Investigation of Feature Extraction for Unsupervised Learning in Human Activity Detection

Wee-Hong Ong, Leon Palafox, Takafumi Koseki
Graduate School of Engineering
The University of Tokyo
Tokyo, Japan
owh@koseki.t.u-tokyo.ac.jp

***Abstract*—The ability to understand what humans are doing is crucial for any intelligent system to autonomously support human daily activities. Technologies to enable such ability, however, are still undeveloped due to the many challenges in human activity analysis. Among them are the difficulties in extracting human poses and motions from raw sensor data, either recorded from visual sensor or wearable sensor and the need to recognize activities not seen before using unsupervised learning. Furthermore, human activity analysis usually requires expensive sensors or sensing environment. With the availability of low-cost *RGBD* (RGB-depth) sensor, the new form of data can provide human posture data with high degree of confidence. In this paper, we present our approach to extract features directly from such data (joint positions) based on human range of movement and the results of tests performed to check their effectiveness to distinguish sixteen (16) example activities are reported. Simple unsupervised learning, *K-means* clustering was used to evaluate the effectiveness of the features. The results indicate that the features based on range of movement significantly improved clustering performance.**

Keywords—human activity detection; human activity discovery; unsupervised learning; clustering; feature extraction; RGBD sensor

I. INTRODUCTION

In any system designed to support human daily activities, be it a smart living environment or assistant robot, understanding of human activities is a fundamental ability. Human activity analysis requires accurate capture of human postures and motions. Two major approaches to capture human poses and motions are uses of vision sensors and wearable devices. Wearable devices are often seen as obtrusive and inconvenient, while vision sensors post the challenges of solving computer vision problems. Solving the problem of extracting human poses reliably from the sensor data has been one of the major challenges in human activity analysis. Recently, the availability of low-cost *RGBD* (RGB-Depth) sensor has enabled accurate capture of human poses and has allowed researchers in human activity analysis to take a big leap to focus on analysis of the available postures and motions data. Features are extracted from the *RGBD* data and learning algorithms are applied to learn and recognize different activities.

In this paper, we present our work to identify a set of features that can be used to distinguish between different activities performed by human. Features have been extracted from the pose information obtained directly from the application programming interface (API) of an *RGBD* sensor. The approach aims to ensure that the set of features will comprise necessary information to distinguish between all possible activities that a human can possibly perform. While it is difficult to model human activities due to its wide variety and complexity, human movements are constrained by the range of movement. Feature extraction for the purpose of human activity analysis will benefit from this knowledge. We believe that given a correct set of features, an intelligent system can distinguish between different activities, and that it is sufficient for intelligent system to be able to distinguish the different activities. Recognition of activities will be achieved through interrogating human or other intelligent systems. This is similar to the way children learn about adult's activities. They could distinguish the different activities and ask about what they are. With this approach, the inputs to the learning system are unlabeled and unsupervised learning can be used. This is suitable in the natural setting of human living environment where intelligent systems can capture infinite data of human activities; however, the data will be unlabeled. The use of unsupervised learning offers the potential for automatic human activity discovery whereby an intelligent system can discover new activities by itself.

The remaining of this paper is organized as follows: *Section II* explains the difference between the work presented in this paper and other related works; *Section III* describes the approach used in our work to extract features based on human range of movements; *Section IV* describes the data used in the experiment and the tests carried out; *Section V* discusses the results obtained from the investigation; finally *Section VI* summarizes the findings from the experiment.

II. RELATED WORKS

Fundamentally human activity analysis is about recognizing the activity being performed from the postures and movements of a person. The actions or movements of the people are captured either with sensors attached to the person or vision sensors. In human activity analysis, vision-based solutions while challenging are preferred due to their unobtrusiveness and rich amount of information from visual data. However, traditional computer vision problems have been haunted by the challenges to perceive three-dimensional information from

two-dimensional images, notably illumination changes and foreground extraction. Aggarwal and Ryoo [5] provide detailed overview of various state-of-the-art research works on human activity recognition. It can be seen that significant efforts have been spent on accurately capture human motion (computer vision problem) and recognizing the activity from the motion (modeling and learning problems). In general, statistical modeling algorithms are developed that match motion sequences by explicitly modeling the probability distribution of an activity. Researchers have used several probability-based algorithms to build activity models. The *hidden Markov model* (*HMM*) and the *conditional random field* (*CRF*) are among the most popular modeling techniques [3].

Alternative methods to extract three-dimensional information from visual images have been explored such as the use of time-of-flight (ToF) camera [1] to recognize activities [6]. However, it was only since two years ago that a low-cost *RGBD* sensor, such as the *Microsoft Kinect* became easily available. Sung et al [4] used the *Microsoft Kinect* to detect human activities. They continued the approach of supervised learning, in which they first extract features from joint positions, orientations and hand movements based on the estimated human skeleton from the *Microsoft Kinect* and develop model from the labeled instances. They trained a *maximum-entropy Markov model* (*MEMM*) with hierarchical structure to learn models of twelve different activities.

The works described so far were based on supervised learning where labeled examples were provided to the learning algorithm. However, in long term, to enable intelligent systems to autonomously learn new activities, they will be required to deal with unlabeled data. For this reason, there have been increasing interest to investigate human activity discovery using unsupervised learning. Huynh et al [8] used clustering to generate a vocabulary of labels from sensor data, which are then used for pattern extraction using topic models to recognize daily routines. They used data from custom made wearable sensors. Stikic et al [7] applied two weakly supervised methods to discover activities from two published dataset obtained from wearable sensors.

The emphasis in the above described works have been on the learning algorithm, and there have been very few reported works based on data that can be obtained from the latest *RGBD* sensor. We investigate the existence of a feature set that can reliably distinguish human postures and motions. We extracted an exhaustive set of features from the estimated human skeleton from *RGBD* sensor API. The features were operated to select an optimal set. We believe that given an effective set of features, simple clustering algorithm can be used to distinguish different activities.

III. Feature Extraction & Learning

For an intelligent system to learn or extract information from a given set of features, the quality of the features is equally, if not more, important than the learning algorithm. The features should ideally contain relevant data suitable for the selected learning algorithm to learn the desired information.

A. *Human Range of Movement*

While it is difficult to model human activities due to its wide variety and complexity, human movements are constrained by the range of movement. Studies in kinematics of human motion [2][9] have identified possible movements around human joints including flexion, extension, lateral flexion, rotation of spinal column (the body movements); flexion, extension, abduction, adduction of shoulder joint (the arm movements); flexion, extension of elbow joint (the forearm movements); flexion, extension of knee joint (the leg movements); flexion, extension, adduction of hip joints (the thigh movements). *Fig. 1* provides some illustrations of human range of movement. In this paper, these angular movements have been used as features for human activity detection. As a side effect, the advantage of using joint angles is that they are scale independent, i.e. the size of the person does not normally affect range of movements.

B. *Features*

Feature extraction in the context of this paper is not about image processing. The raw data were coordinates of 15 joints in human skeleton as shown in *Fig. 2*. These coordinates were determined by the *OpenNI* [10] API from the images (frames) captured from *Microsoft Kinect RGBD* sensor. It is the availability of such data that the work reported in this paper concentrated in extracting features from this form of data.

A few assumptions have been made when considering the feature extraction:

1. Sensor (camera) can be from any angle, however remains stationary during the whole activity duration (2 seconds);
2. Coordinates of the 15 joint positions are available reliably from sensor API;

Three sets of features were extracted: (1) set of 2700 features, (2) set of 2400 features and (3) set of 280 features. The set of 2700 features are simply the x, y, z coordinates of the 15 joint positions. Each activity example was captured for

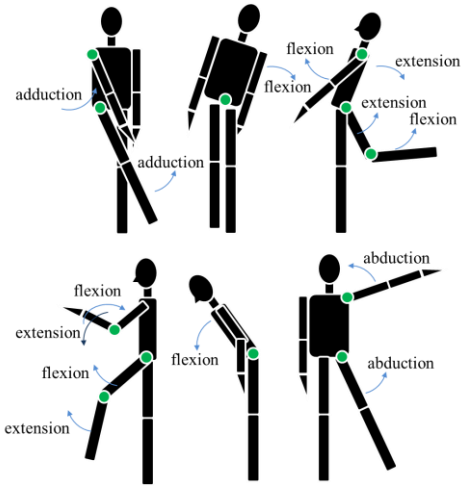


Figure 1. Illustrations of range of movement.

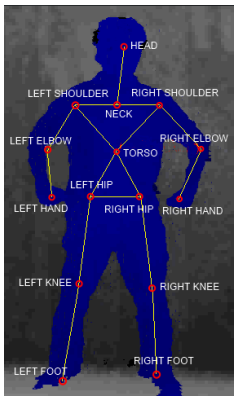


Figure 2. Human skeleton composed from fifteen (15) joint.

a window of 2 seconds at 30fps. Therefore, each example activity has 3 coordinates (x, y, z) for 15 joints in 30 frames per second for 2 seconds giving a total of 3*15*30*2=2700 features. All coordinates were transformed to the local coordinate frame located at the torso joint in the first frame of each example. By fixing the local coordinate frame in the first frame, instead of each frame, the information of translational movement, i.e. the person is not stationary at one location, can be retained. In this set of features, pose or shape information are assumed in all coordinates and temporal information are assumed across the 60 frames.

The set of 2400 features attempts to provide clearer picture of the human pose as compared to that in the set of 2700 features. Here, 40 features were extracted based on human range of movement as described in *Section III.A* and a few other pose and interaction related features. For each pose, i.e. in each frame, the following features were extracted from the coordinates of the joints: x, y, z coordinates of first frame, i.e. initial pose and position (3 features); the normalized (to shoulder width) distance from shoulder to foot at both sides (2 features); angles describing body flexion and turn (4 features); angles describing arms stretch (4 features); angles describing arms and legs bend (4 features); angles describing leg stretch (4 angles); normalized distance from hands to various interaction points on body (19 features). With 40 features per frame, the total number of features in this set was 40*60=2400 features per example activity. Many human activities involve the hand interacting with different parts of the body. It is therefore desirable to use such features to detect certain activity, e.g. drinking will have hand close to the head. In this set of features, the temporal information is assumed across the 60 frames.

The set of 280 features attempts to provide clearer picture of temporal information or movement. It did not sample frames from the 60 frames, but instead determined temporal information from all 60 frames. 7 features were extracted for each of the 40 features from all 60 frames of each example activity: first value; last value; difference between first and last values; speed around middle frames; max speed; acceleration around middle frames; max acceleration. The total number of features in this set was 7*40=280 features per example activity, which contain information from 60 frames.

C. *Unsupervised Learning*

K-means was used to evaluate if the feature extractions were able to make one set of features more distinguishable than the other set of features. *K-means* clustering is one of the simplest unsupervised learning algorithms. It looks for similarity among the examples in the dataset by using simple distance measurement. Given the required number of clusters, *K-means* group the points (examples) in the dataset by minimizing the distance from each data point to a cluster center (centroid). In the tests reported in this paper, *K-means* was used to find the centroids for the 16 activities given a subset of 50 examples. This was the learning phase. Since *K-means* is unsupervised and does not work with the labels, the assignment of centroids to corresponding activities was done as a post-process by assigning the centroid of each cluster to the activity with most membership in the cluster. After the centroids were identified, we performed cross-validation with a separate subset of 30 examples. Simple *Euclidean* distance measurement was used to assign each example to the nearest centroids found in learning phase. The assignment was then checked against the known label of these examples, i.e. was example of activity A being assigned (detected as) to centroid of activity A (as found in learning phase). This completed the cross-validation phase.

IV. DATA & EXPERIMENT

A. *Data*

The data used in the experiment were the coordinates of the 15 joints as shown in *Fig. 2*. *Microsoft Kinect RGBD* sensor was used to capture human activities. *OpenNI* API was used to process the visual input from the *Microsoft Kinect*, detect and provide the 15 joints coordinates in each frame. Sixteen (16) activities as listed below were captured:

1. Bowing
2. Drinking with left hand (standing)
3. Drinking with right hand (standing)
4. Sitting
5. Sitting down
6. Standing
7. Standing up
8. Talking on phone with left hand (standing)
9. Talking on phone with right hand (standing)
10. Walking
11. Wave ‘bye’ with left hand (standing)
12. Wave ‘bye’ with right hand (standing)
13. Wave ‘come’ with left hand (standing)
14. Wave ‘come’ with right hand (standing)
15. Wave ‘go away’ with left hand (standing)
16. Wave ‘go away’ with right hand (standing)

Each activity example has a duration of 2 seconds. A number of the above activities are in common interest of human activity analysis researches, and a number of them are meant to be confusing, e.g. drinking with left hand (2) and talking on phone with left hand (8) are close to each other with subject’s left hand close to his head. At the moment, the data have been recorded for one subject only.

Around 100 examples were recorded for each activities, however for the experiments reported in this paper, 80 examples from each activity were used. Three set of features were extractred from these examples as described in *Section III.B*. Lets call them set of 2700, set of 2400 and set of 280 features, and we have 80 examples (dataset) with each set of features.

B. *Experiment*

K-means clustering was performed on each set of features for the first 50 examples during the learning phase. Then, centroids from the learning phase were tested with the subset of the remaining 30 examples in the cross-validation phase, as described in *Section III.C*. Three rounds of the above tests were conducted. In each round, a confusion matrix was produced. The average of the confusion matrices from the three rounds in each phase was used to calculate the precision, recall and $F_{0.5}$ score. Initial test runs were performed for different number of clusters, K, to get an impression of potential improvement especially during cross-validation. K=32 appeared to be a reasonable value with good performance while not having many unoccupied clusters during the cross-validation. However, ideally we want K=16. Results for both K=16 and K=32 are discussed in next section.

V. RESULTS & DISCUSSION

Tables I to *IV* present the precision, recall and $F_{0.5}$ score for some of the tests performed. All results are average of three runs of *K-means* clustering. The results are grouped for the different set of features described in *Section III.B*. *Fig. 3* and *Fig. 4* give the summary of the performance in learning and cross-validation phases. *Fig. 5* and *Fig. 6* present the confusion matrices for some of the tests performed. In each confusion matrix, each row is an actual activity (actual class), while each column is the cluster (predicted class) assigned to the respective activity, e.g. column 1 is the cluster of activity (1). The corresponding activity to each number is as given in *Section IV.A*. For compactness, the numbers in the confusion matrices have been transformed to gray scale with completely black representing 1 while completely white representing 0. As a reference, the gray scale color bar is given in *Fig. 5*.

Fig. 3 shows the summary of performance in learning phase, i.e. with the subset of 50 examples, for the three sets of features. *Table I* shows the $F_{0.5}$ score achieved during the learning phase. Increasing the number of clusters from 16 to 32 significantly improved the clustering accuracy in all set of features. In general, the set of 2400 features performed better than the set of 2700 features, while the set of 280 features did not help to improve performance when compared to the set of 2400 features. However, at almost ten times less number of features, the set of 280 features performed better than the set of 2700 features. The extraction of features based on range of movement has improved the clustering performance. However, four of the activities were more confused with the extracted features as compared to the set of 2700 features: drinking with left hand (2), standing (6), talking with left hand (8) and walking (10). Looking at the confusion matrix for the result from the set of 2400 features at K=32 given in *Fig. 5*, activity (2) was confused with activity (8) while activity (6) was confused with activity (10). The extracted features lost the necessary information that distinguishes these two pairs of activity.

TABLE I. $F_{0.5}$ SCORE FOR LEARNING WITH DIFFERENT PARAMETERS

No. of features:	2700		2400		280	
Activity **K:**	**16**	**32**	**16**	**32**	**16**	**32**
Bowing	78.1	88.3	73.4	95.9	40.8	99.6
Drinking with left hand	86.4	87.5	60.8	82.6	74.6	69.7
Drinking with right hand	35.8	60.8	78.6	100	38	85
Sitting	80	85.3	69.2	70.3	90.4	99.1
Sitting down	64.5	83.5	79	91.6	87.7	96.8
Standing	72.9	93.4	65.3	74.	33.3	84.3
Standing up	58	74	77.6	78.4	99.6	97
Talking on phone with left hand	48.2	78.2	40.1	71	64.2	65.4
Talking on phone with right hand	48.2	85.9	100	100	0	91.5
Walking	61	96.2	75.2	86	53.1	81.1
Wave 'bye' with left hand	41.7	86.5	61.4	92.5	62.4	82
Wave 'bye' with right hand	51.7	70.5	87.1	97.9	0	88.9
Wave 'come' with left hand	65.6	81.3	66.4	98	96.2	94.7
Wave 'come' with right hand	44.2	90.4	97.7	97.9	98.3	98
Wave 'go away' with left hand	68.3	72.7	67.5	97.3	66	92.8
Wave 'go away' with right hand	35.7	52.1	99.3	99.6	96.1	95
Average:	**58.8**	**80.4**	**74.9**	**89.6**	**62.5**	**88.8**

Fig. 4 shows the summary of the performance in cross-

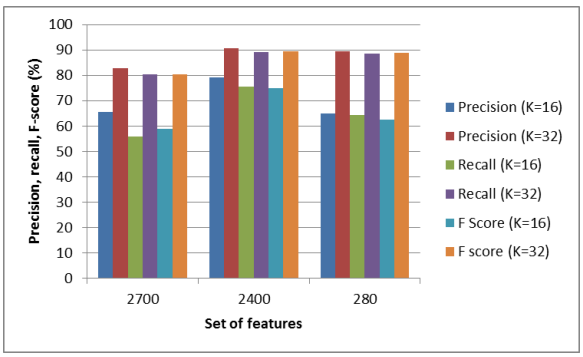


Figure 3. Summary of learning performance for three sets of features.

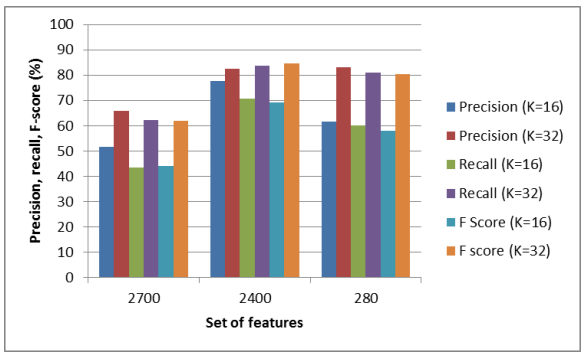


Figure 4. Summary of cross-validation performance for three sets of features.

TABLE II. PRECISION FOR CROSS-VALIDATION WITH DIFFERENT PARAMETERS

No. of features:	2700		2400		280	
Activity K:	**16**	**32**	**16**	**32**	**16**	**32**
Bowing	45.3	83.1	57.8	89.2	37.1	100
Drinking with left hand	100	73.1	46.9	91.8	61.3	54.5
Drinking with right hand	10.7	18.9	71.4	91.6	31.8	66.4
Sitting	100	100	66.7	61.6	100	100
Sitting down	45.3	75	73	100	100	100
Standing	56.4	86.3	61.5	61.6	32.6	69.6
Standing up	61.6	66.9	100	100	100	100
Talking on phone with left hand	83.3	57.8	100	7.2	53.6	52.6
Talking on phone with right hand	39.7	79.5	100	96.5	0	85.1
Walking	67.3	96.7	76.9	84.9	49.5	92.5
Wave 'bye' with left hand	38.9	95.3	75	85.7	59.6	75.6
Wave 'bye' with right hand	10.4	20.9	100	100	0	74
Wave 'come' with left hand	53.6	51.1	55.9	79	83.5	82.8
Wave 'come' with right hand	42.9	65.1	75.6	78.9	80.7	81.3
Wave 'go away' with left hand	71.8	82.1	80	92.3	93.8	98.1
Wave 'go away' with right hand	0	0	100	100	100	95.6
Average:	**51.7**	**65.7**	**77.5**	**82.5**	**61.5**	**83**

validation phase, i.e. with the subset of 30 examples, for the three sets of features. *Table II*, *III* and *IV* show the precision, recall and $F_{0.5}$ score achieved during the cross-validation phase. The same trend of performance increase with feature extraction is observed. The performances with 32 clusters (K=32) remained significantly better than those with 16 clusters (K=16) during the cross-validation. The performance with the set of 2400 features remained superior compared to the other two set of features. At K=32, the highest $F_{0.5}$ score was 97.5%, while the lowest was 66.2%. *Fig. 6* shows the cross-validation confusion matrices for the results from three sets of features at K=32. It is clear that the set of 2400 features was least confusing to the learning algorithm as its confusion matrix, given in *Fig. 6(b)*, has the least off-diagonal elements. Among those activities that were confused, the feature extracted in the set of 280 features appeared to make sitting (4) and standing (7) much less confused as compared to the other two sets of features.

The results indicate the potential of tweaking the features to enable simple clustering algorithm to distinguish between different activities. Nevertheless, *K-means* clustering suffers from the problem of initialization and does not always find the global optimal solution. The clustering performance of *K-means* greatly depends on the initial centroids. It is desirable to adapt a more robust clustering algorithm to evaluate the suitability of extracted features.

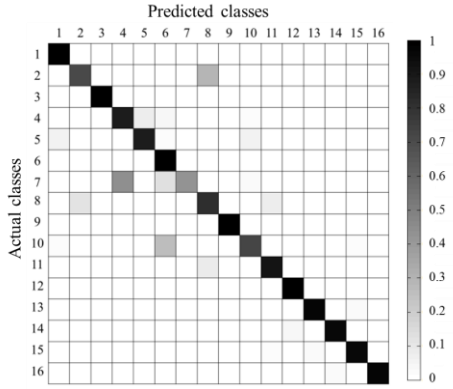


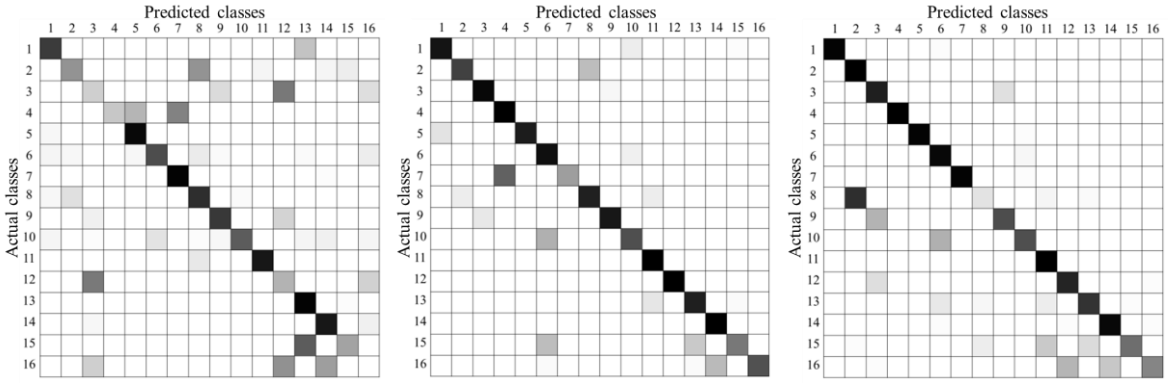
Figure 5. Learning confusion matrix for 2400 features at K=32.

TABLE III. RECALL FOR CROSS-VALIDATION WITH DIFFERENT PARAMETERS

No. of features:	2700		2400		280	
Activity K:	**16**	**32**	**16**	**32**	**16**	**32**
Bowing	53.3	76.7	82.2	92.2	62.2	98.9
Drinking with left hand	24.4	42.2	100	74.4	75.6	100
Drinking with right hand	15.6	18.9	100	96.7	100	87.8
Sitting	10	22.2	66.7	100	66.7	100
Sitting down	32.2	96.7	90	88.9	66.7	98.9
Standing	63.3	70	71.1	94.4	33.3	96.7
Standing up	100	98.9	66.7	37.8	100	98.9
Talking on phone with left hand	61.1	82.2	4.44	86.7	50	11.1
Talking on phone with right hand	25.6	77.8	93.3	91.1	0	70
Walking	38.9	64.4	55.6	68.9	51.1	68.9
Wave 'bye' with left hand	90	91.1	66.7	100	96.7	100
Wave 'bye' with right hand	20	30	66.7	100	0	85.6
Wave 'come' with left hand	65.6	98.9	90	87.8	78.9	80
Wave 'come' with right hand	63.3	91.1	100	100	97.8	96.7
Wave 'go away' with left hand	31.1	35.6	13.3	53.3	16.7	56.7
Wave 'go away' with right hand	0	0	63.3	67.8	66.7	47.8
Average:	**43.4**	**62.3**	**70.6**	**83.8**	**60.1**	**81.1**

TABLE IV. $F_{0.5}$ SCORE FOR CROSS-VALIDATION WITH DIFFERENT PARAMETERS

No. of features:	2700		2400		280	
Activity K:	**16**	**32**	**16**	**32**	**16**	**32**
Bowing	46.7	81.8	61.5	89.8	40.3	99.8
Drinking with left hand	61.8	63.8	52.4	87.7	63.7	60
Drinking with right hand	11.4	18.9	75.8	92.6	36.8	69.8
Sitting	35.7	58.8	66.7	66.8	90.9	100
Sitting down	41.9	78.5	75.8	97.6	90.9	99.8
Standing	57.7	82.5	63.2	66.2	32.8	73.7
Standing up	66.8	71.5	90.9	75.2	100	99.8
Talking on phone with left hand	77.7	61.5	18.9	78.9	52.8	30.1
Talking on phone with right hand	35.7	79.2	98.6	95.3	0	81.6
Walking	58.7	87.9	71.4	81.2	49.8	86.6
Wave 'bye' with left hand	43.9	94.5	73.2	88.2	64.5	79.5
Wave 'bye' with right hand	11.5	22.3	90.9	100	0	76.1
Wave 'come' with left hand	55.7	56.6	60.4	80.6	82.6	82.2
Wave 'come' with right hand	45.8	69	79.5	82.4	83.7	84
Wave 'go away' with left hand	56.9	65	40	80.5	48.7	85.6
Wave 'go away' with right hand	0	0	89.6	91.3	90.9	79.6
Average:	**44.2**	**62**	**69.3**	**84.6**	**58**	**80.5**



(a) (b) (c)

Figure 6. Cross-validation confusion matrix for (a) 2700, (b) 2400, (c) 280 features at K=32.

VI. CONCLUSION

This paper presented the approach to extract features from *RGBD* sensor data based on human range of movement. The results from performing clustering on three sets of features indicated the features extracted based on human range of movement significantly improved clustering performance, as compared to direct use of joint coordinates. For the set of features extracted based on human range of movement, an average $F_{0.5}$ score of 89.6% was achieved in the learning phase and 84.6% was achieved in the cross-validation phase. However, some activities remained confused and further tweak of the feature set will be required. In addition, *K-means* suffers from inconsistent performance highly dependent on its initialization outcome. More robust unsupervised learning algorithm will be required to evaluate the effectiveness of the feature set.

REFERENCES

[1] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-flight sensors in computer graphics," in EUROGRAPHICS, 2009, pp. 119–134. [1]

[2] B. Mackenzie, "(2004) Range of Movement (ROM) [WWW]," available from: http://www.brianmac.co.uk/musrom.htm [Accessed 29/6/2012].

[3] E. Kim, S. Helal, and D. Cook, "Human Activity Recognition and Pattern Discovery," in Pervasive Computing, IEEE, January-March 2010, vol.9, no.1, pp.48-53. [4]

[4] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human Activity Detection from RGBD Images," in Association for the Advancedment of Artificial Intelligence Workshop on Pattern, Activity and Intent Recognition (PAIR), 2011, pp. 47-55. [10]

[5] J.K. Aggarwal, and M.S. Ryoo, "Human activity analysis: A review," in ACM Comput. Surv. 43, 3, Article 16, April 2011. [8]

[6] L.A. Schwarz, D. Mateus, V. Castaneda and N. Navab, "Manifold learning for tof-based human body tracking and activity recognition," in British Machine Vision Conference (BMVC), Aug 2010, pp. 1–11. [9]

[7] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, "Weakly Supervised Recognition of Daily Life Activities with Wearbale Sensors," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 12, December 2011, pp. 2521-2537.

[8] T. Huynh, M. Fritz, and B. Schiele, "Discovery of Activity Patterns using Topic Models," in UbiComp '08 Proceedings of the 10th International Conference on Ubiquitous Computing, 2088, pp. 10-19.

[9] V.M. Zatsiorsky, "Kinematics of Human Motion," Human Kinectics, 1998, ISBN: 0880116765.

[10] http://www.openni.org.