2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks

Unsupervised Human Activity Detection with Skeleton Data From RGB-D Sensor

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
Neural Signal Processing Laboratory
Department of Radiology
University of California, Los Angeles
Los Angeles, CA 90095-7437
leonpalafox@ucla.edu

Abstract— **Human activity recognition is an important functionality in any intelligent system designed to support human daily activities. While majority of human activity recognition systems use supervised learning, these systems lack the ability to detect new activities by themselves. In this paper, we report the results of our investigation of unsupervised human activity detection with features extracted from skeleton data obtained from RGBD sensor. Unlike activity recognition, activity detection does not provide the label however attempts to distinguish one activity from another. This paper demonstrates a suitable set of features to be used with K-means clustering to distinguish different activities from a pool of unlabeled observations. The results show 100% F0.5-score were achieved for six out of nine activities for one of the subjects at low frame rate, while F0.5-score of 71.9% was achieved on average for all activities by four subjects.**

Keywords-human activity detection; unsupervised learning, clustering, RGBD sensor, feature extraction

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 can be achieved through various means. There have been researches on activity recognition based on data from wearable sensors [1], [2], [3]. However, the requirement to wear sensors makes such systems obtrusive and such systems are less preferable than those using visual data. For this reason, a large number of researches in human activity recognition are computer vision based [4]. These works focused on extracting human postures from visual images and modeling of activities from labeled examples.

There are two major issues that hinder the use of such technologies in our natural living environment. First is the high cost of the required infrastructure and high precision vision sensors. Second is the need for labeled observations. Human activity recognition has been using supervised learning algorithms. 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 technique [5]. These algorithms required labeled observations. However, labeled observations are scarce in real-life setting, while there is abundance of unlabeled observations.

For this reason, there have been increasing interest to investigate human activity recognition using unsupervised learning. At current stage, most of the works relating unsupervised learning with activity recognition have been either focusing on solving computer vision problems or they require alternative form of pre-labeled data from wearable sensors or other sources. For example, Song et al. [6] developed an EM-like algorithm on decomposable triangulated graphs to extract human as foreground from background clutter. Huynh et al. [3] 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. [2] applied two weakly supervised methods to discover activities from two published datasets obtained from wearable sensors. Wyatt et al. [1] described their techniques for mining object models from the web and use the information to recognize activities based on the interaction of user with objects. They attached RFID tags to the objects.

With the availability of low-cost RGBD (RGB-Depth) sensor such as the Microsoft Kinect, the first issue of expensive setup is partly resolved. RGBD sensors allow accurate detection of human posture using the depth information. In our work, we use the data from 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 allow the system to function autonomously without requiring retraining or reprogramming to introduce new activities models as in the case of supervised learning. In our previous work [7], 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 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

978-0-7695-5042-8/13 $26.00 © 2013 IEEE
DOI 10.1109/CICSYN.2013.53

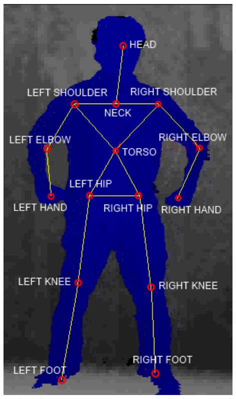

Figure 1. Human skeleton composed from fifteen (15) joints.

to autonomously handle the whole process of activity detection, learning, labeling and recognition. We are currently addressing the detecting stage. In our work [7], the experiments were performed on our own dataset of single subject. In this paper, we adapt the features determined in the earlier work and experiment with dataset from another published work [8]. Their dataset has four subjects. The intention of this paper is to verify that the approach proposed in [7] can be applied on different activities and different subjects. This paper further investigates the effect on the clustering performance when the sampling frame rate is reduced.

The remaining of this paper is organized as follows: Section II explains the features extraction from skeleton data; Section III describes the datasets used and the experiment; Section IV presents the results and discusses them; finally Section V summarizes the findings from the experiment and suggests further works.

II. 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 API of Microsoft Kinect RGBD sensor. 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.

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 [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). In an earlier paper [7], we investigated features extraction from skeleton data. The results showed that features based on range of movements can be used for K-means clustering algorithm to distinguish the different activities taking advantage of the constraint in human range of movement. However, the data used was captured by the authors on single subject only. In this paper, we tested this set of features on third party dataset comprising of four subjects as described in Section III.

Each activity observation was sampled for a window of 2 seconds. At 30fps, each activity observation contains 60 frames at full resolution. In this paper, we tested the effect on clustering outcome when the frames were reduced to 30, 15 and 6, i.e. 15, 7.5 and 3 fps respectively. For each pose, i.e., in each frame, the following features were 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 were formed locally (between joints) and normalized to shoulder width making them view invariant to camera 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 were fourteen 3-dimentional vectors giving 14×3=42 features per frame.

III. Data & Experiment

In this section, we describe the datasets used and the experiment conducted to demonstrate the proposed approach of activity detection from unlabeled observations.

A. Data

Currently, there are a few [8], [10] publicly available skeleton (coordinates of joints) datasets obtained from Microsoft Kinect sensor on human activities. None of these datasets have been adapted as benchmarking dataset for evaluation of human activity research works. The work presented in this paper used the dataset "Cornell Activity Dataset CAD-60" [8]. 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, working on

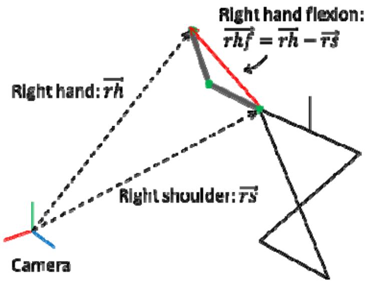
Right hand flexion:
$\overrightarrow{rhf} = \overrightarrow{rh} - \overrightarrow{rs}$
Right hand: $\overrightarrow{rh}$
Right shoulder: $\overrightarrow{rs}$
Camera

Figure 2. Using vector to represent range of movement.

Table I. List of activities

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

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 considered eight of them as listed in Table I (1 to 4, 6 to 9). The other three activities are not atomic in their dataset. We refer an atomic activity as one that cannot be further decomposed into sequence of smaller activities. For examples, drinking comprises of picking up the cup and drink; rinsing mouth comprises of sipping water, gargle and spit; opening pill container comprises of lifting the pill box, twist the cap and there are only three samples per subject. At this stage, we are interested to detect atomic actions or lower-level activities, which will be used to detect higher-level activities eventually. We also considered the still (standing) as one activity. In total, we considered nine activities as illustrated in Fig. 4 and listed in Table I. We sampled 50 observations (of 5, 15, 30 and 60 frames in 2 seconds windows) for each of the activities for each subject. Four datasets were composed at each number of frames: Person 1, Person 2, Person 3, Person 4: each comprising of 50 observations of each of the 9 activities for Person 1, 2, 3, 4 respectively, i.e., each dataset has the size of 450 observations. There were 16 datasets composed from four subjects at 5, 15, 30 and 60 frames per observation.

In addition, we also used our dataset as described in [7] to investigate the effect of reducing frame rate. The CAD-60 dataset contains activities with minimal or small movements, while our dataset contains activities with larger movements, e.g., waving hand.

B. *Experiment*

K-means clustering was performed on each set of the datasets. K-means [6] clustering is one of the simplest unsupervised learning algorithms. K-means minimizes the regular cost function given in Eq. (1).

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

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 and c_j is the centroid of Cluster j, i.e., C_j.

To account for the problem of random initialization, ten rounds were run and the outcome with lowest cost function was taken. For evaluation purpose, the clusters obtained were checked with true labels and identified as predicted classes of corresponding activities. Confusion matrix was constructed from the clustering result for each dataset, and precision, recall and $F_{0.5}$-score values were computed.

IV. Results & Discussion

Fig. 3 gives the average precision, recall and $F_{0.5}$ score of all activities from all four subjects 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, recall and

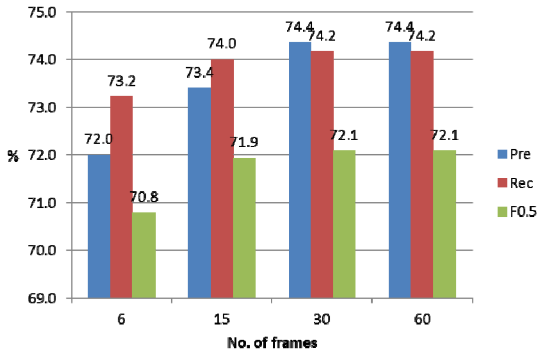

Figure 3. Average precision, recall and $F_{0.5}$ score of all activities from all four subjects at different frames per observation.

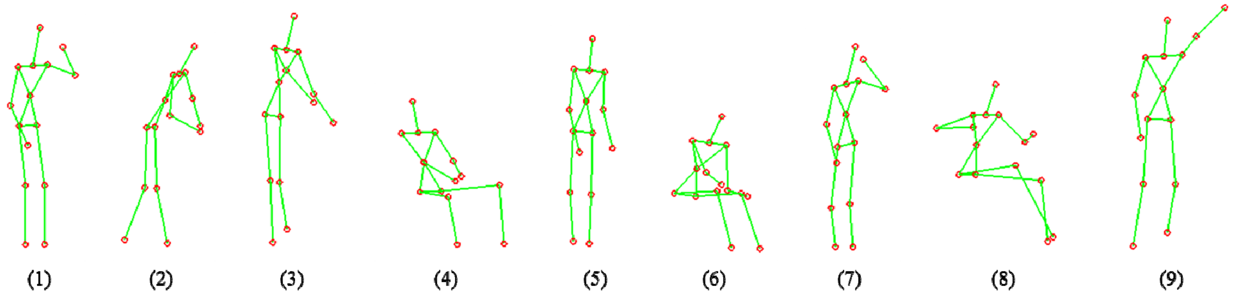

Figure 4. Skeleton of the nine activities (1) brushing teeth, (2) cooking (chopping), (3) cooking (stirring), (4) relaxing on couch, (5) still (standing), (6) talking on couch, (7) talking on phone, (8) working on computer, (9) writing on whiteboard.

$F_{0.5}$ score 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 these datasets 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. In order to test this hypothesis, K-means was run on our dataset [7], which has activities with significant movement, at the four numbers of frames per observation. The overall result is shown in Fig. 5. 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.

We discuss the detail results for the CAD-60 dataset at 15 frames per observation. These datasets of four subjects are referred as Person 1, Person 2, Person 3 and Person 4. As described in Section II, each frame has 42 features. With 15 frames, there are 42×15=630 features per observation. Fig. 6 shows the average precision, recall and $F_{0.5}$ score of the four subjects for each activity at 15 frames per observation. Activity 4 (relaxing on couch), 5 (still), 6 (talking on couch), 8 (working on computer) and 9 (writing on whiteboard) were consistently clustered with precision above 83%, recall above 82.5% and $F_{0.5}$ score above 84.4%. Fig. 7 shows the confusion matrices for the clustering result at 15 frames per observation. The rows are the actual activities as listed in Table I, while the columns are the predicted classes (clusters) for each activity. Looking at Fig. 7, we observe that Activity 4 (relaxing on couch) was clustered into a single homogeneous cluster for Person 2 and 3, while it was clustered into two homogeneous clusters for Person 1 and 4. This resulted in the reduced average recall value. However, in practice, this does not pose problem and two templates can be obtained for the activity with two clusters. Likewise, Activity 8 (working on computer) and 9 (writing on whiteboard) were clustered in homogeneous clusters. Activity 6 (talking on couch) was a little confused with Activity 1 (brushing teeth) for Person 4. This may be due to noise in the values of the six observations of Activity 1. The other clusters of Activity 6 (talking on couch) for Person 4 are homogeneous. Activity 5 (still) was clustered

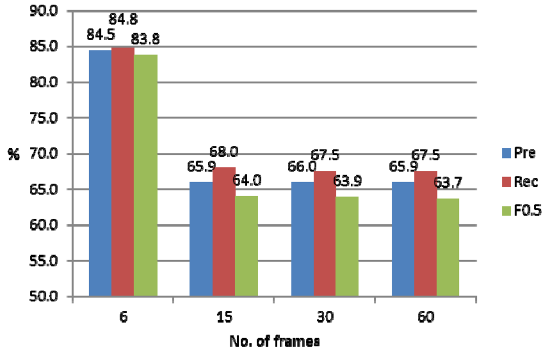

Figure 5. Average precision, recall and F0.5 score of all activities by the single subject in our dataset [7] at different frames per observation.

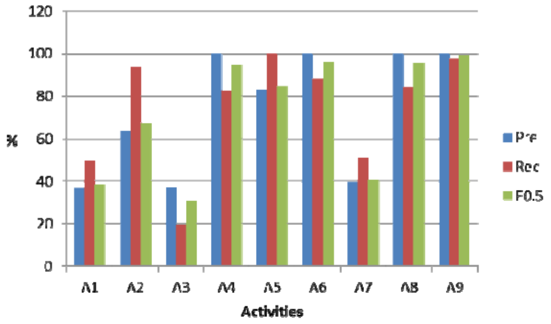

Figure 6. Average precision, recall and $F_{0.5}$ score of the four subjects for each activity at 15 frames per observation.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
A1					49		1		
A2		43	6				1		
A3			27				23		
A4	23			27					
A5					50				
A6						50			
A7					48		2		
A8								50	
A9									50

(a) Person 1

	C1	C2	C3	C4	C5	C6	C7	C8	C9
A1	50								
A2		44	6						
A3		38	12						
A4				50					
A5					50				
A6						50			
A7	50								
A8							14	36	
A9	4								46

(b) Person 2

	C1	C2	C3	C4	C5	C6	C7	C8	C9
A1	50								
A2		50							
A3		50							
A4				50					
A5					50				
A6						50			
A7							50		
A8			18					32	
A9									50

(c) Person 3

	C1	C2	C3	C4	C5	C6	C7	C8	C9
A1			6				44		
A2		50							
A3		46					4		
A4	12			38					
A5					50				
A6			23		1	26			
A7							50		
A8								50	
A9							1		49

(d) Person 4

Figure 7. Confusion matrices for clustering results at 15 frames per observation.

Table II. Precision, recall and $F_{0.5}$ score of the clustering results at 15 frames per observation.

		Person 1			Person 2			Person 3			Person 4		
		Pre	Rec	$F_{0.5}$	Pre	Rec	$F_{0.5}$	Pre	Rec	$F_{0.5}$	Pre	Rec	$F_{0.5}$
1	brushing teeth	0.0	0.0	0.0	48.1	100.0	53.6	100.0	100.0	100.0	0.0	0.0	0.0
2	cooking (chopping)	100.0	86.0	96.8	53.7	88.0	58.2	50.0	100.0	55.6	52.1	100.0	57.6
3	cooking (stirring)	81.8	54.0	74.2	66.7	24.0	49.2	0.0	0.0	0.0	0.0	0.0	0.0
4	relaxing on couch	100.0	54.0	85.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	76.0	94.1
5	still (standing)	34.0	100.0	39.2	100.0	100.0	100.0	100.0	100.0	100.0	98.0	100.0	98.4
6	talking on couch	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	52.0	84.4
7	talking on the phone	7.4	4.0	6.3	0.0	0.0	0.0	100.0	100.0	100.0	50.5	100.0	56.1
8	working on computer	100.0	100.0	100.0	100.0	72.0	92.8	100.0	64.0	89.9	100.0	100.0	100.0
9	writing on whiteboard	100.0	100.0	100.0	100.0	92.0	98.3	100.0	100.0	100.0	100.0	98.0	99.6
	Average:	**69.2**	**66.4**	**66.9**	**74.3**	**75.1**	**72.5**	**83.3**	**84.9**	**82.8**	**66.7**	**69.6**	**65.6**

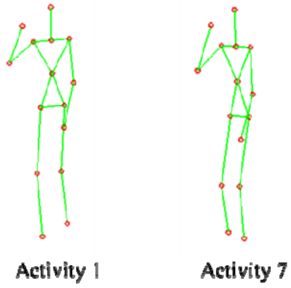

Figure 8. Skeleton of Activity 1 (brushing teeth) and Activity 7 (talking on the phone) captured on Person 1. They are very similar.

into homogeneous clusters for Person 2, 3 and 4, however it was completely confused with Activity 1 (brushing teeth) and 7 (talking on phone) for Person 1. This may be due to K-means had converged to local minima. Activity 1 (brushing teeth) and 7 (talking on phone) were consistently confused in the results for Person 1, 2 and 4. These two activities are very similar to each other as shown in Fig. 8. The average precision, recall and $F_{0.5}$ score are therefore low for these two activities as shown in Fig. 6. Likewise, Activity 2 (chopping) and 3 (stirring) are very similar as shown in Fig. 9, and resulting in low values of precision, recall and $F_{0.5}$ score in Fig. 6.

Table II gives the precision, recall and $F_{0.5}$ score for the results in Fig. 7. For Person 3, six of the activities were clustered with $F_{0.5}$ score of 100%. The values for Activity 8 (working on computer) were computed from one cluster, however the confusion matrix in Fig. 7 shows two homogeneous clusters. If both clusters were accounted for, 100% $F_{0.5}$ score was achieved on Activity 8 for Person 3. Those entries with zero value were due to confusion with other activities. The overall average for all four subjects is given in Fig. 3. At 15 frames, an average precision of 73.4%, recall of 74% and $F_{0.5}$ score of 71.9% were achieved.

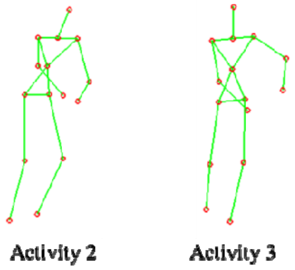

Figure 9. Skeleton of Activity 2 (chopping) and Activity 3 (stirring) captured on Person 3. They are very similar.

V. Conclusion

We used K-means to detect human activities from unlabeled observations with features extracted from skeleton data obtained from RGB-D sensor. This provides a way to perform unsupervised human activity detection. Our results show that K-means can successfully detect five out of nine activities with $F_{0.5}$ score higher than 80%. On average, $F_{0.5}$ score of 71.9% was achieved for all four subjects at 15 frames per observation. Two pairs of the activities are highly similar and the clustering algorithm was regularly confused. Extracting additional features is one solution to make similar activities distinguishable. While these new features may be extracted from visual data, it is useful to consider non visual data. For example, chopping and stirring can be easily distinguishable given sound information. Likewise, brushing teeth and talking on phone can be easily distinguishable if the digital device was communicating with the intelligent system responsible for activity detection. It's not uncommon even for human to be confused with similar activities based solely on visual observation.

However, a major drawback in the method used in this paper is the need to specify the number of clusters, i.e., the value of *k*. This is an important issue to be addressed. For the intelligent system to be completely autonomous, it has to determine the number of cluster by itself. This way, the system can autonomously detect activities from unlabeled observations. There have been various approaches proposed to autonomously find the value of *k*. Various cluster validity indices have been proposed to assist in deciding the value of *k*. This is one future work for us to undertake.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, June 2012, pp. 1290-1297.

[11] 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.