International Journal of Engineering & Technology, 7 (1.1) (2018) 489-493

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Spatial Joint features for 3D human skeletal action recognition system using spatial graph kernels

P.V.V. Kishore[1]*, P. Siva Kameswari [1], K.Niharika[1], M.Tanuja[1], M.Bindu[1], D. Anil Kumar[1], E. Kiran Kumar[1], M. Teja Kiran[1]

[1]Biomechanics and Vision Computing Research Centre, Department of Electronics and Communication Engineering, KLEF Deemed-to-be-University, Andhra Pradesh, India
**Corresponding author E-mail: pvvkishore@kluniversity.in*

Abstract

Human action recognition is a vibrant area of research with multiple application areas in human machine interface. In this work, we propose a human action recognition based on spatial graph kernels on 3D skeletal data. Spatial joint features are extracted using joint distances between human joint distributions in 3D space. A spatial graph is constructed using 3D points as vertices and the computed joint distances as edges for each action frame in the video sequence. Spatial graph kernels between the training set and testing set are constructed to extract similarity between the two action sets. Two spatial graph kernels are constructed with vertex and edge data represented by joint positions and joint distances. To test the proposed method, we use 4 publicly available 3D skeletal datasets from G3D, MSR Action 3D, UT Kinect and NTU RGB+D. The proposed spatial graph kernels result in better classification accuracies compared to the state of the art models.

Keywords: *Human Action Recognition, Skeleton Maps, Spatial Graph Kernels, Graph Matching.*

1. Introduction

Human action recognition is frequently used in real time applications such as indoor or outdoor security by using surveillance videos to identify abnormal persons and dangerous events, health care activities of day by day (living alone) monitoring system and human-computer interaction. In last few years, many researchers are proposed multiple methods for human motion analysis [1] [2]. In recent trends, low cost depth sensors, like Microsoft Kinect depth sensor is playing vital role in human action recognition from 2D color video, 3D skeleton and RGB-D. These sensors allow action recognition problems to be overcome by using 3D positions of joints (skeleton data). In present study, the skeletal data is used as primary input for action recognition.

In this work, we propose a novel algorithm to recognize human actions with the skeletal joint trajectory information acquired from the Microsoft Kinect depth sensors. The proposed method, extracts spatial joint features from a skeleton data. The extracted features in skeleton database, maps into spatial graph to construct a graph kernel for matching. The proposed method has been tested on four widely available datasets such as G3D [3], MSR Action 3D [4], UT Kinect [5] and NTU RGB+D [6]. Further, the robustness of The algorithm is tested by comparing with other state-of-the-art algorithms. Figure 1 shows the outline of the proposed method.

Graphs are the powerful tools to represent structured 3D data. However, graph construction from the 3D data is a complicated task for problems like human motion retrieval. The difficulty in the task is to compute the similarity between the two graphs constructed from their vertices and edges. This work, focuses on human action recognition form a trained database. Spatial graph kernels (SGK) are constructed for every action in the training and testing set. Each SGK consists of vertex kernel and edge kernel.

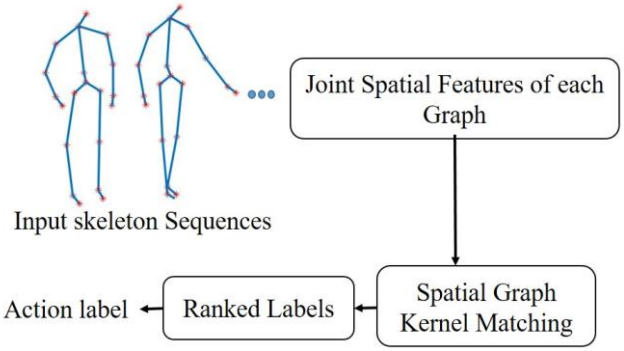


Fig 1. Flow char of the proposed Spatial Graph kernel matching algorithm

The similarity between testing action set and training action set is estimated using spatial graph kernels (SGK). The 3D vertex and edges are attributes modelled with position vector and spatial joint features respectively. The similarity index between SGK's of testing action set and training action set is measured. This similarity index shows the closeness of testing action set with all other actions in the training set.

The rest of the paper is organized as, related work in human action recognition using Kinect sensors is discussed in section 2, the detailed methodology of the proposed model is given in section 3, the section 4 shows the experimental results and finally concluding remarks provides in section 5.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Literature Review

In the last few decades, the human action recognition evolved with the availability of 3D depth sensors like Microsoft Kinect RGB-D sensors [7]. The human silhouette features are easily extracted by using the depth information, which can be concatenated with normalized skeleton features, to improve the motion analysis rate [8]. Kinect sensors captures the depth, which are sometimes combined with RGB data to form an RGB-D video. In recent time these sensors are used to explore human motions [9]. The depth data from Kinect sensor contains, hand trajectories [10], orientations and velocities [11]. Features such as 3D graph joint trajectory locations [12] and joint relative distances [13] are used for human motion analysis.

The features form human actions are classified using support vector machine [14], convolutional Neural Networks [15], Dynamic Time Warping [16], weighted graph matching [17] and Histogram [18]. However, JRD's and RRJRD's based descriptors for human action recognition were successfully used with graph kernel matching in [13] [14]. Inspired from [19], this work uses spatial joint distance features for classifying the human actions.

Human motion recognition in [13] is recognized by representing 3D human joint data as an undirected graphs $\boldsymbol{g}(\boldsymbol{v},\boldsymbol{e})$, with $\boldsymbol{v}$ representing a vertex and $\boldsymbol{e}$ an edge representing path between two consecutive vertices. This model is being explored in this work for human motions on 3D Microsoft Kinect sensor. In [20], human motion in each frame is represented by a graph and matching similarity is calculated between training and test data. Graph based techniques, such as Adaptive Graph Kernels (AGK) in [13], Kuhn – Munkres graph matching algorithm [21] and Dynamic Programming (DP) [22] are used for 3D human motion matching. Graph kernels have received extensive appreciation from researchers on 3D continuous data [23].

These graph kernels compute the similarity between two graphs. However, these methods are not comprehensively meaningful for 3D human motion sequences, since their vertex kernels and edge kernels are not fit for the measurement of similarity between joint relative movements (JRM) upon skeleton data. In this paper, we construct the Spatial-Graph Kernel (SGK) to measure the similarity between 3D human motions represented by the spatial graphs. Our approach shows the superior performance in motion retrieval. This procedure avoids the application of learning algorithm for matching the testing and training action sequences. This works attempts to construct spatial graph kernels (SGKs) on skeleton action sequences for the first time, the method is tested on various publicly available standard databases and performed well on all datasets in recognizing human actions.

3. Proposed Method

The proposed methodology discusses knowledge of the spatial graph kernels systematize on Kinect data. The graph kernels are prepared based on graph constructed from the 3D joint positions of skeleton and distances between joints.

3.1. Joint Spatial Features

Suppose, the human body is represented with P joints, the skeleton with 20 joints is used in this work and is shown in figure 2. The skeleton action sequence $S=\{S_1,S_2,S_3,...,S_T\}$ is performed over T frames. Let $J_i^t=\left[x_i(t),y_i(t),z_i(t)\right]\in\square$ be the ith $(1\le i\le P)$ joint at the tth $(1\le t\le T)$ frame. However, this work uses spatial kernels-based recognition which are built on different joint positions and Euclidian distance between two joints at locations J_i^t and J_1^t. Where i is the joint index.

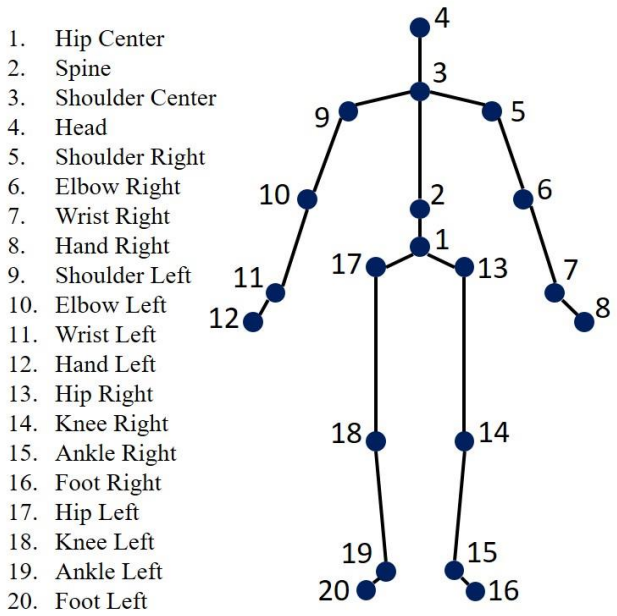


Fig 2. Skeleton template captured by Kinect sensor.

The spatial joint features measure the distance between a pair of ith and 1st joints, over a same frame t. Here, J_1 being the coordinates of the hip joint and J_3 being the coordinates of the shoulder center joint. The ith joint feature J_D^T is the distance between J_i^t and J_1^t, and is normalized to the distance between J_2^t and J_1^t:

$$J_D^T=\frac{J_i^t-J_1^t}{\left\|J_2^t-J_1^t\right\|}\ \forall\ i=1\text{ to }P,\ t=1\text{ to }T \tag{1}$$

Where $J_D^T\in\square^{(P-1)\times T}$ is real matrix with values representing spatial feature distances between P joints over T frames of an action.

3.2. Spatial Graph Construction

Given a skeleton action sequence S, a spatial graph $g(v,e)$ constructed is where v is the vertex set of 3D joint trajectory positions of action sequence S and e is the edge set of joint spatial features in action sequence S. However, in this work a vertex or edge graph matching kernel is being constructed to make the system immune to number of frames. Finally, spatial graph kernel matching between two interlinking graph kernels is performed to identify action label.

3.3. Spatial Graph Kernels

A spatial graph $g(v,e)$, captures the different motions related to an action. To identify action from a dataset, a novel method spatial graph kernels (SGK) is proposed. Spatial graph kernel matching measures the relationship between two graphs in action datasets. The similarity is calculated in graph structure based on vertex and edge matchings.

Let S and S' are the two action sequences, represented as $g(v,e)$ and $g'(v',e')$. Where the vertex and edge kernels in an action dataset are represented as $K_v(v,v')$ and $K_e(e,e')$ respectively. The vertex kernel from a 3D joint trajectory positions of a skeleton sequence S [13] is defined as

$$K_v(v,v') = \exp\left(-\frac{\|v-v'\|_2^2}{2\sigma_1^2}\right) \quad (2)$$

Where v and v' are the vertices of the graphs constructed from the 3D joint trajectory positions of an action sequences S and S' respectively. The gaussian kernel parameter σ_1 is small $(\sigma_1 > 0)$. At the same time, the edge kernel is defined as

$$K_e(e,e') = \exp\left(-\frac{\|e-e'\|_2^2}{2\sigma_2^2}\right) \quad (3)$$

Where e and e' are the edges of the graphs constructed from the joint spatial features J_D^T and $J_{D'}^{T'}$ of an action sequences S and S' respectively. Here, σ_2 is a scale parameter of Gaussian kernel. We introduce one to many matching between the graph kernels $K_v(v,v')$ and $K_e(e,e')$ constructing a matrices of sizes $T_{S_Q} \times T_{S_D}$, where T_{S_Q} and T_{S_D} are the number of frames into the testing set (Query) and training set (Dataset) respectively. The rows of the kernel matrix represent testing set frames and the column represents training set frames. Initiating the cross-value analysis shows the similarity between the testing and training sets. The perfect match gives the highest kernel value.

The action label classification can be maximum kernel value. The decision boundary for the action classifier is set based on the vertex and edge matching scores and are defined as

$$S_v = \frac{1}{T_{S_Q}} \sum_{b \in T_{S_Q}} \arg \max_{r \in T_{S_Q}} (K_v(v,v')) \quad (4)$$

$$S_e = \frac{1}{T_{S_Q}} \sum_{b \in T_{S_Q}} \arg \max_{r \in T_{S_Q}} (K_e(e,e')) \quad (5)$$

Where S_v and S_e are the vertex and edge matching scores respectively for the action dataset. S_v and S_e values are in the range of [0,1]. The value zero represent the nonmatching and one indicates the perfect matching. On behalf of two values, an average of two frameworks is considered as a measure of similarity between the testing and training sets.

4. Experimental Results

The proposed work reports experimental results on human action recognition and tested on different datasets like G3D, MSR Action 3D, UTKinect and NTU RGB+D. The method compares the state of the art methods such as dynamic time wrapping (DTW) [16], weighted graph matching (WGM) [17], adaptive graph kernels [13], histogram [18] and locally preserving positions bag of words(LPP-BOW) [24]. We test the performance of our proposed spatial graph kernels (SGK) algorithm for validating the results with respect to precision-recall and percentage of recognition.

The G3D dataset [3] is a set of skeletons, depth and RGB contains a range of gaming actions captured by Microsoft Kinect sensor. The dataset contains 20 gaming actions performing 10 subjects. i.e., *wave, punch right, steer a car, tennis serve, run, aim and fire gun, kick left, tennis swing forehand, throw bowling ball, walk, defend, golf swing, punch left, tennis swing backhand, flap and clap, jump, climb, crouch, kick right.*

The action performance on the MSR action 3D dataset [4] is of skeleton sequences obtained from Kinect sensor. The dataset contains 20 actions formed by ten subjects, with repetitions of each action into 3 times. The action set is of *side-boxing, hammer, hand catch, tennis swing, draw x, pick up & throw, draw tick, high arm wave, high throw, golf swing, forward punch, two hand wave, side kick, jogging, forward kick, horizontal arm wave, hand clap, bend, draw circle, tennis serve.*

For the UTKinect action dataset [5], every action doing by 10 different subjects (1 female and 9 males) and captured by depth sesors. It contains 10 different actions performed twice: *pick up, pull, throw, sit down, stand up, clap hands, walk, push, wave, and carry*. Each action length contains sample range of 5 to 120 frames. The dataset contains depth map, and RGB image, stable with 20 skeleton joints.

NTU RGB+D dataset [6], is a large-scale action dataset. The dataset contains 56880 action samples containing four different data samples like depth sequences, RGB videos, 3D skeleton and infrared videos. The dataset captured by 3 Microsoft Kinect v.2 camera. The resolution of RGB images are, infrared and depth maps are in, and the 3D skeleton contains 25 major body joints. It contains 60 action sequences formed by 40 subjects, the action categories including daily, mutual and health related actions. Figure 3 shows the sample skeleton database captured from Microsoft Kinect sensor.

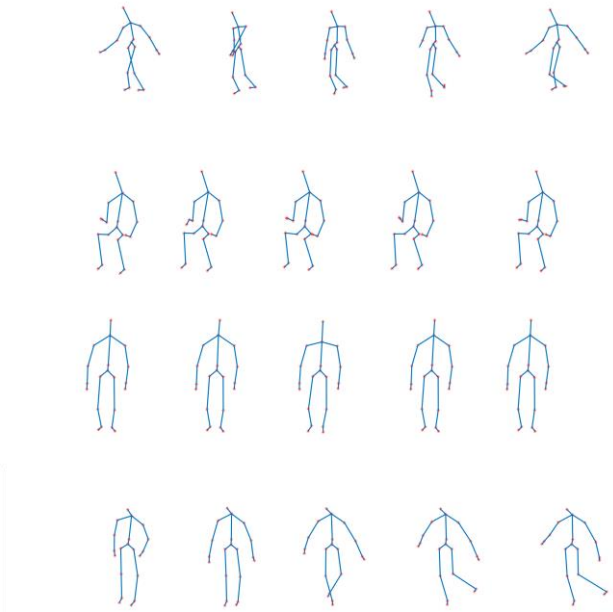
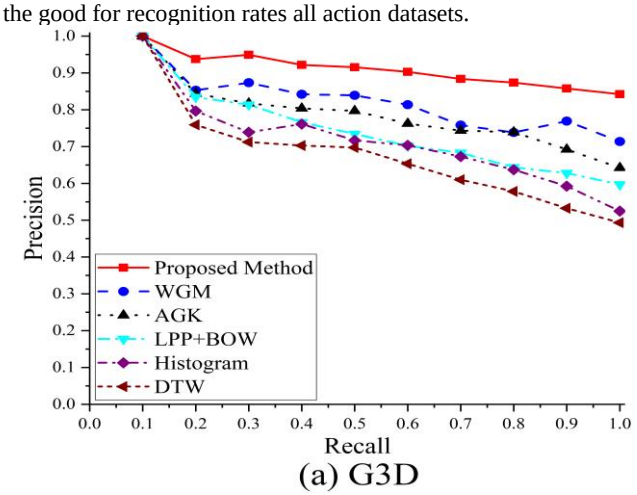
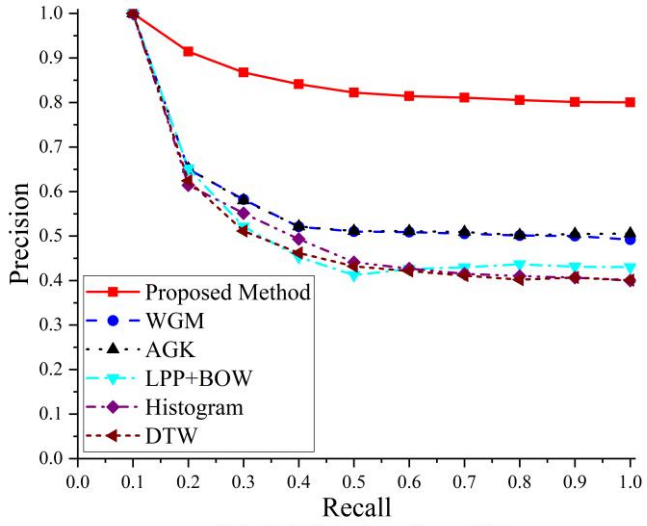


Fig .3. Sample skeleton database captured from Kinect sensor

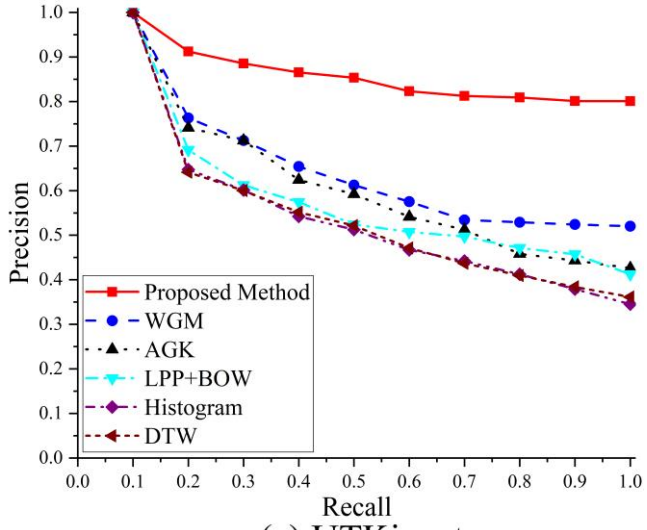
The proposed method measures the closeness of the testing action set with the training action set. Precision and recall values shows the capability of the method in declaring the outcome relativity and truly relativity respectively. All the values are in the range of [0, 1] with a value one indicates the effectiveness of the classifier algorithm. Precision recall curves are plotted for other state-of-the-art methods on the four skeleton action datasets against the proposed method. The fig.4. shows the precision recall curves of proposed SGK method along with the other state of the art methods on four different datasets via G3D, MSR Action 3D, UTKinect and NTU RGB+D. Moreover, the proposed algorithm shows the good for recognition rates all action datasets.



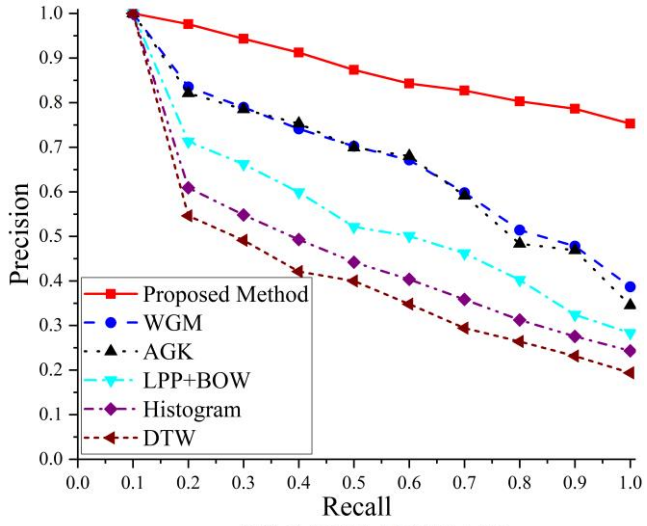
(a) G3D



(b) MSR Action 3D



(c) UTKinect



(d) NTU RGB+D

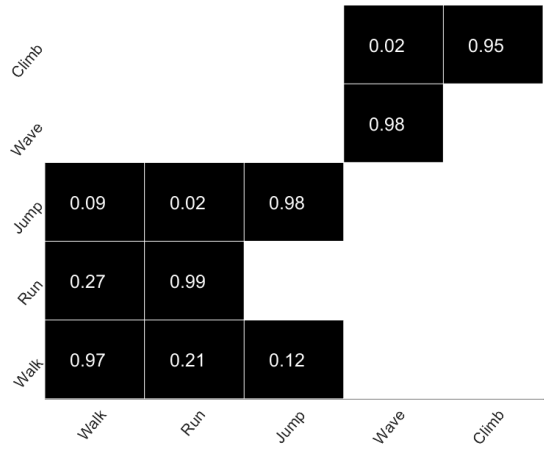
Fig .4. Comparison of different state-of-the-art methods with proposed method on (a) G3D gaming dataset (b) MSR Action 3D dataset (c) UTKinect action dataset (d) NTU RGB+D dataset.

Further, the recognition rates of the proposed method and other state of the art methods are calculated on different datasets and are tabulated in Table-1. Dynamic time wrapping (DTW) achieves good recognition rates, where as it is limited to less number of frames. The recognition rates of weighted graph matching is nearer to our method. But as it is a frame by frame matching, the process takes much more time in recognition. In AGK, only top 50 relative range of joint relative distances (RRJRD's) are used for characterizing the human action. Histogram is a probability-based method, where missing data is unpredictable. In some cases, the action recognition rates were good enough using the LPP and BOW classifies.

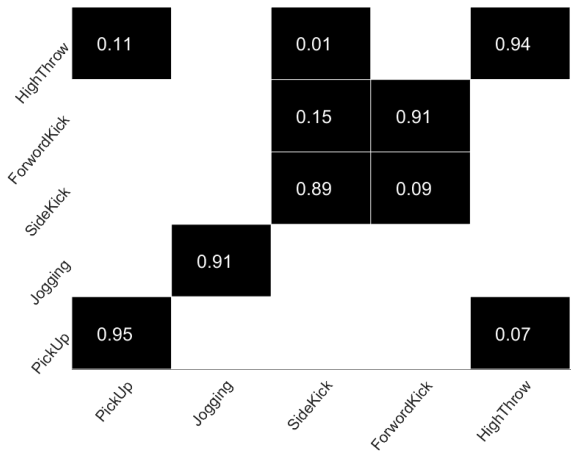
Table-1: List of the averaged recognition with state-of-the-art methods.

Algorithms	Datasets			
	G3D	MSR Action 3D	UT Kinect	NTU RGB+D
dynamic time wrapping [16]	86.3	79.9	72.6	69.8
weighted graph matching [17]	89.2	81.5	84.5	76.3
adaptive graph kernels [13]	84.8	79.5	78.6	73.7
histogram [18]	79.5	75.8	71.5	70.2
LPP+BoW [24]	87.5	81.2	82.4	77.6
Proposed Method	**95.7**	**94.8**	**96.2**	**91.5**

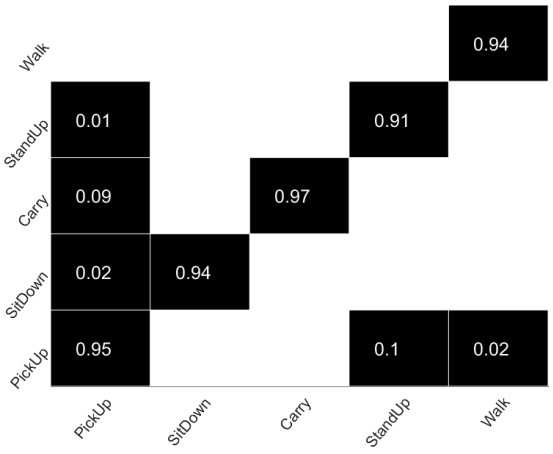
The robustness of the proposed method is tested on four publicly available datasets G3D, MSR Action 3D, UTKinect and NTU RGB+D. The confusion matrix drawn on 4 different datasets with 5 action classes with the proposed method is shown in figure 5. From the fig.5, it can be observed that the action sequences are classified very well with the proposed method. The proposed method exhibits an average of 94.2 recognition rate.



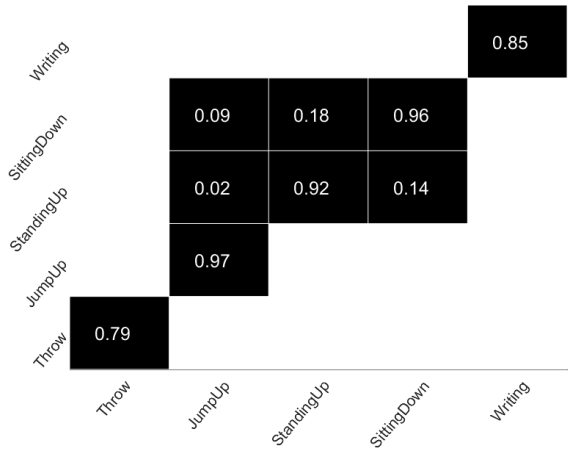
(a) G3D



(b) MSR Action 3D



(c) UTKinect



(d) NTU RGB+D

Fig.5. Confusion matrix of the proposed method on (a) G3D gaming dataset (b) MSR Action 3D dataset (c) UTKinect action dataset (d) NTU RGB+D dataset.

5. Conclusions

In this work, we propose a novel spatial graph kernel matching algorithm for recognizing human action poses from 3D skeletal representations. The skeletal data is represented in the form of a graph with position features and joint distance features as the graphs vertices and edges. The features are used to construct spatial graph kernels which provide a similarity score between the training set and testing set. The proposed method is tested on a large set of 4 publicly available 3D skeletal data. The results show that the proposed model is robust to noisy joints and can detect large number of action sets. The performance of the spatial graph matching model is averaged around 94.2%, which is better than some of the state of the art algorithms used for action recognition.

References

[1] Fujiyoshi, H., Lipton, A. J., & Kanade, T. (2004). Real-time human motion analysis by image skeletonization. IEICE TRANSACTIONS on Information and Systems, 87(1), 113-120.

[2] Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017, February). An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In AAAI (pp. 4263-4270).

[3] Bloom, V., Argyriou, V., & Makris, D. (2016). Hierarchical transfer learning for online recognition of compound actions. Computer Vision and Image Understanding, 144, 62-72.

[4] Li, W., Zhang, Z., & Liu, Z. (2010, June). Action recognition based on a bag of 3d points. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 9-14). IEEE.

[5] Xia, L., Chen, C. C., & Aggarwal, J. K. (2012, June). View invariant human action recognition using histograms of 3d joints. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on (pp. 20-27). IEEE.

[6] Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). 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).

[7] Megavannan, V., Agarwal, B., & Babu, R. V. (2012, July). Human action recognition using depth maps. In Signal Processing and Communications (SPCOM), 2012 International Conference on (pp. 1-5). IEEE.

[8] Papadopoulos, G. T., Axenopoulos, A., & Daras, P. (2014, January). Real-Time Skeleton-Tracking-Based Human Action Recognition Using Kinect Data. In MMM (1) (pp. 473-483).

[9] Patsadu, O., Nukoolkit, C., & Watanapa, B. (2012, May). Human gesture recognition using Kinect camera. In Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on (pp. 28-32). IEEE.

[10] Frati, V., & Prattichizzo, D. (2011, June). Using Kinect for hand tracking and rendering in wearable haptics. In World Haptics Conference (WHC), 2011 IEEE (pp. 317-321). IEEE.

[11] Oikonomidis, I., Kyriazis, N., & Argyros, A. A. (2011, August). Efficient model-based 3D tracking of hand articulations using Kinect. In BmVC (Vol. 1, No. 2, p. 3).

[12] Youssef, C. (2016). Spatiotemporal representation of 3d skeleton joints-based action recognition using modified spherical harmonics. Pattern Recognition Letters, 83, 32-41.

[13] Li, M., Leung, H., Liu, Z., & Zhou, L. (2016). 3D human motion retrieval using graph kernels based on adaptive graph construction. Computers & Graphics, 54, 104-112.

[14] Kishore, P. V. V., Kumar, D. A., Sastry, A. S. C. S., & Kumar, E. K. (2018). Motionlets Matching with Adaptive Kernels for 3D Indian Sign Language Recognition. IEEE Sensors Journal, 1–1. doi:10.1109/jsen.2018.2810449

[15] Kishore, P. V. V., Kumar, K. V. V., Kumar, E. K., Sastry, A. S. C. S., Kiran, M. T., Kumar, D. A., & Prasad, M. V. D. Indian Classical Dance Action Identification and Classification with Convolutional Neural Networks.

[16] Leightley, D., Li, B., McPhee, J. S., Yap, M. H., & Darby, J. (2014, October). Exemplar-based human action recognition with template matching from a stream of motion capture. In International Conference Image Analysis and Recognition (pp. 12-20). Springer, Cham.

[17] Xiao, Q., Wang, Y., & Wang, H. (2015). Motion retrieval using weighted graph matching. Soft Computing, 19(1), 133-144.

[18] Barnachon, M., Bouakaz, S., Boufama, B., & Guillou, E. (2014). Ongoing human action recognition with motion capture. Pattern Recognition, 47(1), 238-247.

[19] Cippitelli, E., Gasparrini, S., Gambi, E., & Spinsante, S. (2016). A human activity recognition system using skeleton data from rgbd sensors. Computational intelligence and neuroscience, 2016, 21.

[20] Kilner, J., Guillemaut, J. Y., & Hilton, A. (2009, September). 3D action matching with key-pose detection. In Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on (pp. 1-8). IEEE.

[21] Ta, A. P., Wolf, C., Lavoue, G., & Baskurt, A. (2010, August). Recognizing and localizing individual activities through graph matching. In Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on (pp. 196-203). IEEE.

[22] Xiao, Q., & Siqi, L. (2017). Motion retrieval based on dynamic bayesian network and canonical time warping. Soft Computing, 21(1), 267-280.

[23] Celiktutan, O., Wolf, C., Sankur, B., & Lombardi, E. (2015). Fast exact hyper-graph matching with dynamic programming for spatio-temporal data. Journal of Mathematical Imaging and Vision, 51(1), 1-21.

[24] Fotiadou, E., & Nikolaidis, N. (2014). Activity-based methods for person recognition in motion capture sequences. Pattern Recognition Letters, 49, 48-54.