Investigation of the Unsupervised Machine Learning Techniques for Human Activity Discovery

Md Amran Hossen[1], Ong Wee Hong[2], and Wahyu Caesarendra[3, *]

[1,2] Faculty of Science, Universiti Brunei Darussalam, Brunei Darussalam
[3] Faculty of Integrated Technologies, Universiti Brunei Darussalam, Brunei Darussalam
wahyu.caesarendra@ubd.edu.bn

Abstract. Human activity recognition has been considered as the main capability of an intelligent system in understanding of human activities. Human activity recognition focuses on classifying activities with predefined models learned from labelled data based on supervised or semi-supervised approaches. These approaches have assumed the availability of abundant labelled activity observations. In real-world scenarios, labelled activity observations are difficult to obtain given the undefined number of human activities and their wide variation between different subjects. The desirable approach is an un-supervised one in which an intelligent system can discover new activities from unlabeled observations. This work aimed to evaluate the performance of several clustering algorithms to effectively distinguish different daily activities for human activity discovery. Clustering algorithms used include k-means, spectral, hierarchical and BIRCH clustering. Activity observations were represented as a sequence of postures with 3D skeletal joint locations derived from the Kinect depth map, and then different clustering algorithms were applied to the data. The approach is evaluated on a lab recorded dataset and a publicly available dataset. Overall mean precision, recall and F1-score for both datasets were above 58%, 68%, 61% respectively. K-means and agglomerative clustering with ward linkage achieved highest precision, recall and f1-score on both datasets which demonstrated the potential of using clustering algorithms to distinguish and group different activities for activity discovery without using labeled data.

Keywords: Human activity discovery, Clustering human activities, Machine learning, Unsupervised learning.

1 Introduction

Human activity analysis has been an important area of computer vision research for the past few decades. It has a wide range of applications in various domains, including human-robot interaction, video surveillance, gesture recognition, home behavior analysis and healthcare monitoring. Most of the research works in human activity analysis have focused on human activity recognition (HAR). The HAR systems mainly use supervised or semi-supervised approaches to recognize a limited set of activities they have been trained on. It's a challenging task to train models for each and every activity as

human activity can be numerous and be carried out with wide variation. Another challenge is labelling all learning data, which is required for supervised learning. With supervised learning approaches, it is assumed that there are abundant la-belled observations of activities. However, in real life due to the wide variety of human activities, it is impossible to acquire labelled sample of all possible daily activities. A much less studied aspect in human activity analysis is the human activity discovery. Eunju et al. [1] has pointed out that human activity analysis comprises of different aspects including human activity discovery and human activity recognition. Human activity discovery is the ability of an intelligent system to find new activities from a pool of unlabeled and unknown activities. From the pool of unknown activities, human activity discovery groups the same or similar activities using unsupervised learning, and each group of a new activity can then be used for human activity recognition with the help of active learning (ask human). In other words, human activity discovery requires the ability to autonomously differentiate or distinguish between different activities without knowing their labels.

Survey papers on HAR have shown that the majority of the approaches on HAR are focused on supervised, semi-supervised and hybrid learning approaches [2]–[5]. Many research works have applied unsupervised learning in HAR. Clustering algorithms have been used to determine crucial postures for constructing activity feature vectors to recognize human activities with high precision [6][7]. Unsupervised learning algorithms have been used in various contexts for mapping temporal motion dynamics from a fixed or varying length of skeleton sequences [8][9]. These approaches have used unsupervised learning algorithms to extract features to improve the accuracy of their HAR systems, which were eventually trained using labelled data.

A few researchers have proposed the use of unsupervised algorithms for human activity discovery. They have predominantly used data from accelerometer [10]–[13] and smart home systems [14][15]. Their approaches either require attaching sensors on different parts of the human limbs or attaching sensors to the household objects such as chairs, cupboards and doors. The requirement to attach sensors on human body is cumbersome, uncomfortable, and not desirable in human's everyday life. The use of objects to indirectly label the activities, for example going out of home when the door sensor has detected the movement of the doors, has limited the use of such approach to identifying highly distinguishable activities and demanding large number of environment sensors to be installed.

Consequently, very few works have studied unsupervised human activity discovery based on visual data, particular the skeleton data from a single depth sensor. To the authors' knowledge, the most related research on human activity discovery based on skeleton data is done by Ong et al. [16]. Incremental k-means clustering was used by Ong et al., for the discovery of human activity using human range of movement features which were extracted from skeleton data of the postures in consecutive frames of each activity sample. Though they were able to distinguish several unknown activities using their technique, they have only used k-means clustering in their method. K-means clustering has assumed that the data distribution is spherical. The nature of the distribution of human activities data is unknown. Investigation with other clustering algorithms to

solve the problem is desirable. In this paper, we investigated the effectiveness of four widely used clustering algorithms in human activity discovery.

In this work, different unsupervised machine learning techniques were evaluated to group different activities within a pool of unlabeled and unknown activities. For the investigation, methods preferred were based on visual data that do not require people to wear sensors or devices to capture motion data, as in daily living environment it is unlikely for a person to wear sensors. Specifically, we obtained human skeleton data from a depth vision sensor as the features for the unsupervised learning algorithms. It's worth pointing out that recording skeleton data from depth sensor, i.e. without the color images, has the advantage of preserving a good degree of privacy for the person being recorded by the sensor. The main contributions of this study are that it investigates the application of different clustering algorithms in human activity discovery based on visual data without labels, and highlights areas for further research.

The paper is structured as follows. Section 2 explains how human activity discovery was portrayed as a clustering problem. Section 3 describes the experiments carried out for the investigation. The findings of this study are summarized in section 4. Brief discussion is presented in section 5 and finally, this work is concluded in section 5.

2 Application of clustering for human activity discovery

In this section we describe the formulation of human activity discovery as a clustering problem. Assume the data points in Fig. 1(a) represent different samples or instances of human activities. There are three different activities, for examples standing, walking and waving right hand, represented with different symbols in the figure. Each activity instance comprises of a set of features extracted from depth images, which will be described in detail in the following paragraph. A clustering algorithm is applied on all the activity instances without knowing their label. The objective of the clustering is to distinguish and separate the three different activities into three coherent groups as shown in Fig. 1(b).

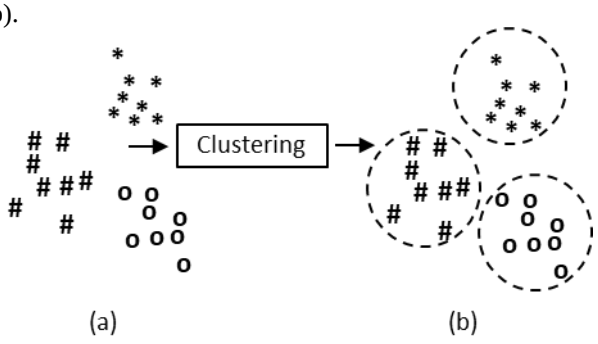

Fig. 1. Application of clustering for human activity discovery (a) unlabeled activity sequences (b) similar activities grouped into respective clusters

These groups of activities discovered by the clustering algorithm can then be used for human activity recognition by using active learning and state of the art HAR techniques. Depending on the clustering algorithm, the number of clusters may be specified by human, or determined by the clustering algorithm. In this study, we have defined an activity instance as a fixed duration of an activity. This assumption has been widely used in HAR research. Based on HAR literatures, an activity can be recognized from an observation of within a few seconds duration. For activity that may last longer, such as walking, an observation of 1 to 3 seconds is sufficient to recognize the activity. If the sensor data is 30 fps, that will translate to an instance of 30 to 90 frames. Each frame is a single posture of the activity represented as a skeleton comprising of various body joints. The number of joints is depending on the algorithm that has been used to extra skeleton data from the depth images. Each joint is represented by its 3D position in XYZ coordinates. An activity instance is, therefore, represented by a feature vector comprising of f frames × j joints × 3 coordinates features. A geometric transformation Fig. 2 was used to translate the skeleton in each frame, with hip center joint being translated to the center of the coordinate frame (0,0,0) to make the activity instances view invariant. Fig 2(b) and Fig 2(d) are the translation of Fig 2(a) and Fig 2(c) respectively which shows two different frames of walking that were recorded at different locations in the sensor's field of view.

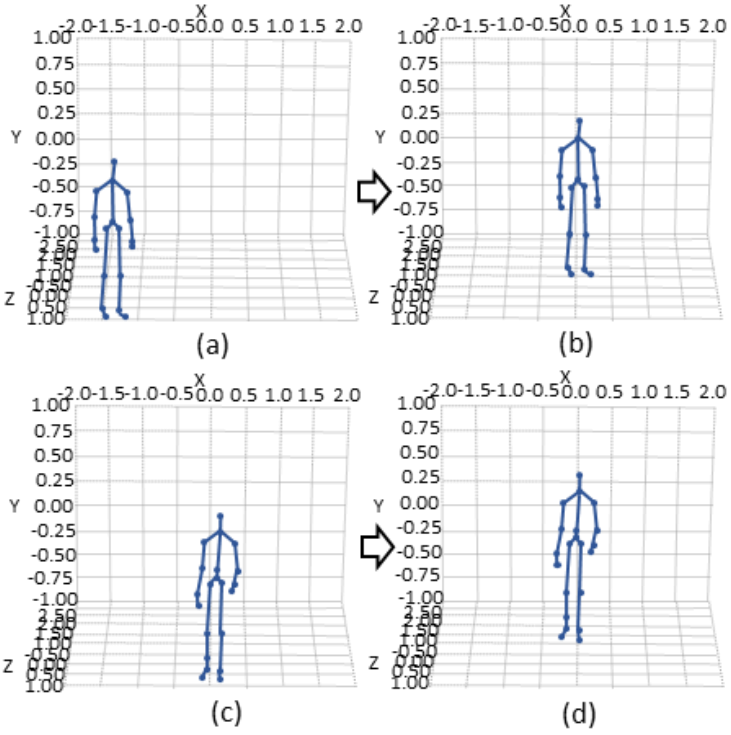

Fig. 2. (a) and (c) are frames of same activity performed at different positions in the sensor's field of view. (b) and (d) are the translation of (a) and (c)

2.1 Clustering

In this sub section, we briefly describe the four clustering algorithms we have investigated in this work. In general clustering is the process of grouping comparable objects from a given set of objects based on certain measure of resemblance, so that intra-class similarity is enhanced, and inter-class similarity is reduced.

K-means. Clustering looks for a finite number (*K*) of clusters in a dataset by categorizing *n* data points in *d* dimensions into *K* clusters while maintaining the cost function low in Eq. 1

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \left\| X_i^{(j)} - C_k \right\|^2 \quad (1)$$

J is the objective function, number of clusters is denoted with *K*, *n* is the number of objects in the dataset, $\left\| X_i^{(j)} - C_k \right\|^2$ is a chosen distance metric between a data point $X_i^{(j)}$ in cluster *k* and the centroid c_k of cluster *k*.

Spectral. Clustering is a technique for identifying groups of nodes in a graph by looking at the edges that connect them. Given data points $X = X_1, X_2, \ldots\ldots\ldots X_n$, for each pair of data points $i, j \in X$, a similarity (weight) $S_{ij} = S_{ji} \geq 0$ is assigned. In Spectral clustering, a graph $G = (V, E, W)$ is composed with *V* containing the vertices (data points), *E* containing the edges and *W* containing the edge weights. *W* is also known as the adjacent matrix. Spectral clustering applies k-means clustering on the eigenvalues of the graph Laplacian of the neighboring matrix *W* to obtain the clusters.

Hierarchical. Clustering constructs nested clusters by combining (agglomerative) or splitting (divisive) data points in a tree or dendrogram. Divisive clustering begins with all data points in single cluster known as the root, which is then split into a set of child clusters until each cluster is a single data point. Agglomerative clustering begins with each data point as a cluster and proceed to combine the most similar pair of clusters until all the data points are unified into a single cluster. For this study, investigations were performed using agglomerative clustering with ward, complete and average linkages methods.

Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH). Clustering creates a compact summary of the original dataset as clustering feature (CF) entries, which is subsequently clustered instead of the given dataset. For a set of *N d*-dimensional objects, the clustering feature CF is a 3-D vector encapsulating characteristics about the data points and it is defined as

$$CF = (N, \overrightarrow{LS}, SS) \quad (2)$$

Where *N* is number of objects, $\overrightarrow{LS} = \sum_{i=1}^{N} \vec{X}_i$ is the linear sum of the data points $\vec{X}_i$ and $SS = \sum_{i=1}^{N} (\vec{X}_i)^2$ is the squared sum of the data points $\vec{X}_i$. BIRCH operates agglomerative clustering on the CF starting with each data point being a cluster. For two separate clusters, C_1 and C_2 with clustering feature $CF_1 = (N_1, \overrightarrow{LS_1}, SS_1)$ and $CF_2 = (N_2, \overrightarrow{LS_2}, SS_2)$, the clustering feature for the cluster that has been created by combining C_1 and C_2 will be

$$CF_1 + CF_2 = (N_1 + N_2\ ,\ \overrightarrow{LS_1} + \overrightarrow{LS_2},\ SS_1 + SS_2) \tag{3}$$

3 Experiments

There are a few publicly accessible human activities datasets with skeletal data which include MSR daily activity Dataset [18], UTKINECT dataset [7] and CAD60 dataset [17]. In the MSR daily activity dataset and the UTKINECT dataset, subjects completed each activity only once for most of the activities. To perform clustering, we require sufficient data points for each activity. Table 1 shows the summary of the characteristics of the three publicly available datasets and a dataset that we have collected. Among the three publicly available dataset, we have chosen to evaluate clustering performance on the CAD60 dataset.

Table 1. Summary of publicly available datasets and our dataset. *For 10 of the activities in CAD60

Datasets	Total samples	Samples/ class	Classes	Subjects	Modalities
MSRDailyActivity3D	320	20	16	10	RGB+D+3DJoints
CAD60	1200	120*	12	4	RGB+D+3DJoints
UTKINECT	150	15	10	10	RGB+D+3DJoints
Our dataset	2295	135	17	3	RGB+D+3DJoints

The CAD60 dataset comprises of 12 daily activities performed by four subjects with skeleton data of 15 joints. We have eliminated two activities that have only a few observations. We used 10 activities from the CAD60 dataset that we could sample sufficient instances, i.e. 120 instances for each activity. Activities from the CAD60 dataset that we have used are standing, brushing teeth, talking on phone, drinking water, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard and working on computer. Fig. 3 shows sample RGB images of the ten activities from CAD60 dataset. We have only used the skeleton data provided in the dataset. Data in CAD60 were obtained from OpenNI library that provides skeleton data with 15 joints. For CAD60, we have sampled 30 instances consisting of 30 frames each for each subject giving 120 instances for each activity. The feature vector for each activity instance comprised of 30 frames × 15 joints × 3 coordinates, i.e. 1,350 features.

Fig. 3. Instances of activities from the CAD60 dataset [17]. (1) brushing teeth, (2) writing on the board, (3) working in computer, (4) cooking (chopping), (5) opening pill container, (6) making phone calls, (7) having a drink, (8) sitting, (9) cooking (stirring)

To evaluate human activity discovery on more activities, we have collected a dataset ourselves. We have intentionally added activities that have locomotion activity, i.e. the walking and are motion intensive such as jumping, kicking and waving hands. We note activities in CAD60 have limited motion. We have collected the datasets using the Microsoft Kinect SDK that provides skeleton data with 20 joints, in contrast to the 15 joints in CAD60. For our dataset, three subjects in an indoor environment performed seventeen activities: standing, raising the right hand, raising the left hand, kicking the right leg, kicking the left leg, waving the right hand, waving the left hand, doing jumping jacks, walking, sitting down, being seated, standing up, making a phone call, drinking, picking up, sitting, and reading a book, and sweeping the floor. We recorded a sequence of each activity using a single stationary Kinect at 30 frames per second for at least 2 minutes. We have recorded both RGB and depth images, however we have only used the skeleton data derived from the depth images in this work. Sample RGB images from the dataset are shown in Fig. 4. For each action, 45 observations consisting of 70 frames each were sampled from the recording of each individual giving 135 instances for each activity. The feature vector for each activity instance comprised of 70 frames × 20 joints × 3 coordinates, i.e. 4,200 features. This differs from the feature vectors of CAD60 for the purpose of investigation.

Fig. 4. Sample images from 16 activities in our dataset. (1) raising right hand, (2) raising left hand, (3) kicking right leg, (4) kicking left leg, (5) waving right hand, (6) waving left hand, (7) jumping jacks, (8) walking, (9) sitting down, (10) seated, (11) standing up, (12) phone call, (13) having a drink, (14) picking up from the floor, (15) seated and reading book, (16) sweeping the floor.

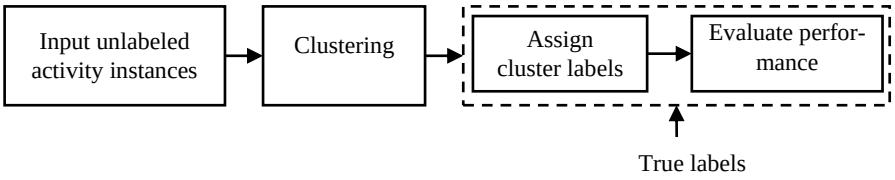

Fig. 5. Experimental flow. True labels were only used for evaluation purposes.

In Fig. 5, we have summarized the process involved in the experiment. The input to the clustering algorithm is all the transformed instances in a dataset. In this study, we have specified the number of clusters. Each of the clusters generated by the clustering technique was anticipated to be a group of the instances for an activity. True labels of the activity instances were used to evaluate the performance of the clustering algorithm. For evaluation purpose, each cluster was labelled as one of the activities in the dataset based on the label of its majority instances. For instance, if true labels for a dataset comprising of six instances from two activities were [0, 0, 0, 1, 1, 1] and the instances

were clustered into two clusters as [***1***, ***1***, ***0***, ***0***, ***0***, ***1***]. There were one activity 0 and two activity 1 in cluster ***0***; two activity 0 and one activity 1 in cluster ***1***. Based on majority, cluster ***0*** would be labelled activity 1, and cluster ***1*** would be labelled activity 0 giving the predicted labels as [0, 0, 1, 1, 1, 0]. Then we compared the predicted labels with ground truth, i.e. true labels, to evaluation the performance of the clustering. Evaluations were performed applying the K-means, Spectral, Agglomerative clustering with several linkage techniques (ward, average and complete linkage) and BIRCH clustering on both datasets. Random centroid initialization was used in K-means and spectral clustering. The K-means and spectral clustering algorithms were repeated ten times and average results are reported. For agglomerative and BIRCH clustering, the clustering results do not change in each run. Precision, recall and F1-score were computed for each result.

4 Results

Fig. 6 summarizes the result of each clustering algorithm performed on our dataset. Each bar represents the average precision, recall and F1-score of the clustering results for all the 17 activities. The rightmost bar shows the overall mean precision, recall and F1-score of all clustering algorithms, which were 58%, 68%, and 61% respectively. The agglomerative clustering with ward linkage and BIRCH clustering have achieved the same precision, recall and F1-score were 62%, 70% and 63% respectively, which was the highest performance among the clustering algorithms used.

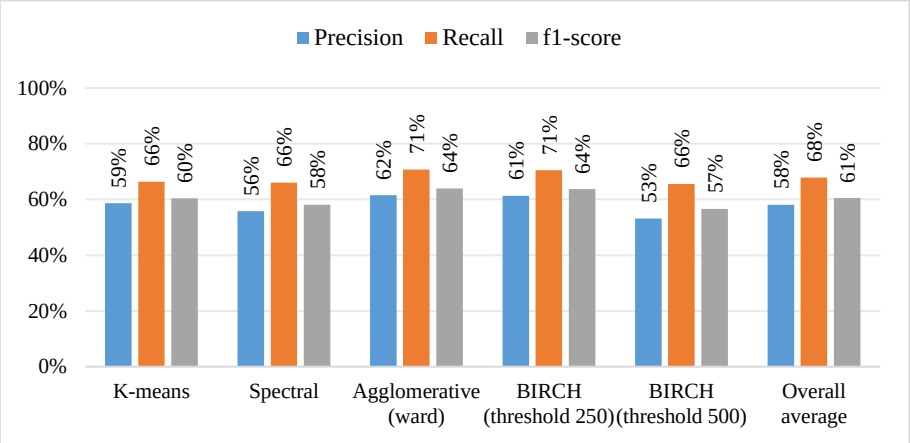

Fig. 6. Precision, recall and F1-score of different clustering algorithms on our dataset.

Fig. 7 shows the precision, recall and F1-score for the results on CAD60 dataset. Overall mean precision, recall and F1-score for all clustering algorithms were 61%, 69% and 63% respectively as shown by the rightmost column. Similar to the results on our dataset, agglomerative clustering (using ward and average linkage) and BIRCH

clustering have achieved the highest precision, recall and F1-score at 62%, 70% and 64% respectively.

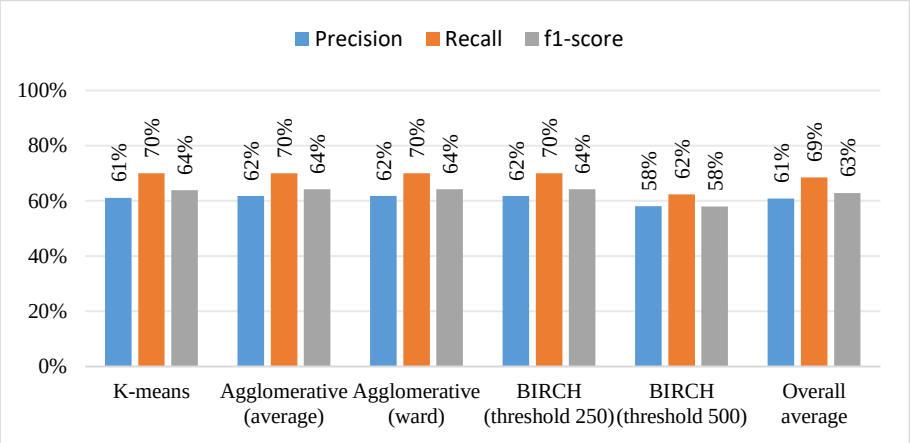

Fig. 7. Precision, recall and F1-score of different clustering algorithms on CAD60 dataset.

Comparing the results on both datasets, we observe that precision, recall and F1-score of k-means, agglomerative clustering with ward linkage and BIRCH clustering results were consistent and relatively high. However, clustering results for spectral clustering differed significantly on the two datasets. Determining two of the required parameters for BIRCH clustering (branching factor and threshold) was challenging. We have determined these parameters empirically for both datasets. Fig 8 summarizes the overall clustering performance on the two datasets. It can be observed that overall precision, recall and F1-score on our dataset was lower than that on CAD60 dataset. This is due to that our dataset contains more activities with more variation and a few highly similar activities. Fig. 9 shows the precision, recall and F1-score achieved by k-means, spectral, agglomerative with ward linkage and BIRCH (threshold of 250) clustering for every activity in our dataset.

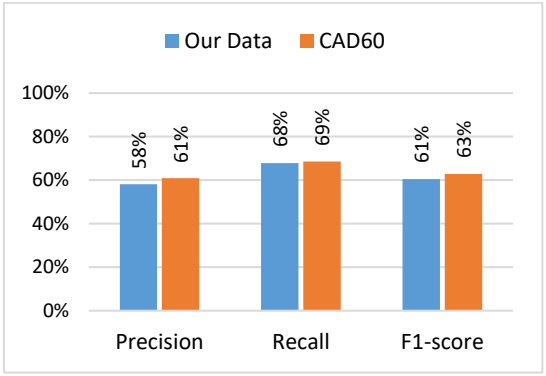

Fig. 8. Average precision, recall and F1-score achieved on both datasets.

For brevity, we have omitted the results of the other agglomerative and BIRCH clustering that were lower. It can be observed that activities which were grouped well by all of the clustering algorithms were: raise left hand, kick right leg, kick left leg, waiving right hand, waiving left hand, jumping jacks, seated, sitting and reading and sweeping floor. A few activities were poorly clustered including walking, drinking and sitting down. There were activities which were clustered well by some clustering algorithms while other clustering algorithms had poorly grouped them. For example, standing was clustered well by k-means, agglomerative and BIRCH clustering while with spectral clustering it was confused with other activities.

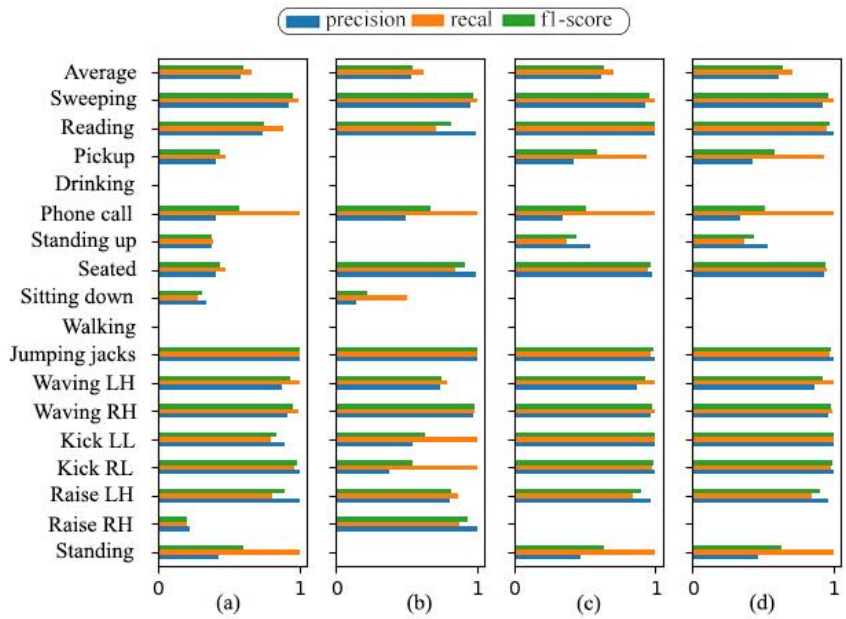

Fig. 9. Precision, recall and f1-score for every activity in our dataset. (a) k-means, (b) Spectral, (c) Agglomerative with ward linkage (d) BIRCH with threshold of 250.

5 Discussion and Analysis

Fig. 6 summarizes the result of each clustering algorithm performed on our dataset. Each bar represents the average precision, recall and F1-score of the clustering results for all the 17 activities. The clustering results on CAD60 dataset were on average better than that on our dataset. Our dataset contains more activities with more similar activities and more complex activities (walking, jumping) in comparison to the CAD60 dataset. To investigate further on why certain activities were not clustered well, the confusion matrices of the clustering results for k-means, spectral, agglomerative with ward linkage and BIRCH (threshold 250) clustering performed on our dataset are shown in Fig. 10(a), Fig. 10(b), Fig. 10(c) and Fig. 10(d) respectively. The activities are enumerated as in Fig. 4 and standing is enumerated as 0. It can be observed that walking (label 8)

and standing (label 0) were grouped in the same cluster by all the four clustering algorithms. The two activities appear similar when looking at postures. The features used in this study lack the motion or time series information. Raising right hand (label 1), drinking (label 13) and talking on the phone (label 12) were grouped in the same cluster by k-means, agglomerative and BIRCH as shown in Fig. 10(a), Fig. 10(c) and Fig. 10(d) respectively, however spectral clustering could effective cluster raising right hand (label 1) as shown in Fig. 10(c). Drinking and talking on the phone appear similar when looking at the postures. Raising right hand shares similarity in terms of right hand being raised. Coincidently, all subjects are right-handed hence drinking and talking on the phone were performed with right hand.

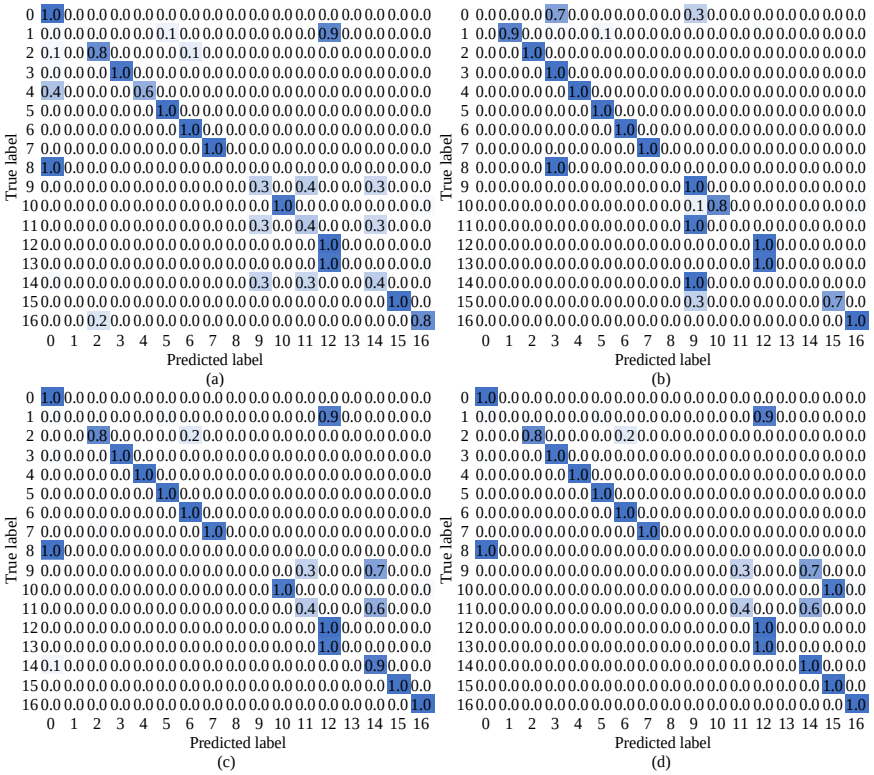

Fig. 10. Confusion matrix for the clustering results on our dataset with (a) k-means, (b) spectral, (c) agglomerative with ward linkage and (b) BIRCH with threshold 250.

Fig. 11 shows the confusion matrices for the results on CAD60 dataset. The activities labels are as shown in Fig. 3. It can be seen that drinking (label 3), brushing teeth (label 4) and talking on the phone (label 1) were confused by all clustering algorithms. These three activities are highly similar with their hands raised to the head. Cooking (chopping, label 7) was confused with cooking (stirring, label 8) by all clustering algorithms. The two activities are similar when viewed as postures.

The confusion matrices for the clustering results on both datasets show that k-means, agglomerative clustering (ward linkage) and BIRCH clustering (thresh hold 250) performed consistently on both datasets. However, BIRCH clustering performance varied significantly with different threshold values. It was challenging to determine the optimum threshold value. We experimented with different threshold values to find the threshold value that produced the best result. This is not feasible in real life application when the algorithm does not know the true labels to evaluate the clustering outcome. Spectral clustering performance varied significantly between the two datasets. On our dataset, the performance of spectral clustering was similar to that of the other clustering algorithms. However, spectral clustering has performed poorly on CAD60 dataset. Among all the clustering algorithms, agglomerative with ward linkage and k-means have produced consistent and relatively good results. One issue with k-means is that the result is dependent on random initialization of the centroids. In real life application, we can only rely on a single run of k-means, which may not be reliable.

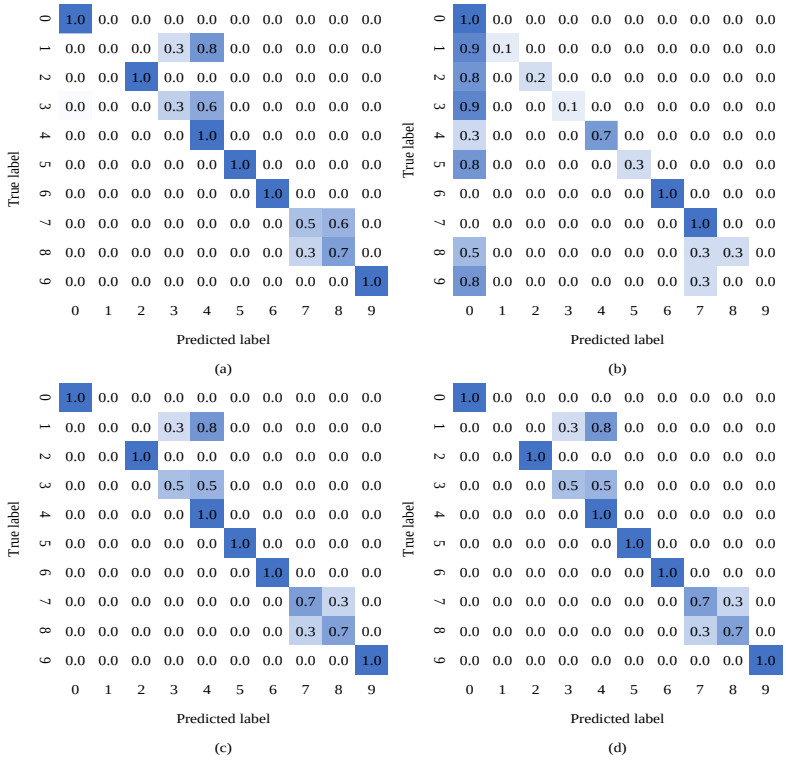

Fig. 11. Confusion matrix for the clustering results on (a) k-means, (b) spectral, (c) agglomerative with ward linkage and (b) BIRCH with threshold of 250.

The CAD60 dataset was used by many researchers to validate their methods on human activity recognition. Methods used by researchers include maximum entropy Mar-

kov model (MEMM), bag of words (BoW) method used with multi-class Support Vector Machine (SVM) and Hidden Markov Models (HMM). The performance of this study is compared with few of the algorithms. The average precision and recall score are shown in Table 2.

Table 2. Overall precision and recall score of this study compared with other methods

Authors	Method	Precision	Recall
Sung et al. [18]	Supervised with MEMM	68%	56%
Gaglio et al. [19]	Supervised with HMM and SVM	77%	77%
Shan and Akella [20]	Supervised with SVM and BoW	94%	95%
Cippitelli et al. [6]	Supervised with SVM and BoW	94%	94%
Proposed	Unsupervised/k-means/Agglomerative	61%	69%

6 Conclusion

This study was conducted with an aim to investigate the performance of several clustering algorithms to identify the most suitable one to explore unlabeled human activity data. Clustering algorithms were evaluated on two datasets. The average precision, recall and F1-score was 58%, 68%, and 61% on the lab recorded dataset. The precision, Recall and F1-score on the publicly accessible CAD60 dataset were 61%, 69% and 63%. The overall performance at current stage is lower than many states of the art supervised methods available to classify human activities. However, the clustering results have shown performance in par or better than the performance of classifiers trained with supervised learning on the CAD60 dataset (see Table 2). The results from this work have demonstrated the potential of clustering algorithms for human activity discovery without requiring labels. From the results, spectral clustering performed poorly, BIRCH clustering requires setting of parameters that would be difficult without the knowledge of the dataset, and k-means is subject to uncertainty of random centroids initialization. Agglomerative clustering with ward linkage appears to be the preferred clustering algorithm given the features based on joint positions in fixed number of frames.

As this is an exploratory work, a number of assumptions and conditions have been used to simplify the experiments. This study highlights that more problems need to be addressed besides improving the clustering performance which will be carried out as extension of this study. Firstly, the assumption of knowing the number of clusters or activities will be an issue in real life application. It is necessary to explore clustering algorithms that do not require knowing the number of clusters. Secondly, it is necessary to deal with the different durations of different activities. This is an unsolved problem even in HAR using supervised learning. In many cases, HAR has made assumption on activity duration based on the dataset. Thirdly, more useful features need to be extracted that capture the motion and time series information of the activities. Current set of features was effective to distinguish activities that are different in the postures involved,

however failed to distinguish activities that have similar postures that are distinguished by their motion (e.g. stirring food and chopping food).

References

1. Kim, E., Helal, S. & Cook, D.: Human Activity Recognition and Pattern Discovery. *IEEE Pervasive Comput.* **9**, 48–53 (2010).
2. Aggarwal, J. K. & Xia, L.: Human activity recognition from 3D data: A review. *Pattern Recognit. Lett.* **48**, 70–80 (2014).
3. Baisware, A., Sayankar, B. & Hood, S.: Review on Recent Advances in Human Action Recognition in Video Data. in *2019 9th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19)* 1–5 (IEEE, 2019). doi:10.1109/ICETET-SIP-1946815.2019.9092193.
4. Chen, L., Wei, H. & Ferryman, J.: A survey of human motion analysis using depth imagery. *Pattern Recognit. Lett.* **34**, 1995–2006 (2013).
5. Trong, N. P., Minh, A. T., Nguyen, H., Kazunori, K. & Le Hoai, B.: A survey about view-invariant human action recognition. in *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* 699–704 (2017). doi:10.23919/SICE.2017.8105762.
6. Cippitelli, E., Gasparrini, S., Gambi, E. & Spinsante, S.: A Human Activity Recognition System Using Skeleton Data from RGBD Sensors. *Comput. Intell. Neurosci.* **2016**, (2016).
7. Xia, L., Chen, C. & Aggarwal, J. K.: View invariant human action recognition using histograms of 3D joints. in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 20–27 (IEEE, 2012). doi:10.1109/CVPRW.2012.6239233.
8. Mohammadzade, H. & Tabejamaat, M.: Sparsness embedding in bending of space and time; a case study on unsupervised 3D action recognition. *J. Vis. Commun. Image Represent.* **66**, 102691 (2020).
9. Zheng, N. *et al.:* Unsupervised representation learning with long-term dynamics for skeleton based action recognition. *32nd AAAI Conf. Artif. Intell. AAAI 2018* 2644–2651 (2018).
10. Abdallah, Z. S., Gaber, M. M., Srinivasan, B. & Krishnaswamy, S.: AnyNovel: detection of novel concepts in evolving data streams: An application for activity recognition. *Evol. Syst.* **7**, 73–93 (2016).
11. Fang, L., Ye, J. & Dobson, S.: Discovery and Recognition of Emerging Human Activities Using a Hierarchical Mixture of Directional Statistical Models. *IEEE Trans. Knowl. Data Eng.* **14**, 1–1 (2019).
12. Gjoreski, H. & Roggen, D.: Unsupervised online activity discovery using temporal behaviour assumption. *Proc. - Int. Symp. Wearable Comput. ISWC* **Part F1305**, 42–49 (2017).
13. Kwon, Y., Kang, K. & Bae, C.: Unsupervised learning for human activity recognition using smartphone sensors. *Expert Syst. Appl.* **41**, 6067–6074 (2014).
14. Huynh, T., Fritz, M. & Schiele, B.: Discovery of activity patterns using topic models. in *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp '08* 10 (ACM Press, 2008). doi:10.1145/1409635.1409638.
15. Ye, J., Fang, L. & Dobson, S.: Discovery and recognition of unknown activities. *UbiComp 2016 Adjun. - Proc. 2016 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.* 783–792 (2016) doi:10.1145/2968219.2968288.

16. Ong, W. H., Koseki, T. & Palafox, L.: Unsupervised human activity detection with skeleton data from RGB-D sensor. *Proc. - 5th Int. Conf. Comput. Intell. Commun. Syst. Networks, CICSyN 2013* 30–35 (2013) doi:10.1109/CICSYN.2013.53.
17. Wang, J., Liu, Z., Wu, Y. & Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 1290–1297 (2012) doi:10.1109/CVPR.2012.6247813.
18. Sung, J., Ponce, C., Selman, B. & Saxena, A.: Unstructured human activity detection from RGBD images. *Proc. - IEEE Int. Conf. Robot. Autom.* 842–849 (2012) doi:10.1109/ICRA.2012.6224591.
19. Gaglio, S., Re, G. & Morana, M.: Human Activity Recognition Process Using 3-D Posture Data. IEEE Transactions on Human-Machine Systems 45, 586-597 (2015).
20. Shan, J. & Akella, S.: 3D human action segmentation and recognition using pose kinetic energy. 2014 IEEE International Workshop on Advanced Robotics and its Social Impacts (2014). doi:10.1109/arso.2014.7020983