International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

Performance Analysis of Human Action Recognition System between Static k-Means and Non-Static k-Means

Tin Zar Wint Cho, May Thu Win

University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar

Abstract: ***In this paper, the human actions are classified in skeleton data from Kinect sensor based on joint distance features. To raise the accuracy rate of postures analysis, the proposed system uses the static k-means algorithm that it takes the static initial K centroids at the first estimates instead of using the non-static (traditional) k-means that it takes the randomized starting centroids at all time. Further, to improve the performance and accuracy, artificial neural network (ANN) is applied to label the classes of the human poses and discrete Hidden Markov Model (HMM) is also used to correctly recognize the human actions based on the sequence of known poses. Experiments with two different datasets (public dataset UTKinect and New dataset) show that the proposed approach produces good performance results and accuracy rate from the action class models.***

Keywords: Skeleton joint data, Static k-means, Artificial Neural network, Hidden Markov model

1. Introduction

Among the various trends of computer vision applications, the problematic issues of recognizing human activities in videos has become popular due to its applications areas such as surveillance systems, healthcare monitoring systems, and human-computer interaction systems. Although the advancements in the availability and acquisition of video data have increased, the improvement of automated human activity recognition is still limited. Moreover, the developments in sensor technology, the invention of depth sensors (Microsoft Kinect), have improved inexpensive images acquisition in the form of three modalities (color, depth, skeleton) which offer a significant perception of the human activities and environment.

First of all, this paper is an extension of work originally presented in [1]. The human action recognition using skeleton joint from the Kinect sensor is reviewed. This sensor provides twenty three dimensional joint positions for each frame in real-time [2]. Using the human skeleton from Kinect output a promising direction as the movements of the human skeleton can determine many actions.

The tracked human joint positions in a real time dance classification system are used [3]. In this system, the Principle Component Analysis (PCA) based on the upper-body joint positions is applied to estimate the torso surface, and the human pose is represented by the angles between the torso surface and the positions of limb joint. To distinguish the temporal structure of actions, it also works Fourier transform over time.

Human actions using eigenjoints are recognized in this system [4]. The joints position difference between joint-pairs within one frame, the joints of one frame and the initial frame, and the joints of two successive frames is computed. PCA is also used to extract “eigenjoints”, which are the main information of human pose. Lastly, to classify the actions on these features, a nearest neighbor-based classifier is used.
The top six informative joints using the Sequence of most informative joints (SMIJ) based on the difference of angle and velocity are found [5], and the feature vectors with these joints are constructed. The histogram of joint position coordinates is extracted and the hip joint is used as the origin [6].

In [7], a human skeleton data is interpreted as a graph and its edge weights based on the distances between all of the joint pairs are computed. Additional edges are also included to connect consecutive skeleton joints. Later, the entire sequence of skeleton is represented as a spatio-temporal graph. A pyramidal representation based on spectral graph wavelet transform (SGWT) [8] is used to collect joints' trajectories at different scales (in time and space). PCA is also applied to make the high dimensionality representation. Classification task is performed by using the standard SVM.

Another attempt to capture the related information among the joints is represented by using the covariance matrix of a skeleton sequence [9]. In this system, a sequence of skeleton joints using a fixed length temporal window is also represented in terms of a covariance matrix that encodes the shape of the joint probability distribution with the set of random variables. To study sequential dependencies of the joint positions, a hierarchical representation is implemented. Action classification is performed by linear SVM.

The key difference between the methodology [10] and other systems using HMM is that in [2] the emission probabilities are replaced by deep neural networks which comprise several layers of features. In [11], to extract relevant features for the actions recognition, only the joints of the left-hand, the right-hand and the pelvis are used. A convolutional neural network classifier is also applied in which an alternating structure of convolution and subsampling layers is used for classification.

2. Proposed System

It is of great challenge to recognize human activity from video sequences due to large variations such as scaling, occlusion, motion style, performance duration, etc. A key issue is that which features is more informative for this task. In this system, the target contribution would be to overcome the above mentioned problems by defining skeletal joints features extraction and the static k-means algorithm which takes the static centroids at the first time and reduces the random centroids at all time to increase the accuracy of postures selection. Furthermore, a supervised neural network is used to define the labels for each posture and the hidden Markov model is applied to recognize the action as correctly as possible. The overview structure of the proposed system is shown in Figure 1.

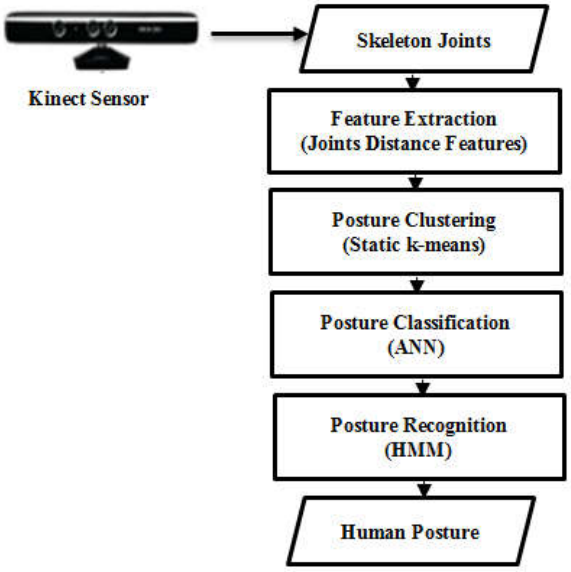

Figure 1: System overview of the proposed system

2.1 Feature Extraction

The invariant distance between the location of the sensor and appearance of the person can be provided by using the skeletal joints data from Kinect sensor [12]. The skeleton joints tracking algorithms from this sensor can exactly extract the joint locations from both front view and back view. A human posture by the positions of twenty skeleton joints data in each frame is represented as follows.

$$HP_t = \{j_t^1, j_t^2, \ldots, j_t^{20}\} \quad (1)$$

where j_t^i is the i^{th} joint location at time t and each joint point consists of 3D joint coordinates x_t^i, y_t^i, and z_t^i.

The differences of joint coordinates are caused by the variant distances between the sensor and the person. This invariant can be removed by subtracting the hip-center joint coordinate from every joint position in each frame as this hip-center point is stable in main postures of human actions. The joints coordinates from this transformation are as follows.

$$j_t^{ki} = j_t^i - j_t^{hipcenter}, \quad 1 \le i \le n \quad (2)$$

where j_t^{ki} is the i^{th} joint point in time t and n is the number of joints.

These joint features are also a set of joints distance feature vectors in which each joint links to the hip-center joint position. A joint distance feature (f) of the action sequence for each frame is defined and a set of distance feature vectors (F) in every frame (m) for an action is computed as follow:

$$f = \{ j_t^{k1}, j_t^{k2}, \ldots, j_t^{kN} \} \quad (3)$$

$$F = \{f_1, f_2, \ldots, f_m\} \quad (4)$$

2.2 Posture Clustering

In this phase, the complication of similar poses is decreased and the simplification of the postures representation is increased without using all of the redundant postures. A popular clustering technique based on the metric with squared Euclidean distance is applied to reduce the similarity of the postures sequence [1]. Example of a duplicated sequence of human posture for the "Walking" action is shown in Figure 2.

In this system, instead of using the non-static (traditional) k-means algorithm that takes randomly the K centroids at the initial estimates in every time, the static k-means technique is used to improve the performance and reduce the random centroids at all times. This approach takes the defined centroids statically at the first step and then the next steps are processed as the non-static (traditional) k-means.

If an action has N joint feature vectors [f_1, f_2, f_3, . . ., f_N], this algorithm processes N feature vectors by grouping together into K clusters for each vector and K vectors [C_1, C_2, C_3, . . . , C_K] in which the centroids of each cluster is presented. Five cluster-identifiers are used in this system.

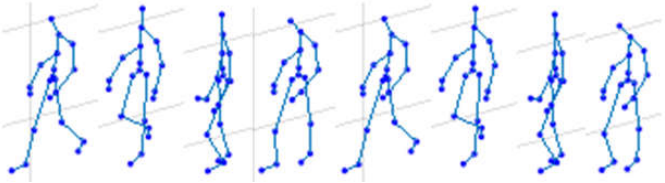

Figure 2: Example of a posture sequence with repetition of the "walking" action

2.3 Posture Classification

After the sequences of similar pose have been eliminated by means of k-means clustering algorithm, the artificial neural network (ANN) is applied which makes the proposed system more intelligent and correctly determines each feature of human pose with the corresponding classes [1].

In this system, 20 3D skeleton joint coordinates in the input layer are included, and there are two hidden layers in which first layer have 40 nodes and second layer have 30 nodes respectively. In the output layer, there are ten output classes. A flow diagram of an ANN applied in this system is shown in Figure 3.

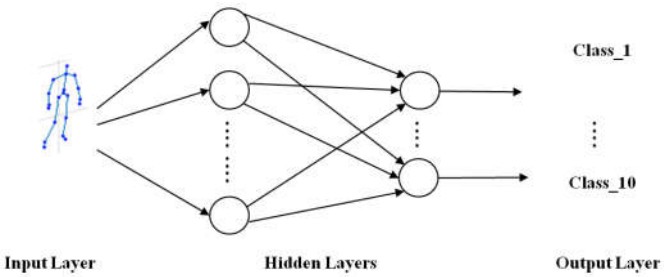

Figure 3: A flow of an ANN in the proposed system

2.4 Posture Recognition

In the recognition phase, the discrete Hidden Markov Models (HMMs) is used to correctly identify the various occurrences of the sequence of human posture after the postures are labeled using the ANN [1]. The unknown state sequences using HMM based on a known observation sequence are established. The given observed posture sequences are extracted from the earlier steps. The three parameters are included in this model.

$$\lambda = (A, B, \pi) \quad (5)$$

In this phase, there are many distinct states at time t, $S = \{S^1_{t1}, S^2_{t2}, \ldots, S^N_{tN}\}$. The prior probabilities, $\pi = \{\pi_i\}$ is:

$$\pi_i = P\,[S^i_{ti}]\,,\ 1 \le i, t \le n \quad (6)$$

where π_i is the probability for the initial state of a state sequence that are assumed as the equal probability distribution and n is the number of different states.

The transition probability of states, $A = \{a_{ij}\}$, from the state S_i to the state S_j, is :

$$a_{ij} = P\,[S_j \mid S_i]\,,\ 1 \le i, j \le n. \quad (7)$$

Assume that the number of different observation symbols per state is R, the set of relevant symbols are $U = \{u_1, u_2, \ldots, u_R\}$, and the probability distribution of the observation symbol in state j, $B = \{b_j(k)\}$ is

$$b_j(k) = P\,[u_k \mid S_j]\,,\ 1 \le j \le n,\ 1 \le k \le R. \quad (8)$$

Figure 4 shows the configuration of the HMM for the proposed system. To train and test different actions, the several classes of joint features are applied to the Hidden Markov Models (HMM) to encode each action as a sequence of the postures and build a discrete HMM respectively.

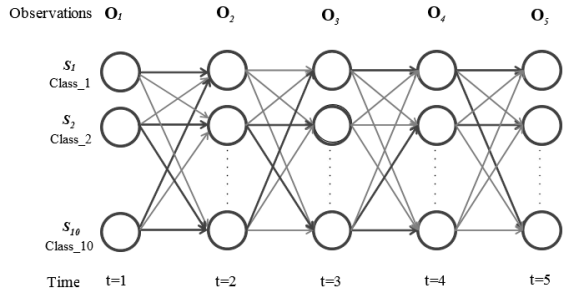

Figure 4: The structure of an HMM used in this system

Once the respective HMM has been trained on the given posture sequences of each human action, a new posture sequence of an action is tested compared to the training set of HMMs and well recognized based on the largest posterior probability from all of these models. The overview process of the human action recognition system in the training and testing phase are shown in Figure 5 and Figure 6 respectively.

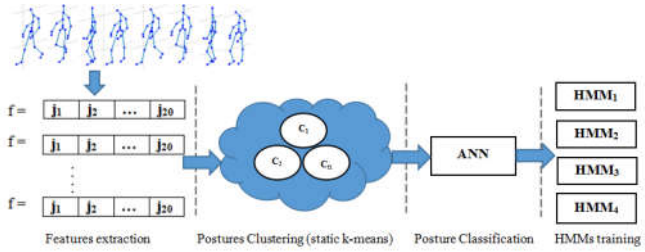

Figure 5: Human Action Recognition Process in Training Phase

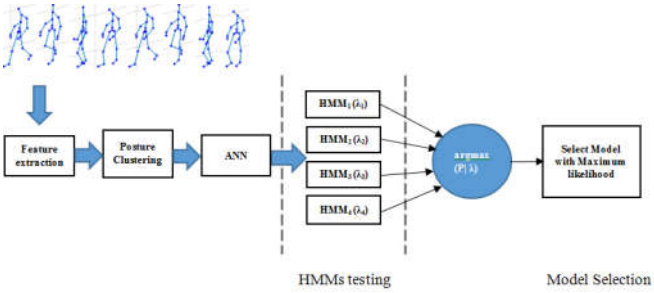

Figure 6: Human Action Recognition Process in Recognition Phase

3. Experiments

In this section, the proposed approach has been evaluated on the public dataset UTKinect-Action3D and a new dataset recorded by the Kinect sensor. From these datasets, skeleton joints distance features are extracted and these features are clustered based on k-means (non-static and static) with the five cluster-identifiers to reduce the similar postures. The class-label of each human pose is classified using artificial neural network, and the corresponding discrete hidden Markov model is constructed to correctly recognize the posture sequences.

3.1 UTKinect Action3D Public Dataset

In this dataset, three channels (skeleton joint positions, color, and depth) are included. There are ten actions (walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, clap hands) by ten subjects with two instances. Among them, four actions (*walk, sit down, stand up, and pick up)* are used *that are* corresponding to the system.

3.2 New Dataset

This dataset consists of the skeleton data captured from the depth sensor (Microsoft Kinect). The Kinect is located at a fixed distance and fixed height from the ground to capture the whole subject body. In this dataset, there are four actions (walking, sitting, standing, and bending) by ten subjects with two instances.

3.3 Experiment on the UTKinect Dataset

The evaluation is performed on the *400* different sequences of posture in which there are *four actions* by *ten subjects* with *two instances* and *five clusters*. In the *training* phase, there are *280* distinct posture sequences that contain *four actions* by *seven subjects* with *two instances and five clusters*. The *testing* set has *120* different sequences which include *four actions* by *three subjects* with *two instances* and *five clusters*.

Initially, the experimentation is performed on the training set using non-static k-means in which all of the centroids are not the same at all time. When the training data is trained on one of the random centroids, the testing results are different at the first time and the next time.

The evaluation using this method is performed on the training set and this process is repeated 10 times. The experimental results are shown in Figure 7. In this experiment, the average correctness rate of the two actions ("walking" and "bending") are reasonable, however, the average accuracy rate of "sitting" and "standing" actions are reduced.

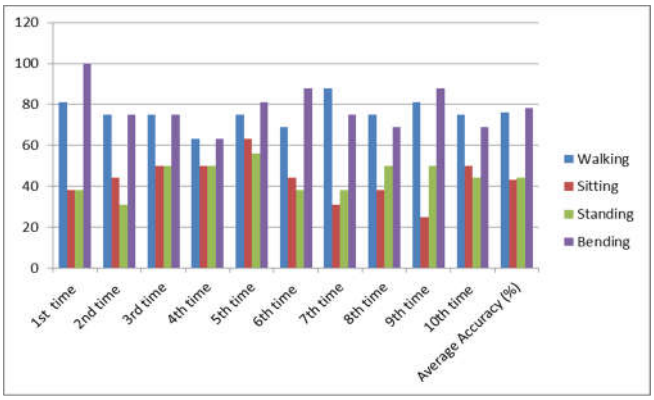

Figure 7: Recognition rate on the training set of UTKinect dataset using non-static k-means

Hence, the proposed static k-means method is applied which takes statically the defined centroids at the first time to improve the performance and accuracy rate. The experiment using this proposed technique is performed on the training set and it is also compared with the non-static k-means process.

The experimental results show that the overall accuracy rate of all actions are much better than the previous method, especially, the accuracy rate of the "sitting" and "standing" actions are obviously great. The evaluation on the training set with the non-static and static k-means algorithm is also shown in Figure 8.

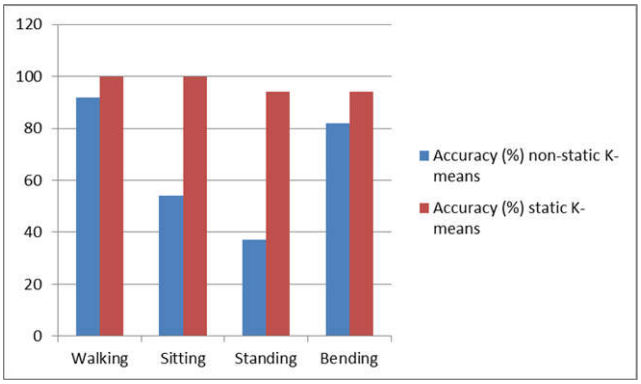

Figure 8: Comparison using the non-static and static k-means on the training set of the UTKinect dataset

Next, the experiment using non-static k-means is also evaluated on the testing set and then, it is repeated 10 times. In this evaluation, the average accuracy rate of each action is not high, specially, the accuracy rate of the "sitting" action is worst. The experimental results are shown in Figure 9.

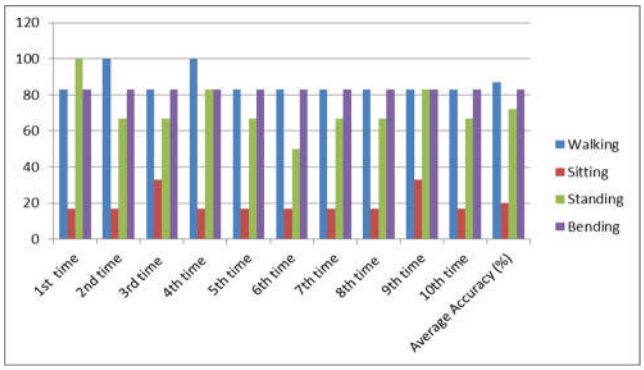

Figure 9: Recognition rate on the testing set of UTKinect dataset using non-static k-means

Then, the experiment is performed on the testing set by using the proposed method (static k-means). From this experiment, all of the accuracy rates for each action are significantly high. The comparison with the non-static and static k-means technique on the testing set is shown in Figure 10. The confusion matrices on the training set and testing set with the static k-means are also shown in Table 1 and 2.

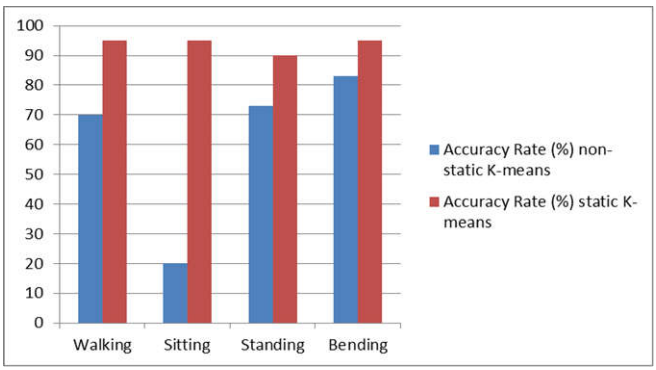

Figure 10: Comparison using the non-static and static k-means on the testing set of the UTKinect dataset

Table 1: Confusion Matrix on the Training set of UTKinect dataset with Static k-means

Type of Action	*Accuracy Rate (%)*			
	Walking	*Sitting*	*Standing*	*Bending*
Walking	100	-	-	-
Sitting	-	100	-	-
Standing		6	94	-
Bending	-	-	6	94

Table 2: Confusion Matrix on the Testing set of UTKinect dataset with Static k-means

Type of Action	*Accuracy Rate (%)*			
	Walking	*Sitting*	*Standing*	*Bending*
Walking	95	-	-	5
Sitting	5	95	-	-
Standing	-	10	90	-
Bending	-	-	5	95

3.4 Experiment on the New Dataset

The experiment is also evaluated on the new dataset in which 400 different posture sequences (*four actions, ten subjects, two instances and five clusters*) are used. There are *280* distinct posture sequences (*four actions, seven subjects, two instances and five clusters*) in the training phase. In the *testing* phase, there are *120* different sequences (*four actions, three subjects, two instances and five clusters*).

First of all, the evaluation using the non-static k-means is performed on the training set. The experimental results are shown in Figure 11. In this experimentation, the average accuracy rate of "walking" and "bending" action are high and the "sitting" and "standing" accuracy rate are low.

The experimentation using the proposed method (static k-means) is also tested. The comparison between these two methods is shown in Figure 12. From this experiment, the results are shown that the overall accuracy rate of all actions of the proposed method is significantly high.

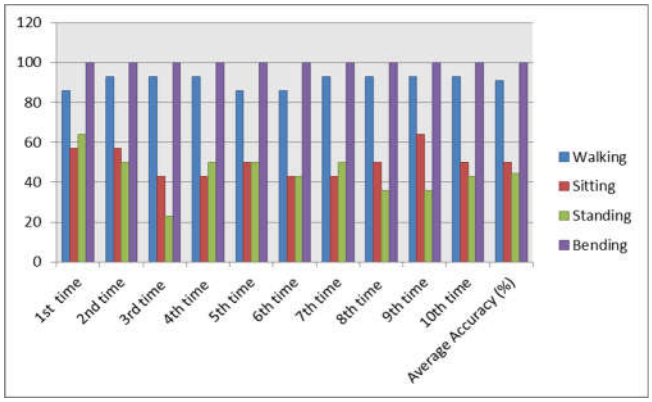

Figure 11: Recognition rate on the training set of New dataset using non-static k-means

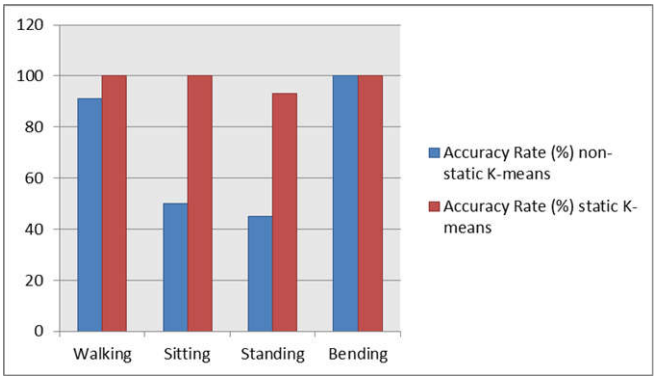

Figure 12: Comparison with the non-static and static k-means on the training set of the New dataset

Then, the experimentation with non-static k-means is performed on the testing set. From this evaluation, the accuracy rate of the "standing" action is worst and the accuracy rate of all actions is shown in Figure 13. The evaluation with static k-means is also performed on the testing set. Figure 14 compares the results between the non-static and static k-means. This experimental result showed that the "standing" accuracy rate using static k-means is higher than the non-static k-means. The confusion matrices with the static k-means on this dataset on the training and testing set are also shown in Table 3 and 4 respectively.

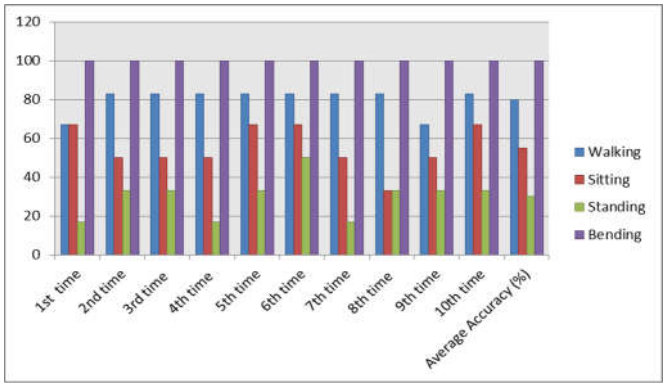

Figure 13: Recognition rate on the testing set of New dataset using non-static k-means

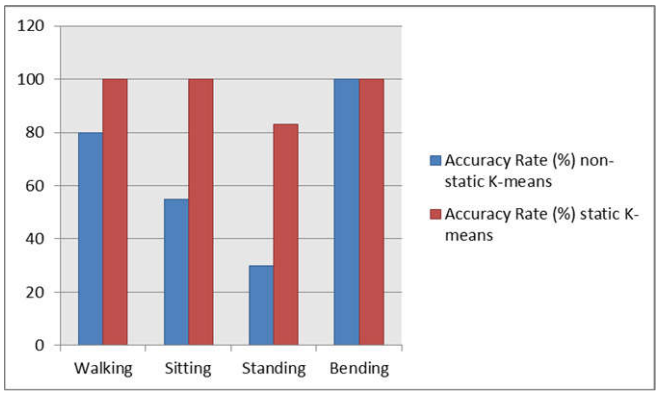

Figure 14: Comparison with the non-static and static k-means on the testing set of the New dataset

Table 3: Confusion Matrix on the Training set of New dataset with Static k-means

Type of Action	*Accuracy Rate (%)*			
	Walking	*Sitting*	*Standing*	*Bending*
Walking	100	-	-	-
Sitting	-	100	-	-
Standing	-	-	93	7
Bending	-	-	-	100

Table 4: Confusion Matrix on the Testing set of New dataset with Static k-means

Type of Action	*Accuracy Rate (%)*			
	Walking	*Sitting*	*Standing*	*Bending*
Walking	100	-	-	-
Sitting	-	100	-	-
Standing	-	-	83	17
Bending	-	-	-	100

Through the evaluation, the experimental results on both the training and the testing set show that the static k-means correctly recognizes a sequence of human action better than the traditional k-means and the performance of the proposed method is significantly improved more than the non-static k-means.

4. Conclusion

The human action recognition system in skeleton joints data is proposed. The joints distance features is extracted and these features are grouped together to reduce the similar human postures using the static k-means algorithm that statically takes only the initial centroids at the first time. Then, the artificial neural network is applied to classify the class-labels of each posture. Furthermore, a sequence of human action is correctly recognized by using the discrete hidden Markov Model (HMM). The evaluation is effectively performed on the public dataset UTKinect and a new dataset by means of these techniques. From the experimental results, the overall average accuracy and performance of the proposed method are improved more than the non-static k-means method.

References

[1] Tin Zar Wint Cho, May Thu Win, Aung Win. "Human Action Recognition System based on Skeleton Data", IEEE International Conference on Agents (ICA), 2018
[2] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. CVPR, 2011.
[3] M. Raptis, D. Kirovski, and H. Hoppe. Real-time classification of dance gestures from skeleton animation. In Proceedings of the SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '11, page 147, New York, USA, 2011. ACM Press.
[4] X. Yang and Y. Tian. EigenJoints-based Action Recognition Using Na¨ıve-Bayes-Nearest-Neighbor. In CVPR HAU3D Workshop, 2012.
[5] F. Oi, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition, Journal of Visual Communication and Image Representation 25 (1) (2014).
[6] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference on, pages 20–27, 2012.
[7] T. Kerola, N. Inoue, K. Shinoda, Spectral graph skeletons for 3D action recognition, in: Proc. of Asian Conference on Computer Vision (ACCV), Springer, 2014
[8] D. K. Hammond, P. Vandergheynst, R. Gribonval, Wavelets on graphs via spectral graph theory, Applied and Computational Harmonic Analysis
[9] M. E. Hussein, M. Torki, M. A. Gowayyed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in: Proc. of International Joint Conference on Arti_cial Intelligence (IJCAI), AAAI Press, 2013
[10] X. Yang, Y. Tian, Eigenjoints-based action recognition using Naive-Bayes-Nearest-Neighbor, in: Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2012.
[11] E. P. Ijjina, C. K. Mohan, Human action recognition based on MOCAP information using convolution neural networks,
[12] Bangli Liu, Hui Yu, Xiaolong Zhou, and Honghai Liu. "Combining 3D Joints Moving Trend and Geometry Property for Human Action Recognition". In Proc. IEEE Systems, Man, and Cybernetics (SMC), 2016.