q, f^q) * k_{GAK}(f^d, f^d)}} \quad (10)$$

The measurement of K_{GAK} gives the similarity of query action in the dataset. The GAK is applied on distance, angle, and velocity constructing kernels K_{GAK}^{JRDF} , K_{GAK}^{JRAF} and K_{GAK}^{JRVF} .

The final average AR is $AL_s = \frac{K_{GAK}^{JRDF} + K_{GAK}^{JRAF} + K_{GAK}^{JRVF}}{3}$.

IV. DATASETS, EXPERIMENTS AND RESULTS

In this section, we evaluate the effectiveness of the proposed fused JRDF, JRAF and JRVF features against individual features. Four different datasets: our own KLU3D Action dataset and three publicly available benchmark datasets and find performance of our algorithm against other previous algorithms.

A. Datasets

The experiments performed on KLU3D Action dataset. It is captured by Microsoft Kinect sensor which gives 3D coordinate location information. The human skeleton in our dataset has 20 joints from head to toe. However, the KLU3D Action dataset has 53 actions captured by five human subjects, the actions labels are : *answering a phone call, basketball, bowling, bowling throw, boxing, breaking stick, brushing teeth, catching ball, clapping, cleaning utensils, combing hair, coughing, counting money, cricket batting, crouching, double handwaving, drinking tea, drinking water, eating, forward bending, image capture, jogging, kicking soccer ball, kneel down, knocking door, left bending, left hand-waving, left kick, left punch, lifting, marchpast, opening drink bottle, opening water bottle, painting(roller), playing drum, playing flute, playing guitar, playing harmonium, reading, right bending, right kick, right punch, rotating head, rubbing hands, running, salutation, sneezing, stagger walk, tearing a paper, tiptoe, walking, walk-turn-left 90, walking-turn-right 90, wall painting*. Figure.3. shows the skeleton representation of sample actions from KLU3D Action dataset.

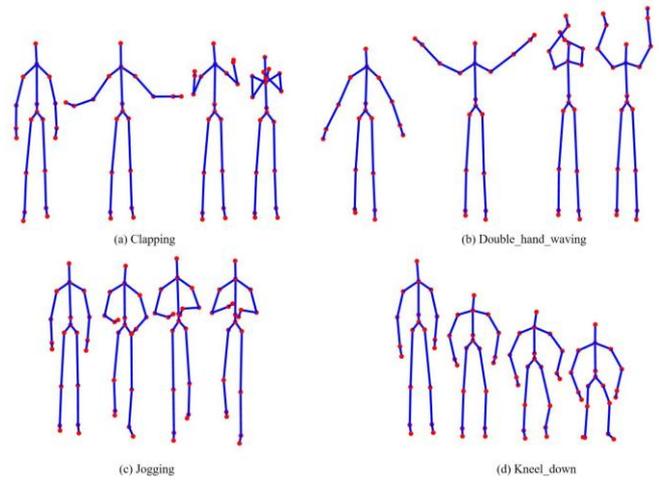


Figure.3. Visualization of skeletal representation of KLU3D Action database captured by Microsoft Kinect sensor.

The other three benchmark datasets [1-3]. The [1] dataset contains 56,000 actions performed 60 different subjects. [2] containing 80,000 frames contains 20 actions performed by 10 subjects and [3] captured 27 actions with containing 861 video sequences captured 8 subjects.

B. Comparison with various popular 3D features

The performance of our proposed feature fusion of distance, angle, and velocity (JRDF, JRAF and JRVF) with GAK is compared with the performance of the other various 3D features like joints, surface, distance, angle, velocity with GAK on four datasets. The recognition rates of the proposed algorithm with different 3D features with above mentioned four datasets as shown in table I. The obtained GAK kernel averaged individual (JRDF, JRAF and JRVF) similarity scores. Furthermore, we quantitatively analyzed the permeance of GAK over the 3D features.

Table I: Recognition rates of various features as input to GAK on different datasets.

Features	NTU		UTD-	KLU3D
	RGB-D	G3D	MHAD	Action
Distance (JRDF)	71.9	89.7	83.7	75.2
Angle (JRAF)	74.1	92.4	87.9	77.3
Velocity (JRVF)	75.9	92.3	85.6	80.7
JRDF+JRAF	79.4	95.7	90.1	84.5
JRAF+JRVF	83.2	94.9	91.7	88.3
JRVF+JRDF	82.1	94.4	90.7	87.3
JRDF+JRAF+JRVF	89.7	97.2	92.5	90.4

Table I shows the AR of individual features with GAK over all datasets. The proposed framework tested on different subjects and different action frame rates.

C. Comparison of previous algorithms with proposed framework

In this session we compared our proposed features on GAK and other previous classifiers. The classifier models were incepted from literature and the average recognition rates for different test subjects testing are shown in table II.

The experiment shows that our proposed method GAK achieved good recognition than the other previous classifiers such as Adaptive Kernel Matching (AKM), Histogram, Dynamic Time Wrapping (DTW), Multiple Kernel Learning (MKL), Weighted Graph Matching (WGM), LSTM, RNN and CNN. For CNN, the features were converted into color coded RGB images for training and testing. Only with few sets of (3 sets) training data, convolutional neural network came close to our method. Similarly, the above-mentioned algorithms were implemented on four datasets considered in this paper.

D. Comparison of previous algorithms with different features

The performance of our proposed GAK classifier on spatio-temporal joint features fused together was tested and

Table II. Recognition rates in classifying various actions of different datasets with state- of-the-art classification algorithms

	NTU RGB-D	G3D	UTD-MHAD	KLU3DSkeleton
Adaptive Kernel Matching (AKM)	82.7	91.2	85.3	84.1
Histogram	78.9	87.3	82.8	79.2
Dynamic Time Wrapping (DTW)	80.3	89.7	82.6	81.7
Multiple Kernel Learning (MKL)	82.1	91.8	86.3	85.6
Weighted Graph Matching (WGM)	86.8	94.9	89.3	88.4
Long Short-Term Memory (LSTM)	85.9	95.2	89.4	87.9
Recurrent Neural Network (RNN)	84.2	93.5	88.7	86.8
Convolutional Neural Network (CNN)	88.4	97.6	91.9	90.1
Proposed Method	89.7	97.2	92.5	90.4

Table III. Comparison of various state-of-the-art methods with different features on different datasets.

Methods	Features	Recognition			
		NTU RGB+D	G3D	UTD MHAD	KLU3D Action
H. Wang [8]	joints+Edges+Surfaces	79.5	--	--	--
Adnan [12]	Spherical Angles	--	92.89		--
C. Li [13]	Joint Distance	76.2	--	88.1	--
Liu J [14]	Joint location	77.4	--	94.7	--
P. Wang [15]	joint trajectory maps	76.32	96.02	87.9	--
Proposed	Distance+Angle+Velocity	89.7	97.2	92.5	90.4

V. CONCLUSION

In this work, we lay out a novel framework for skeleton based action recognition based on a clear model of 3D skeleton joints in a spatio-temporal domain. JRDF, JRAF are spatial features and JRVF are temporal features, which are extracted as distance, angle and velocity on the 3D joint skeleton. These features showed exceptional ability over popular 3D features for action recognition. We applied the proposed fused spatio-temporal joint features to a GAK based classifier, which showed improved accuracies in recognizing actions on our own KLU3D Action dataset and other three publicly available action datasets, NTU RGB+D, G3D and UTD MHAD. The proposed framework performed exceptionally well on all four datasets. For our own KLU3D Action dataset, the recognition rate is 91.5%. The results show an improvement in overall recognition rate using the proposed framework due to the addition of JRVF temporal

features. The proposed model has shown outstanding performance when compared to other previous AR methods.

features. The proposed model has shown outstanding performance when compared to other previous AR methods.

REFERENCES

1. A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, Proceedings of the CVPR, (2016).
2. S. Fothergill, H.M. Mentis, S. Nowozin, P. Kohli, Instructing people for training gestural interactive systems, Proceedings of the ACM Conference on Computer-Human Interaction (ACM HCI), (2012).
3. V. Bloom, D. Makris, V. Argyriou, G3D: A gaming action dataset and real time action recognition evaluation framework, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), (2012), pp. 7–12.
4. 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 Journal, vol. 18, no. 8, pp. 3327–3337, Apr. 2018.



Fusing Spatio-Temporal Joint Features for Adequate Skeleton Based Action Recognition using Global Alignment Kernel

5. S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2330–2343, Sep. 2018.
6. Wang, P., Li, W., Li, C., & Hou, Y. (2018). Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*.
7. Kumar, E. Kiran, P. V. V. Kishore, A. S. C. S. Sastry, M. Teja Kiran Kumar, and D. Anil Kumar. "Training CNNs for 3-D Sign Language Recognition With Color Texture Coded Joint Angular Displacement Maps." *IEEE Signal Processing Letters* 25, no. 5 (2018): 645-649.
8. Wang, Hongsong, and Liang Wang. "Beyond Joints: Learning Representations From Primitive Geometries for Skeleton-Based Action Recognition and Detection." *IEEE Transactions on Image Processing* 27, no. 9 (2018): 4382-4394.
9. Eweiki, Abdalrahman, Muhammed S. Cheema, Christian Baukhage, and Juergen Gall. "Efficient pose-based action recognition." In *Asian Conference on Computer Vision*, pp. 428-443. Springer, Cham, 2014.
10. Kumar, D. Anil, A. S. C. S. Sastry, P. V. V. Kishore, E. Kiran Kumar, and M. Teja Kiran Kumar. "S3DRGF: Spatial 3-D Relational Geometric Features for 3-D Sign Language Representation and Recognition." *IEEE Signal Processing Letters* 26, no. 1 (January 2019): 169–173.
11. V.V. Kishore, P, P Siva Kameswari, K Niharika, M Tanuja, M Bindu, D Anil Kumar, E Kiran Kumar, and M Teja Kiran. "Spatial Joint Features for 3D Human Skeletal Action Recognition System Using Spatial Graph Kernels." *International Journal of Engineering & Technology* 7, no. 1–1 (December 21, 2017): 489.
12. Adnan Salih, Al Alwani, and Chahir Youssef. "Spatiotemporal Representation of 3D Skeleton Joints-Based Action Recognition Using Modified Spherical Harmonics." *Pattern Recognition Letters* 83 (November 2016): 32–41.
13. Li, Chuankun, Yonghong Hou, Pichao Wang, and Wanqing Li. "Joint Distance Maps Based Action Recognition With Convolutional Neural Networks." *IEEE Signal Processing Letters* 24, no. 5 (May 2017): 624–628.
14. Liu, Jian, Naveed Akhtar, and Ajmal Mian. "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition." *arXiv preprint arXiv:1711.05941* (2017).
15. Kumar, Eepuri Kiran, P. V. V. Kishore, Maddala Teja Kiran Kumar, Dande Anil Kumar, and A. S. C. S. Sastry. "Three-Dimensional Sign Language Recognition With Angular Velocity Maps and Connived Feature ResNet." *IEEE Signal Processing Letters* 25, no. 12 (December 2018): 1860–1864.