An Unsupervised Approach for Human Activity Detection and Recognition

Wee-Hong Ong, Takafumi Koseki
Department of Electrical Engineering and Information Systems, Graduate School of Engineering
The University of Tokyo, Hongo 7-3-1, Bunkyo-ku
Tokyo 113-8656, Japan
owh@ieee.org, koseki@koseki.t.u-tokyo.ac.jp

Leon Palafox
Department of Radiology
University of California, Los Angeles
Los Angeles, CA 90095-7437
leonpalafox@ucla.edu

***Abstract*—Human activity recognition is an important ability in any system that supports human in performing their daily activities. However, current supervised approach in human activity recognition is difficult to be deployed in the natural human living environment where labeled observations are scarce. In this paper, we demonstrate the use of K-means clustering and simple template models to achieve human activity detection and recognition in an unsupervised manner. The features used are extracted from the skeleton data obtained from an inexpensive RGBD (RGB-Depth) sensor. Our results show an average detection performance of 80.4% precision and 83.8% recall. The availability of an unsupervised approach in human activity recognition will make possible the wide adoption of human activity recognition in the natural human living environment.**

Keywords-human activity detection; human activity recognition; unsupervised learning, RGBD sensor

I. INTRODUCTION

Human activity support or assisted living systems are useful to address various social needs in the modern society such as personal assistant and elderly care. For such system to sensibly support us, it is useful for the systems to understand what we are doing. Human activity recognition is therefore an important component in such systems.

Human activity recognition (HAR) is an important area of computer vision research [1]. There are at least two major issues that hinder the use of HAR technologies in our natural living environment such as a regular home. First is the high cost of the required infrastructure and high precision vision sensors. Second is the need for labeled observations. The majority of human activity recognition technologies use supervised learning algorithms. These algorithms require labeled observations. However, labeled observations are scarce in our natural living environment, whereas there is abundance of unlabeled observations.

With the availability of inexpensive RGBD (RGB-Depth) sensors such as the Microsoft Kinect [2], the issue of expensive setup and sensors is partly resolved. RGBD sensors allow accurate detection of human posture using the depth information. In our work, we use the data from a RGBD sensor to address the second issue and deal with unlabeled observations. We are working on unsupervised human activity recognition based on visual data. With unsupervised human activity recognition, an intelligent system can autonomously detect new activities. This allows the system to function autonomously without requiring retraining or reprogramming to introduce new activity models as in the case of supervised learning.

In this paper, we adapt the features determined in our earlier work [3] and use a simple clustering algorithm to detect activities in an unsupervised manner. The detected activities are modeled using simple templates for recognition of unseen observations. We experiment with our dataset and a third party dataset [4]. The intention of this paper is to verify that the approach proposed based on simple clustering and templates can be applied to different activities and different subjects. We also investigate the effect on the clustering performance when the sampling frame rate is reduced.

The remaining of this paper is organized as follows: Section II describes related works in human activity recognition. Section III explains the proposed unsupervised approach. Section IV describes the datasets used and the experiment carried out. Section V presents the results and discusses them. Finally, Section VI summarizes the findings from the experiments and suggests further works.

II. RELATED WORKS

Human activity recognition can be achieved through various means. There have been researches on activity recognition based on data from wearable sensors [5], [6], [7]. However, the requirement to wear sensors makes such systems obtrusive and less preferable to those using visual data. For this reason, a large number of researches in human activity recognition are computer vision based [1]. These works focused on extracting human postures from visual images and modeling of activities from labeled observations.

However, in the natural human living environment, labeled observations are rare. For this reason, there have been increasing interest to investigate human activity recognition using unsupervised learning to deal with unlabeled observations that are abundance in the real-life situations.

For example, Wyatt et al. [5] described their techniques for mining object models from the web and used the information to recognize activities based on the interaction of

a user with the objects. They attached RFID tags to the objects. Stikic et al. [6] applied two weakly supervised methods to discover activities from two published datasets obtained from wearable sensors. Huynh et al. [7] 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. Song et al. [8] developed an EM-like algorithm on decomposable triangulated graphs to extract human as foreground from background clutter.

We can see from the above works that relate unsupervised learning to activity recognition have been either focusing on solving computer vision problems in an unsupervised manner, or they require alternative form of pre-labeled data from wearable sensors or other sources.

In our earlier work [3], we investigated the extraction of features from skeleton data obtained from low-cost RGBD sensor and demonstrated the use of simple clustering algorithm, K-means, to detect activities from unlabeled observations. The outcome was groups of unlabeled observations of the same activity. Post operation can be called on these groups of activities to obtain their templates or models. These templates do not have label, however, they can be labeled through human-machine interaction. This approach will enable an intelligent system to autonomously handle the whole process of activity detection, learning, labeling and recognition.

In this paper, we demonstrate the potential of simple clustering algorithm and template modeling for unsupervised approach in human activity detection and recognition on a total of twenty five activities performed by five different subjects.

III. Unsupervised Approach for Human Activity Detection and Recognition

In this section, we describe the various technologies used in the unsupervised approach for human activity detection and recognition. In particular, we focus on the feature extraction, activity detection and recognition.

A. Features from Skeleton Data

Feature extraction in the context of this paper is not about image processing. The raw data are coordinates of 15 joints in human skeleton as shown in Fig. 1. These coordinates have been obtained directly from the OpenNI SDK [9] for the RGBD sensor, Kinect. For an intelligent system to learn or extract information from a given set of features, the quality of the features is as important as the learning algorithm. The features should ideally contain relevant data suitable for the selected learning algorithm to learn the desired information.

In our work, each activity observation is sampled for a window of two seconds. At 30fps, each activity observation contains 60 frames at full resolution. In this paper, we test the effect on clustering outcome when the frames per observation are reduced to 30, 15 and 6. The features are formed based on human range of movements [3].

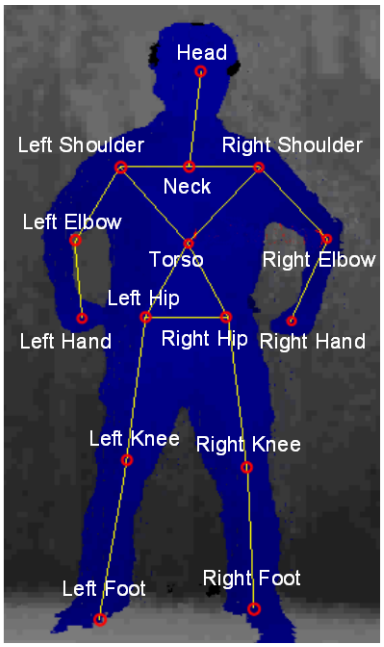

Figure 1. Human skeleton composed from fifteen (15) joints.

For each pose, i.e., in each frame, the following features are extracted from the coordinates of the joint positions: four vectors describing body flexion and turn; four vectors describing arms abduction and flexion; four vectors describing leg abduction and flexion and two vectors describing interaction between hands and head. The vectors are formed locally (between joints) and normalized to shoulder width making them view invariant to camera position and scale invariant to the size of the subject.

Fig. 2 illustrates a vector to represent right hand flexion. $\overrightarrow{rh}$ and $\overrightarrow{rs}$ are the vectors that represent the joint coordinates of right hand and right shoulder, respectively, in camera coordinate frame. A transformation to local coordinate frame is not required when taking vectors formed from two joints, e.g. the right hand flexion $\overrightarrow{rhf}$. There are fourteen 3-dimentional vectors giving 14×3=42 features per frame.

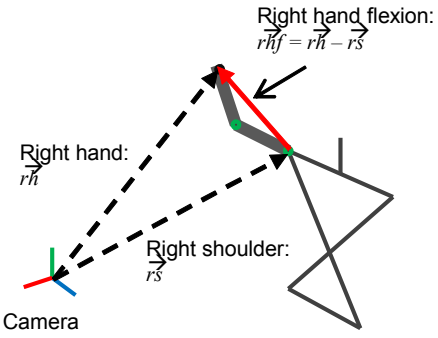

Figure 2. Using vector to represent range of movement.

B. Activity Detection by Clustering

Activity detection finds new activities, which have not been modeled, from unlabeled observations. We use K-means [10] clustering to group similar activities and discriminate one from the other. K-means clustering is one of the simplest unsupervised learning algorithms. K-means minimizes the cost function given in Eq. (1).

$$\boldsymbol{J} = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

where k is the number of clusters, n is the number of data points (observations), $x_i^{(j)}$ is ith data point in Cluster j $C_j = [x_1, x_2, \cdots, x_i, \cdots, x_N]$ and c_j is the centroid of C_j.

To account for the problem of random initialization in K-means, ten rounds were run and the outcome with the lowest cost function was taken.

C. Activity Recognition using Templates

Given that we have used K-means to discriminate one activity from another, an easy way to model each activity is to use the centroid of the cluster as the template for each activity.

$$\boldsymbol{c_j} = mean(x_i \in C_j) \quad (2)$$

where c_j is the centroid of the cluster $C_j = [x_1, x_2, \cdots, x_i, \cdots, x_N]$ and N is the number of members in cluster C_j.

An unseen observation is compared to all available templates to recognize it. The unseen observation is recognized as the activity corresponding to the nearest template to it. We use squared Euclidean distance to compute the similarity of an unseen observation to an activity centroid.

$$\boldsymbol{d_j}(x) = \sum_{p=1}^{P} \left(x_{(p)} - c_{j(p)}\right)^2 \quad (3)$$

where $d_j(x)$ is the squared Euclidean distance between an observation $x = [x_{(1)}, x_{(2)}, \cdots, x_{(p)}, \cdots, x_{(P)}]$ and the centroid $c_j = [c_{j(1)}, c_{j(2)}, \cdots, c_{j(p)}, \cdots, c_{j(P)}]$, and P is the number of dimensions for both x and c_j.

Simply assigning an observation to its nearest centroid will have problem in rejecting an activity that does not belong to any of the available centroids. To address this issue, we record the maximum radius of each activity cluster in the detection phase.

$$\boldsymbol{r_j} = max\left(d_j(x_i \in C_j)\right) \quad (4)$$

where $\boldsymbol{r_j}$ is the maximum radius of cluster $C_j = [x_1, x_2, \cdots, x_i, \cdots, x_N]$ and N is the number of members in cluster C_j.

An observation is only recognized as the activity corresponding to its nearest centroid if it's also within the radius. Therefore, an activity template is given by

$$\boldsymbol{H_j} = \left(\boldsymbol{c_j}, \boldsymbol{r_j}\right) \quad (5)$$

where $\boldsymbol{H_j}$ is the template or model of the activity j.

IV. Data & Experiment

In this section, we describe the datasets used and the experiment conducted to demonstrate the unsupervised approach of activity detection and recognition from features based on skeleton data from RGBD sensor.

A. Data

The work presented in this paper used the dataset "Cornell Activity Dataset CAD-60" [4]. CAD-60 is the earliest publicly available dataset obtained from Microsoft Kinect. CAD-60 consists of twelve daily activities: rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard and working on computer. The data was collected from four subjects: two males (referred as Person 1 and Person 4) and two females (referred as Person 2 and Person 3). One of the females is left-handed (Person 3). Still (standing) and random activity samples by each subject are also included in the dataset. All of the data was collected in a regular household setting with no occlusion of body from the view of sensor.

The CAD-60 dataset comprises of RGB images, depth images and skeleton data (coordinates of joint positions and orientations). We have only used the skeleton data of the joint positions in the work reported in this paper. Out of the twelve activities, we have considered eight of them as listed in Table I (1 to 4, 6 to 9). The other three activities have too few observations for the clustering algorithm. We also considered the still (standing) as one activity. In total, we have considered nine activities as illustrated in Fig. 3 and listed in Table I.

In addition, we have also used our dataset as described in [3]. This dataset has a single subject (referred as Person 5). The CAD-60 dataset contains activities with minimal or little movement, while our dataset contains activities with significant movements, e.g., waving hand. There are sixteen activities in this dataset, as illustrated in Fig. 4 and listed in Table II. "sit" refers to the stationary state of sitting on the chair. "stand" refers to the stationary state of standing on feet. "sit down" refers to the transition action from stand to sit, whereas "stand up" refers to the transition from sit to stand. There are three waving gestures. "wave bye" is the gesture of waving sideway. "wave come" is the gesture of waving the hand inward from high to low position. "wave go" is the gesture of waving the hand outward from low to high position. "(left)" indicates the activity being carried out with left hand, whereas "(right)" indicates a right handed activity.

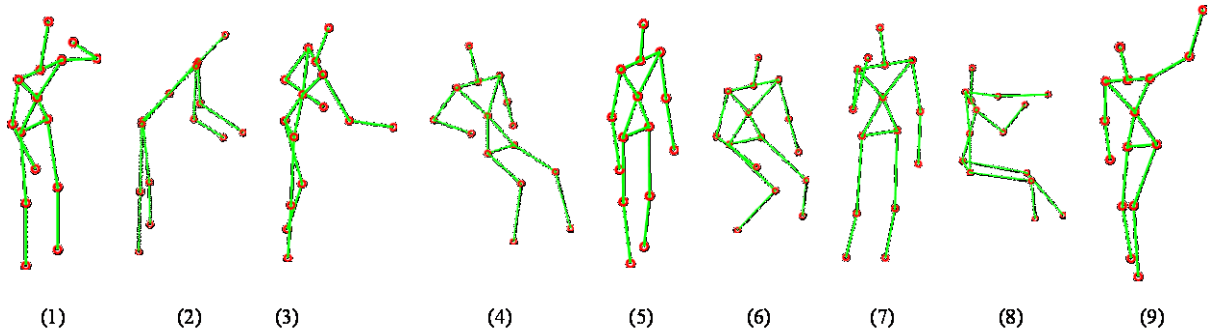

Figure 3. Snapshots of one random frame of the skeleton of the nine activities by Person 1 to 4 (1) brushing teeth, (2) cooking (chopping), (3) cooking (stirring), (4) relaxing on couch, (5) still (standing), (6) talking on couch, (7) talking on phone, (8) working on computer, (9) writing on whiteboard.

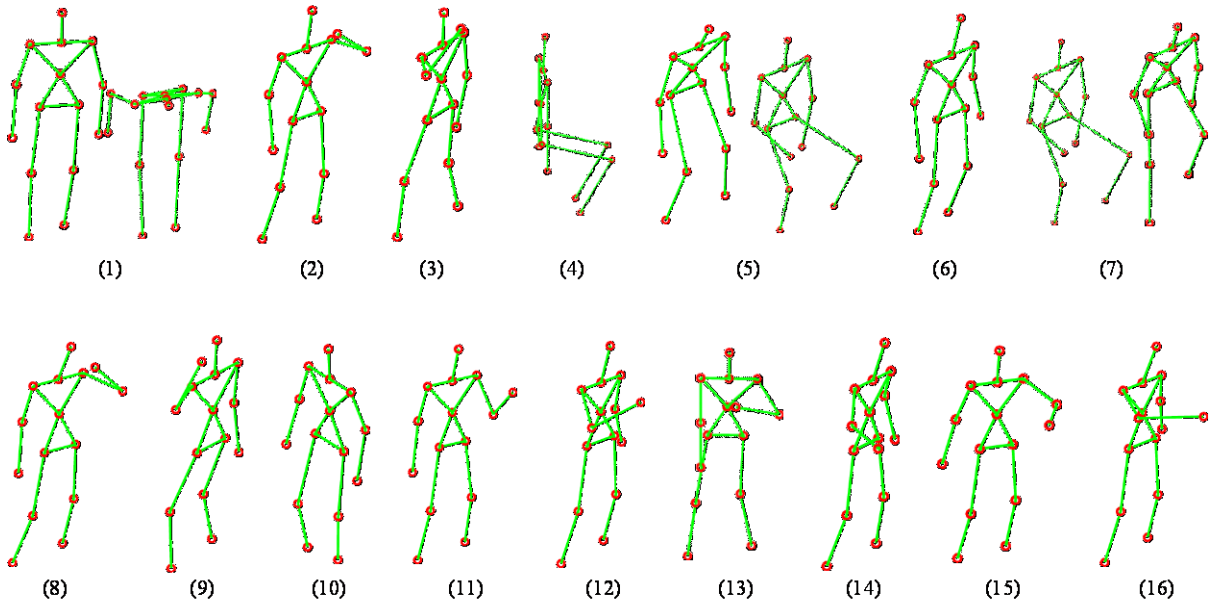

Figure 4. Snapshots of one* random frame of the skeleton of the sixteen activities by Person 5 (1) bowing, (2) drinking (left), (3) drinking (right), (4) sit, (5) sit down, (6) stand, (7) stand up, (8) talking on phone (left), (9) talking on phone (right), (10) walking, (11) wave bye (left), (12) wave bye (right), (13) wave come (left), (14) wave come (right), (15) wave go (left), (16) wave go (right). *Two frames from the beginning and end of an observation are taken for activities (1), (5) and (7).

Table I. LIST OF ACTIVITIES FOR PERSON 1 TO 4.

1. brushing teeth
2. cooking (chopping)
3. cooking (stirring)
4. relaxing on couch
5. still (standing)
6. talking on couch (sitting)
7. talking on the phone
8. working on computer
9. writing on whiteboard

Table II. LIST OF ACTIVITIES FOR PERSON 5.

1. bowing	9. talking on phone (right)
2. drinking (left)	10. walking
3. drinking (right)	11. wave bye (left)
4. sit	12. wave bye (right)
5. sit down	13. wave come (left)
6. stand	14. wave come (right)
7. stand up	15. wave go (left)
8. talking on phone (left)	16. wave go (right)

B. *Experiment*

We first evaluated the clustering, i.e. activity detection, performance at different frames per observation. All activities from the dataset of each subject were pooled together, and K-means was ran on the dataset to discriminate different activties and find groups of activities. From this experiment, we determined an appropriate reduced frame rate to work on.

We used the clustering results at the reduced frame rate, as suggested from the outcome of the experiment above, to test the recognition performance of the templates. Each dataset (for each subect) was divided into 70% as training set and 30% as test set. The clustering was performed on the training set, and the template for each cluster was

obtained. The templates were used to recognize unseen observations in the test set.

In both experiments above, we report the average results of five runs. K-means is sensitive to random initialization and the outcome was different in each run.

V. RESULTS & DISCUSSION

Fig. 5 gives the average precision and recall score of all nine activities from the four subjects (Person 1 to 4) of CAD-60 dataset at different frames per observation. It can be seen that reducing the frames from 60 to 30 has no effect on the performance of clustering while reducing to 15 frames only lowered the precision and recall by not more than 1%. Further reduction to 6 frames only lowered the clustering performance by further 1%. This however is anticipated as majority of the activities in the dataset do not involve much motion.

For activities with significant motion, e.g., waving hand, we expect the clustering performance to degrade significantly at low frame rates. Fig. 6 shows the clustering performance on our dataset at different frames per observation. This dataset has activities with significant movements. It is surprising to see that at six frames per observation, the clustering result was significantly better than the results at higher frame rates. The inconsistency between the results for the two datasets indicates that at six frames per observation, the clustering outcome is highly dependent on the nature of the data. We can however observe that the clustering performances at 15, 30 and 60 frames per observation were consistent in both datasets. The results on both datasets indicate that at 15 frames per observation, the clustering result was close to that at full resolution of the sensor, i.e., 60 frames per observation. The result suggests that 15 frames per observation is a suitable reduced frames per observation without compromising the clustering performance.

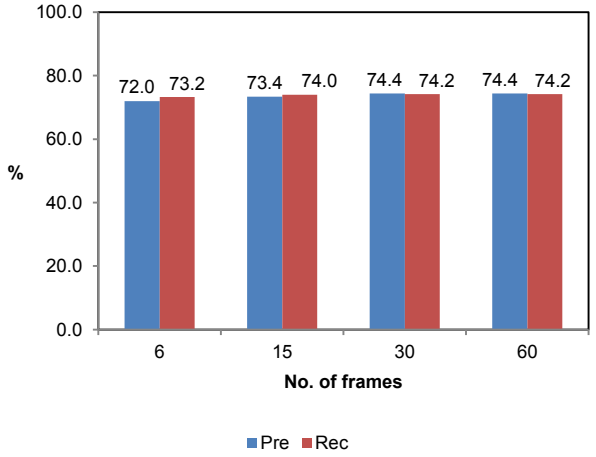

Figure 5. Average precision and recall of all nine activities for four subjects (Person 1 to 4) at different frames per observation.

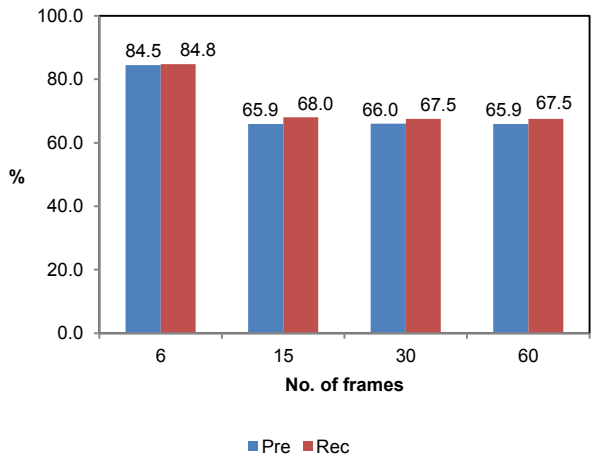

Figure 6. Average precision and recall of all sixteen activities for single subject (Person 5) at different frames per observation.

From here on, we discuss the results at 15 frames per observation. Fig. 7 shows the average precision and recall of the clustering of all the activities for each of the five subjects. The average clustering performance across all subjects and all activities has the precision of 80.4% and recall of 83.8%. The dataset of Person 5 has the lowest precision and recall at 68.6% and 75.6% respectively. This is due to the fact that there are significantly more activities to discriminate in this dataset and that there are significantly more movements in the activities.

Table III and the clustering columns in Table IV show the detailed clustering result for each activity for each subject. They are presented in two separate tables due to their different list of activities. The results in Fig. 7 correspond to the bottom row in Table III and the bottom row of the clustering columns in Table IV. From these tables, we see the clustering performance was good in most of the activities, with many achieving 100% precision and recall.

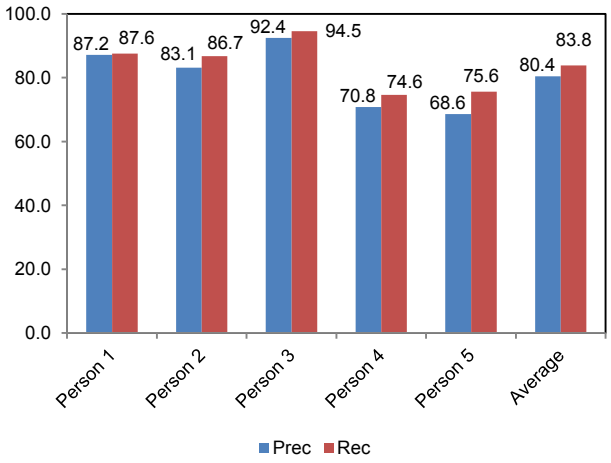

Figure 7. Average precision and recall of the clustering of all activities for all five subjects at 15 frames per observation.

Table III. PRECISION AND RECALL OF THE CLUSTERING RESULT FOR PERSON 1 TO 4.

		Person 1		Person 2		Person 3		Person 4		Average	
		Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
1	brushing teeth	46.1	77.5	67.8	99.3	89.0	93.9	66.5	63.9	67.3	83.7
2	cooking (chopping)	98.5	90.0	90.4	94.3	90.0	98.2	77.9	87.9	89.2	92.6
3	cooking (stirring)	90.9	99.3	84.3	72.5	79.0	80.0	48.9	64.6	75.7	79.1
4	relaxing on couch	100	100	100	100	100	100	100	100	100	100
5	still (standing)	100	98.6	100	100	98.6	100	53.8	80.0	88.1	94.6
6	talking on couch	90.0	100	100	100	100	100	100	77.9	97.5	94.5
7	talking on the phone	79.2	42.9	38.3	40.0	75.3	78.6	19.9	40.0	53.2	50.4
8	working on computer	80.0	80.0	97.8	100	100	100	90.2	100	92.0	95.0
9	writing on whiteboard	100	100	69.2	74.3	100	100	80.0	56.8	87.3	82.8
	Average:	**87.2**	**87.6**	**83.1**	**86.7**	**92.4**	**94.5**	**70.8**	**74.6**	**83.4**	**85.8**

Table IV. PRECISION AND RECALL OF THE RECOGNITION RESULTS FOR PERSON 5.

Person 5		Clustering		Recognition	
		Pre	Rec	Pre	Rec
1	bowing	81.6	98.6	84.6	100
2	drinking (left)	56.4	85.0	56.7	94.2
3	drinking (right)	11.4	14.3	10.0	10.8
4	sit	79.1	98.6	79.8	98.3
5	sit down	98.4	85.4	99.1	93.3
6	stand	76.2	76.1	76.1	78.3
7	stand up	58.6	57.9	60.0	60.0
8	talking on phone (left)	33.0	22.5	46.4	26.7
9	talking on phone (right)	55.9	87.9	52.4	88.3
10	walking	83.6	92.9	100	88.3
11	wave bye (left)	64.9	77.5	74.2	79.2
12	wave bye (right)	84.8	93.9	92.0	95.8
13	wave come (left)	69.5	74.6	70.7	73.3
14	wave come (right)	78.9	99.3	82.3	98.3
15	wave go (left)	87.5	79.3	89.2	77.5
16	wave go (right)	78.3	65.7	80.0	64.2
	Average:	**68.6**	**75.6**	**72.1**	**76.7**

K-means clustering outcome is sensitive to the randomness in its initialization. However, the effect is observed only with activities that are very similar. For example, significant confusion was observed between Activity 1 (brushing teeth) and 7 (talking on the phone) as they are very similar as illustrated in Fig. 8. Activity 2 (chopping) and 3 (stirring) are also very similar as illustrated in Fig. 9. In certain runs of K-means, Activity 5 (standing) was confused with a few standing activities such as Activity 1 (brushing teeth) and 7 (talking on the phone).

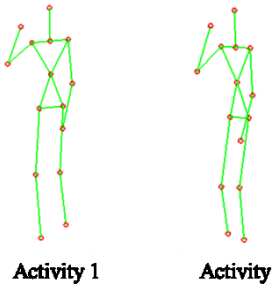

Figure 8. Skeleton of Activity 1 (brushing teeth) and Activity 7 (talking on the phone). They are very similar.

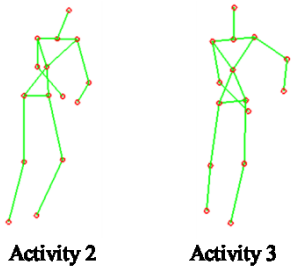

Figure 9. Skeleton of Activity 2 (chopping) and Activity 3 (stirring). They are very similar.

The consequence is that we observe low value of precision and recall in these activities in Table III.

In the case of the dataset for Person 5, there are more activities that are similar to each other. The detailed clustering results for Person 5 are given in Table IV. In this dataset, Activity 2 (drinking with left hand) and 8 (talking on phone with left hand); Activity 3 (drinking with right hand) and 9 (talking on phone with right hand), are very similar. The results were compiled from five runs of the clustering. To avoid clutter, we show only two confusion matrices out of the five runs to illustrate the state of confusion. The rows are the actual activities and the columns are the clusters obtained. In Fig. 10, we see that Activity 2 (drinking with left hand) and 8 (talking on phone with left hand) were confused with each other, and Activity 3 (drinking with right hand) was confused with Activity 9 (talking on phone with right hand). In this run, Activity 1 (bowing) was significantly confused with Activity 6 (stand). However, in another run, Activity 1 and 6 were not confused as shown in the confusion matrix in Fig. 11. In this run, Activity 8 (talking on phone with left hand) and 11 (wave bye with left hand) were confused instead. The effect of the random initialization of K-means is observed. However, for most activities, the effect was not significant.

Given the clusters from the detection phase, a template as described in Section III was obtained for each cluster. The intelligent system does not know whether the clusters were homogeneous or if they contain more than one activity. These templates were used to recognize unseen observations in the test sets as described in Section IV. The

P5	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
A1	55													1		
A2		41						15								
A3									49	3		4				
A4				56												
A5	4				47					4		1				
A6	52									2				2		
A7				1			54						1			
A8		28						26			2					
A9									53			3				
A10	2						2			51				1		
A11								3			53					
A12										1		55				
A13								1			4		51			
A14												1		55		
A15								1					51		4	
A16										1		2		7		46

Figure 10. Confusion matrix for clustering result for Person 5 in one of the five runs.

P5	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
A1	55									1						
A2		45						7			4					
A3									49	2		5				
A4				55	1											
A5	4				48					4						
A6						55				1						
A7				1			54							1		
A8		22						9		2	23					
A9									55			1				
A10						2				53				1		
A11								1			55					
A12									8	1		45				2
A13								1					55			
A14														56		
A15		1									1				54	
A16										1		1		9		45

Figure 11. Confusion matrix for clustering result for Person 5 in another one of the five runs.

recognition performance is summarized in Fig. 12. These results are dependent on the results of the clustering. The templates from those clusters containing confused activities would not be able to recognize well. We can see from Fig. 12 that the recognition performance is consistent with the detection performance in Fig. 7. An average precision of 81.4% and recall of 83.3% were achieved. Table V and the recognition columns in Table IV give the detailed results. We observe similar performance to that of the clustering phase. The recognition performance was generally good in most activities apart from those highly similar ones.

VI. CONCLUSION

We presented the results of our study of an unsupervised approach for human activity recognition. The approach uses K-means to detect unknown activities and uses template models to recognize known activities. The clustering achieved an average of 80.4% precision and 83.8% recall. The recognition performance achieved an average precision of 81.4% and recall of 83.3%. These results suggest the potential of performing unsupervised human activity recognition using just the skeleton data from an inexpensive RGBD camera.

There are a few issues we have identified to be addressed in our further works. The first is the ability of the system to determine the number of clusters, *k* value, by itself. There are a number of ways to determine the *k* value including the use of cluster validity indices. Another issue is the confusion of highly similar activities. By addressing the confusion issue, the problem of sensitivity to random initialization will be minimized.

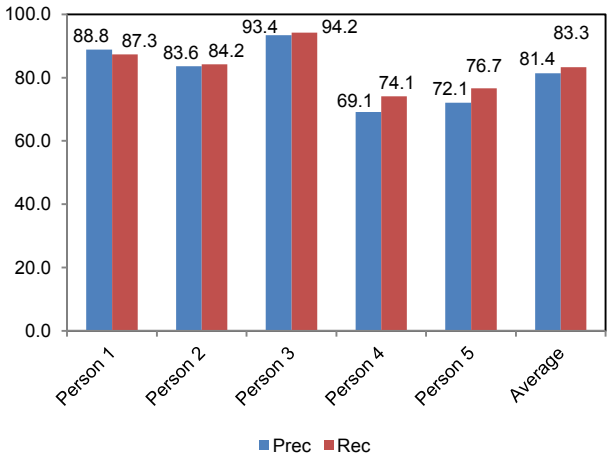

Figure 12. Average precision and recall of the recognition of all activities for all five subjects at 15 frames per observation.

Table V. PRECISION AND RECALL OF THE RECOGNITION RESULTS FOR PERSON 1 TO 4.

		Person 1		Person 2		Person 3		Person 4		Average	
		Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
1	brushing teeth	47.6	78.3	66.2	100	90.2	100	63.7	63.3	66.9	85.4
2	cooking (chopping)	100	94.2	80.2	95.0	90.0	97.5	78.9	93.3	87.3	95.0
3	cooking (stirring)	96.7	98.3	86.5	70.0	80.0	80.0	53.7	65.0	79.2	78.3
4	relaxing on couch	100	99.2	100	99.2	100	98.3	100	96.7	100	98.3
5	still (standing)	100	98.3	100	96.7	100	100	62.2	80.0	90.6	93.8
6	talking on couch	90.0	97.5	100	92.5	100	97.5	80.0	72.5	92.5	90.0
7	talking on the phone	85.2	42.5	39.2	40.0	80.0	76.7	17.7	40.0	55.5	49.8
8	working on computer	80.0	79.2	100	94.2	100	98.3	90.2	96.7	92.6	92.1
9	writing on whiteboard	100	98.3	80.0	70.0	100	99.2	75.5	59.2	88.9	81.7
	Average:	**88.8**	**87.3**	**83.6**	**84.2**	**93.4**	**94.2**	**69.1**	**74.1**	**83.7**	**84.9**

REFERENCES

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

[2] Z. Zhang, "Microsoft kinect sensor and its effect," IEEE MultiMedia, vol. 19, no. 2, pp. 4–10, 2012.

[3] W. Ong, L. Palafox, and T. Koseki, "Investigation of Feature Extraction for Unsupervised Learning in Human Activity Detection," in Bulletin of Networking, Computing, Systems, and Software, North America, 2, jan. 2013.

[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.

[5] D. Wyatt, M. Philipose, and T. Choudhury, "Unsupervised activity recognition using automatically mined common sense," in Proceedings of the National Conference on Artificial Intelligence, July 2005, Vol. 20, No. 1, p. 21.

[6] 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.

[7] 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.

[8] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," in Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003, 25(7), 814-827.

[9] OpenNI., "OpenNI | The standard framework for 3D sensing," http://openni.org/, 2010, [Accessed: 2012-04-30].

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, June 1967, Vol. 1, No. 281-297, p. 14.