Unsupervised Activity Detection based on Human Range of Motion Features

Wee-Hong Ong, Takafumi Koseki

Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

owh@koseki.t.u-tokyo.ac.jp

***Abstract* — Human activity detection is usually achieved through supervised learning where a model is learned from given samples of each activity. With this approach, it is difficult for intelligent systems to automatically discover and learn new activities. We attempt to distinguish human activity using unsupervised learning. The approach relies on the fact that given appropriate features, the distinction between different activities can be observed from simple distance measurement. In this paper, we present our investigation to extract features from the coordinates of human 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 investigation on features that can be used to distinguish between different activities performed by human. Features were extracted from the pose information obtained directly from the application programming interface (API) of an RGBD sensor. The approach aims to identify the features that 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 [2],[9]. Feature extraction for the purpose of human activity analysis will benefit from this constraint. 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 interaction with 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 and learn new activities by itself.

II. 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

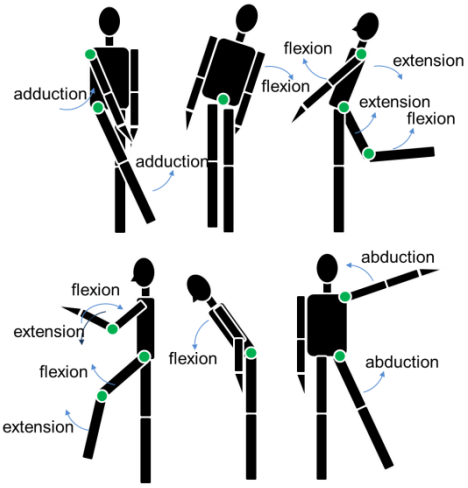


Fig. 1. Illustrations of range of movement.

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 view invariant.

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) Joint Coordinates, (2) Range of Movement and (3) Temporal Properties of Range of Movement. The set of Joint Coordinates (JC) are simply the x, y, z coordinates of the 15 joint positions. Each activity example was sampled from a window of 2 seconds comprising 15 frames. Therefore, each example activity has 3 coordinates (x, y, z) for 15 joints in 15 frames giving a total of 3*15*15=675 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 15 frames.

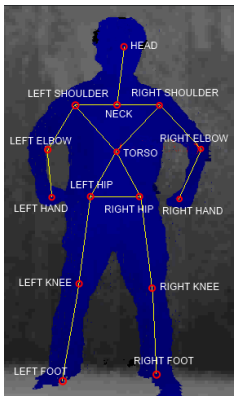


Fig. 2. Human skeleton composed from fifteen (15) joint.

The set of Range of Movement (ROM) attempts to provide clearer picture of the human pose as compared to that in the set of JC. Here, 36 features were extracted based on human range of movement as described in *Section II.A*. For each pose, i.e. in each frame, the following features were extracted from the coordinates of the joints: angles describing body flexion and turn (4 angles); angles describing arms abduction and flexion (4 angles); angles describing leg abduction and flexion (4 angles). Taking advantage of the rigid link nature of human limbs, a 3D vector was used to represent each ROM angle. The vectors were normalized to shoulder width to make the values scale invariant to the size of the subject. Note also the computation is view invariant, i.e. it doesn't matter where the camera is.

There were 12 3D angles giving 36 features per frame. With 36 features per frame, the total number of features in this set was 36*15=540 features per example activity. In this set of features, the temporal information is assumed across the 15 frames.

The set of Temporal Properties of Range of Movement (ROM-T) features attempts to provide clearer picture of temporal information or movement. It did not sample frames from the 15 frames, but instead determined temporal information from all 15 frames. 7 features were extracted for each of the 36 features from all 15 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

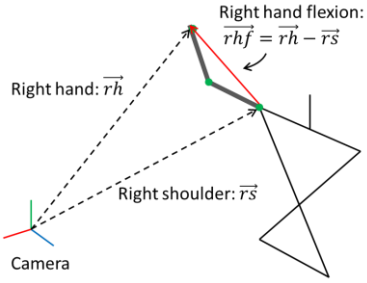


Fig. 3. Using vector to represent ROM.

was 7*36=252 features per example activity, which contain information from 15 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). Square Euclidean distance was used in the test, and the K-means minimized the regular cost function in (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 (samples), $x_i^{(j)}$ is ith data point in cluster j and c_j is the centroid of cluster j.

III. 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 extracted from these examples as described in *Section III.B*. Lets call them Joint Coordinates, Range of Movement and Temporal Properties of Range of Movement features, and we have 80 examples (dataset) with each set of features.

B. Experiment

K-means was used to find the centroids for the 16 activities given a subset of 50 examples. This has been called the learning phase in this paper. 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.

Three rounds of the above tests were conducted. In each round, a confusion matrix was produced with seeded random initial controids. The best outcome of the three rounds was recorded including the precision, recall and $F_{0.5}$ score.

IV. Results & Discussion

Tables I to *III* present the precision, recall and $F_{0.5}$ score for the three sets of features during learning and cross-validation phrases. *Fig. 4* and *Fig. 5* give the summary of the performance in learning and cross-validation phases. The values are the average precision, recall and $F_{0.5}$ score for all activities. *Fig. 6* shows the confusion matrices for

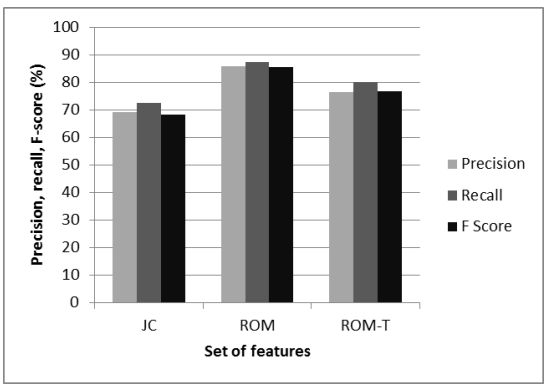


Fig. 4. Summary of learning performance for three sets of features.

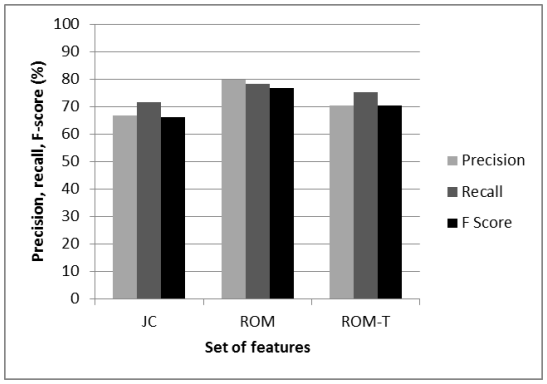


Fig. 5. Summary of cross-validation performance for three sets of features.

cross-validation result. 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 III.A*. For compactness, the numbers in the confusion matrices have been transformed to gray scale with completely black representing 1 (100%) while completely white representing 0. As a reference, the gray scale color bar is shown in *Fig. 6*.

There is clear indication in *Fig. 4* and *Fig. 5* that the set of features based on human range of movement (ROM) enabled significantly better clustering than the set of joint coordinates (JC) given the same input and clustering algorithm. The average $F_{0.5}$ score during learning was 85.6% with ROM and 68.2% with JC. In cross-validation, the average $F_{0.5}$ score was 76.7% with ROM and 66.3% with JCD. ROM has achieved improved clustering with smaller number of features (540 features) compared to JC (675 features). The set of temporal properties of range of movement (ROM-T) has performance better than JC but worse than ROM. However, ROM-T has significantly smaller number of features (252 features) compared with the other two sets of features.

A number of activities were difficult to distinguish by all three sets of features. From Table I to III, and the confusion matrices in *Fig. 6*, we observed that activities "drinking with left hand" (2) was confused with "talking on phone with left hand" (8) and "drinking with right hand" (3) was confused with "talking on phone with right hand" (9) in all three sets of features. ROM significantly improved detection of "sitting" (4), which has significantly different leg flexion compared to other activities. ROM-T has improved detection of "walking" (10), which is the only activity that translates its location.

The effect of the features on different nature of activities requires further study to identify effective combination of features capable to distinguish activities of different nature.

TABLE I. Learning (L) and Cross-Validation (CV) Precision

Feature set:	JC		ROM		ROM-T	
	L	CV	L	CV	L	CV
Bowing	98	90.6	98	90.6	100	96.7
Drinking with left hand	54.9	50.8	67.2	56.9	62.3	56.9
Drinking with right hand	0	0	0	0	67.9	40.7
Sitting	100	100	100	100	100	100
Sitting down	98	100	98	100	98	100
Standing	94.3	88.2	94.3	88.2	89.3	85.7
Standing up	100	100	100	96.8	100	100
Talking on phone with left hand	0	0	73.2	88.9	5.56	0
Talking on phone with right hand	76.9	69	51	54.5	76.4	61.4
Walking	100	100	100	100	100	100
Wave 'bye' with left hand	84.7	88.2	98	90.9	76.6	90.3
Wave 'bye' with right hand	58.1	62.5	92.6	85.7	0	0
Wave 'come' with left hand	50	47.4	100	57.8	100	76.9
Wave 'come' with right hand	98	73.2	100	73.2	100	77.8
Wave 'go away' with left hand	0	0	98	91.7	97.1	92.3
Wave 'go away' with right hand	96.1	100	100	100	50	50
Average:	**69.3**	**66.9**	**85.7**	**79.7**	**76.4**	**70.5**

V. CONCLUSION

This paper investigated features extraction 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 considerably improved clustering performance, as compared to direct use of joint coordinates. The approach reduces number of features without compromising clustering performance. For the set of features extracted based on human range of movement, an average $F_{0.5}$ score of 85.6% was achieved in the learning phase and 76.7% was achieved in the

TABLE II. Learning (L) and Cross-Validation (CV) Recall

Feature set:	JC		ROM		ROM-T	
	L	CV	L	CV	L	CV
Bowing	100	96.7	100	96.7	100	96.7
Drinking with left hand	100	100	78	96.7	86	96.7
Drinking with right hand	0	0	0	0	72	36.7
Sitting	50	63.3	98	100	98	100
Sitting down	98	90	98	90	100	96.7
Standing	100	100	100	100	100	100
Standing up	54	86.7	98	100	98	100
Talking on phone with left hand	0	0	60	26.7	2	0
Talking on phone with right hand	100	96.7	100	100	84	90
Walking	62	56.7	72	56.7	90	93.3
Wave 'bye' with left hand	100	100	100	100	98	93.3
Wave 'bye' with right hand	100	100	100	100	0	0
Wave 'come' with left hand	100	90	100	86.7	98	100
Wave 'come' with right hand	96	100	96	100	90	93.3
Wave 'go away' with left hand	0	0	100	36.7	68	40
Wave 'go away' with right hand	98	63.3	98	63.3	96	66.7
Average:	**72.4**	**71.5**	**87.4**	**78.3**	**80**	**75.2**

TABLE III. Learning (L) and Cross-Validation (CV) $F_{0.5}$ Score

Feature set:	JC		ROM		ROM-T	
	L	CV	L	CV	L	CV
Bowing	98.4	91.8	98.4	91.8	100	96.7
Drinking with left hand	60.4	56.4	69.1	62	66	62
Drinking with right hand	0	0	0	0	68.7	39.9
Sitting	83.3	89.6	99.6	100	99.6	100
Sitting down	98	97.8	98	97.8	98.4	99.3
Standing	95.4	90.4	95.4	90.4	91.2	88.2
Standing up	85.4	97	99.6	97.4	99.6	100
Talking on phone with left hand	0	0	70.1	60.6	4.1	0
Talking on phone with right hand	80.6	73.2	56.6	60	77.8	65.5
Walking	89.1	86.7	92.8	86.7	97.8	98.6
Wave 'bye' with left hand	87.4	90.4	98.4	92.6	80.1	90.9
Wave 'bye' with right hand	63.5	67.6	94	88.2	0	0
Wave 'come' with left hand	55.6	52.3	100	61.9	99.6	80.6
Wave 'come' with right hand	97.6	77.3	99.2	77.3	97.8	80.5
Wave 'go away' with left hand	0	0	98.4	70.5	89.5	73.2
Wave 'go away' with right hand	96.5	89.6	99.6	89.6	55.3	52.6
Average:	**68.2**	**66.3**	**85.6**	**76.7**	**76.6**	**70.5**

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 deterministic unsupervised learning algorithm will be required to evaluate the effectiveness of the feature set.

REFERENCES

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

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

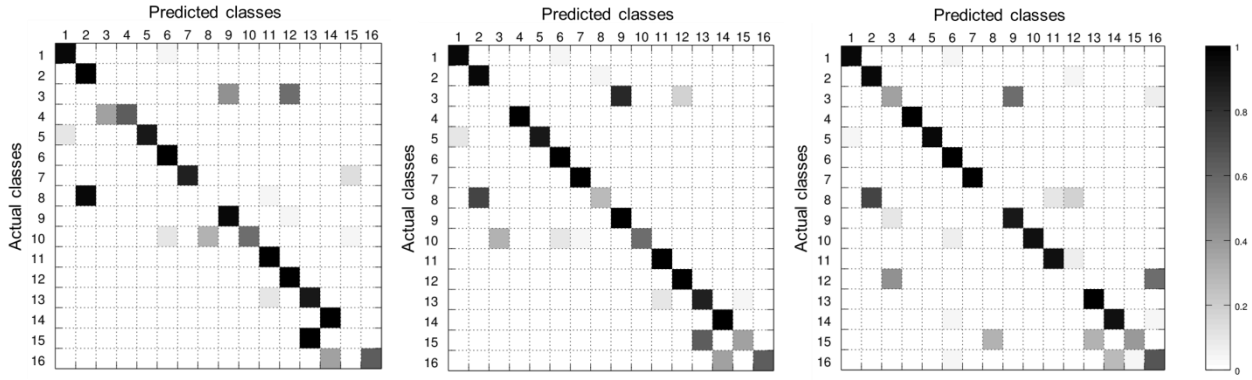
[6] L.A. Schwarz, D. Mateus, V. Castaneda and N. Navab, "Manifold learning for tof-based human body tracking and activity recognition," in British Machine Vision Conference (BMVC), Aug 2010, pp. 1–11.

[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.



(a) (b) (c)

Fig. 6. Cross-validation confusion matrix for (a) JC, (b) ROM, (c) ROM-T.