電気学会論文誌 C（電子・情報・システム部門誌）
IEEJ Transactions on Electronics, Information and Systems
Vol.134 No.11 pp.1724–1730 DOI: 10.1541/ieejeiss.134.1724

Paper

An Incremental Approach of Clustering for Human Activity Discovery

Wee-Hong Ong[*a)] Non-member, Leon Palafox[**] Non-member
Takafumi Koseki[*] Member

(Manuscript received Dec. 30, 2013, revised May 7, 2014)

One of the challenges in human activity recognition is the ability for an intelligent system to discover the activity models by itself. In this paper, we propose an incremental approach to discover human activities from unlabeled data using K-means. The approach does not require prior specification of the number of clusters, or k-value, and has the ability to reject random movements or noise. Simple algorithm is used making the approach easy to implement without requiring any prior knowledge in the data. We evaluated the effectiveness of the approach and the results show more than 30% improvement in precision and 19% improvement in recall when compared to the results obtained using a non-incremental approach with cluster validity index. The achievement in human activity discovery will enable the wide adoption of human activity recognition technologies in the natural human living environment where labeled data are not available.

Keywords: human activity detection; human activity discovery; unsupervised learning; clustering; RGB-D sensor

1. Introduction

Human activity recognition (HAR) is the ability of an intelligent system to understand what people are doing. Significant achievements in HAR technologies [(1)(2)] have been made in applications where the systems can be provided with models learned in a supervised manner. However, we have not seen wide adoption of HAR technologies in our homes, where such technologies will play an important role in assisted living.

One of the reasons that hinders the use of HAR in our homes is the wide variation in the nature of human daily activities. It is difficult to come up with a definite set of human activity models that will suit everyone. Another reason is the lack of labeled data in our homes. In such environment, a supervised approach in learning activity models cannot be directly applicable. Ideally we want the system to learn the activity models by itself. This is the aim of human activity discovery.

Human activity discovery is the autonomous learning of activity models in an unsupervised fashion. Given the increased uncertainty in learning from unlabeled data, human activity discovery is less developed than the supervised approach in learning activity models. Clustering algorithms are widely used in solving unsupervised learning problems. An important issue in all clustering algorithms is the need to determine the number of clusters in the data. There have been various approaches developed to address the issue and the suitability of each approach depends on the nature of the data.

a) Correspondence to: Wee-Hong Ong. E-mail: owh@koseki.t.u-tokyo.ac.jp
[*] Graduate School of Engineering, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[**] Department of Radiology, University of California
Los Angeles, CA 90095-7437

In this paper, we propose an incremental approach of clustering for human activity discovery. This approach uses a simple algorithm and is easy to implement. It does not require prior specification of the number of clusters. Unlike methods that attempt to "estimate an optimal number of clusters", our approach attempts to "find good clusters" from the data. The approach is designed to reject noise, i.e. random movements in human activity discovery.

The contributions of this paper are twofold. First, a novel approach to incrementally "find good clusters" is proposed. The approach is generalized to perform clustering of noisy data without prior specification of the number of clusters. It is applicable in other application domains besides human activity discovery. Second, we demonstrate the effectiveness of the approach in human activity discovery. Given that each individual person may perform an activity in different manner, the algorithm is expected to discover activity models for each individual separately. For this reason, the experiment is conducted on data from individual subject. The development of such algorithm will enable the wide adoption of HAR technologies in the regular househould settings and facilitate assisted living.

The remaining of this paper is organized as follows: Section 2 describes related current works on human activity discovery. In Section 3, we describe the existing non-incremental approach we have used to compare with our proposed approach. In Section 4, we describe our proposed incremental approach of clustering for the purpose of human activity discovery. In Section 5, we describe the data used and the experiments conducted to evaluate the proposed approach. We discuss the results in Section 6 and conclude the findings in Section 7.

2. Related Works

Owing to the increased difficulty in dealing with unlabeled

© 2014 The Institute of Electrical Engineers of Japan.

data, there have been significantly less works in the unsupervised approach in human activity recognition than those using the supervised approach. A number of in-depth surveys and reviews [(1)–(3)] on human activity recognition (HAR) technologies have been published. From these articles, we observe that HAR has been extensively studied in the field of computer vision. We also note the works reported in these surveys are based on the supervised approach.

In the domain of human activity discovery, i.e. the unsupervised approach, a good number of the works are sensor-based [(4)–(7)]. These works predict activities from the signals obtained from sensors embedded in the environment, attached to the user's body or tagged on objects. A sensor-based approach cannot be conveniently implemented in the existing home setting. Our interest is focused on vision-based human activity discovery.

Two notable works in the vision-based unsupervised approach for human activity recognition are that of Niebles et al. [(8)] and Cui et al. [(9)]. In both works, the number of activities was specified a priori. In unsupervised learning, clustering in particular, automatically determining the number of clusters has been one of the most difficult problems [(10)]. The typical way to estimate the number of clusters, or k-value, is to perform the clustering for the range of $k = 2$ to n and evaluate the clustering results at each k-value, where n is the number of data points in the dataset. The k-value that gives the best clustering result based the chosen criteria is taken as the optimal k-value.

There have been various criterion [(11)] used to evaluate the clustering outcome. They are referred as cluster validity indices. The approach using the cluster validity indices to determine the number of clusters are however not taking an incremental approach. They lack the ability to deal with data that have significant noise, such as the random movements in the case of human activity recognition. In our earlier work [(12)], we investigated the use of cluster validity indices for human activity discovery. We evaluated the effectiveness of five indices [(13)–(17)]. Out of the five indices, the most effective index to determine k in our dataset was the Hartigan index. While we had proposed to use threshold with cluster homogeneity measures to reject random movements, it was found difficult that the process also significantly rejected clusters of valid activities.

In this paper, we propose an incremental approach that can automatically discover clusters from noisy data without the specification of k-value a priori. The approach avoided the need to use threshold in determining the clusters to be rejected. The approach is also more computational efficient when compared to the non-incremental approach in determining the number of clusters. Unlike the non-incremental approach, the proposed incremental approach looks for "good clusters" rather than finds an "optimal k-value".

We describe our approach as incremental due to its behavior that incrementally finds new clusters from modified dataset. We have avoided using the term "incremental clustering", which generally refers to dealing with dynamic data, i.e. on-line mode. Our incremental approach can be used in batch mode, as demonstrated in this paper, as well as on-line mode. In on-line mode, new data points can be added to the modified dataset. This will require a mean to identify if the new data points can be assigned to existing clusters. This, however, is beyond the scope of this paper.

3. Discovery of Activities using Cluster Validity Index

In this section, we describe the existing non-incremental approach to determine the number of clusters that we have used to compare with the proposed incremental approach.

As discussed in Section 2, we had found that the Hartigan index [(15)] is most suitable for our dataset when estimating the number of clusters using cluster validity indices. We will compare the proposed incremental approach with the use of Hartigan index to discover clusters.

To use a cluster validity index to estimate a suitable k-value, we first run clustering for a range of k-values, i.e. from $k = 2$ to $k = k_{max}$. For each k-value, we compute the cluster validity index. In the case of Hartigan (Ha) index, the index is calculated as below:

$$Ha(k) = \left(\frac{trace\,(SSW\,(k))}{trace\,(SSW\,(k+1))} - 1\right)(n-k-1) \cdots (1)$$

$$SSW\,(k) = \sum_{j=1}^{k}\sum_{x_i \in C_j}\left(x_i^{(j)} - c_j\right)\left(x_i^{(j)} - c_j\right)^T \cdots\cdots\cdots (2)$$

where n is the number of data points in the dataset, and $SSW(k)$ is the (sum-of-square) within-cluster scatter matrix, $x_i^{(j)}$ is ith data point in Cluster j and c_j is the centroid of Cluster j, i.e., C_j. Data points are column vectors.

The objective of using Hartigan index to determine k-value is to take the k-value at which Ha crosses a threshold. $Ha \leq 10$ is typically used and we have used this value in our work reported in this paper.

4. Proposed Incremental Discovery of Activities

In this section, we describe our proposed incremental approach to discover human activities from unlabeled observations. This approach is motivated by the way children learn about their environment in an incremental manner. For example, while there are many activities going on, children do not learn all of them at once. They learn them over time.

4.1 Algorithm Fig. 1 gives the psuedo code of our proposed incremental approach to discover clusters of activi-

```
inc_discovery( X, MinPt ):
    n = size of X;
    k = sqrt( n );
    Set the value of MinPt;
    While k ≥ 2 do:
        Cluster X into k clusters;
        Evaluate the homogeneity of each cluster that
        has at MinPt members;
        Collect the most homogeneous cluster, C*;
        Remove the members of C* from X, i.e.
        X = X − C*;
        Compute new k = sqrt( n* ) where n* is the size
        of trimmed X;
```

Fig. 1. Incremental Discovery of Activities

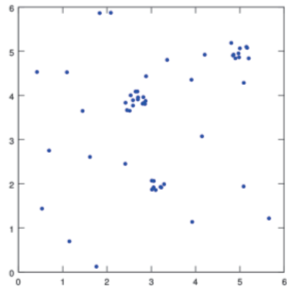
ties. It takes the dataset X and the minimum point per cluster *MinPt* parameter as input. It returns a set of clusters that the algorithm has discovered. Not all data points in the dataset will be clustered.

The basic idea is to start with a sufficiently high value of k, k_{max}. k_{max} can be up to n, the number of data points in the dataset. However, it helps to restrict the computation time by setting a reasonable value for k_{max}. There is no concrete guideline for the choice of k_{max}, however many researchers had referred to Mardia et al. [18] as stating the rule of thumb for setting $k = \sqrt{\frac{n}{2}}$. We have chosen $k_{max} = \sqrt{n}$, which includes the value of k suggested by the said rule of thumb.

For each k-value, clustering is performed on the dataset and the homogeneity of each of the clusters is evaluated. The choice of clustering algorithm and cluster homogeneity measure will be described in the following subsections. The most homogeneous cluster is collected and the members removed from the dataset. A new value of k is then computed from the new population $n*$ of the pruned dataset.

The process gradually lower the value of k until $k = 2$. To ensure the clusters have sufficient observations to model the activities, a minimum points or minimum membership parameter ($MinPt$) can be imposed.

Fig. 3 shows the comparison between the non-incremental approach using cluster validity index, e.g. Hartigan Index, and our incremental approach. The figure shows the outcome of each iteration of the algorithm. In each iteration, the left hand side shows the outcome from the non-incremental approach, i.e. using cluster validity index, while the right hand side is the outcome from the incremental approach.



Dataset $n = 53, k_{max} = \sqrt{53} \cong 7$

Fig. 2. Fictitious Dataset for Illustration of Incremental Approach in Clustering in Fig. 3.

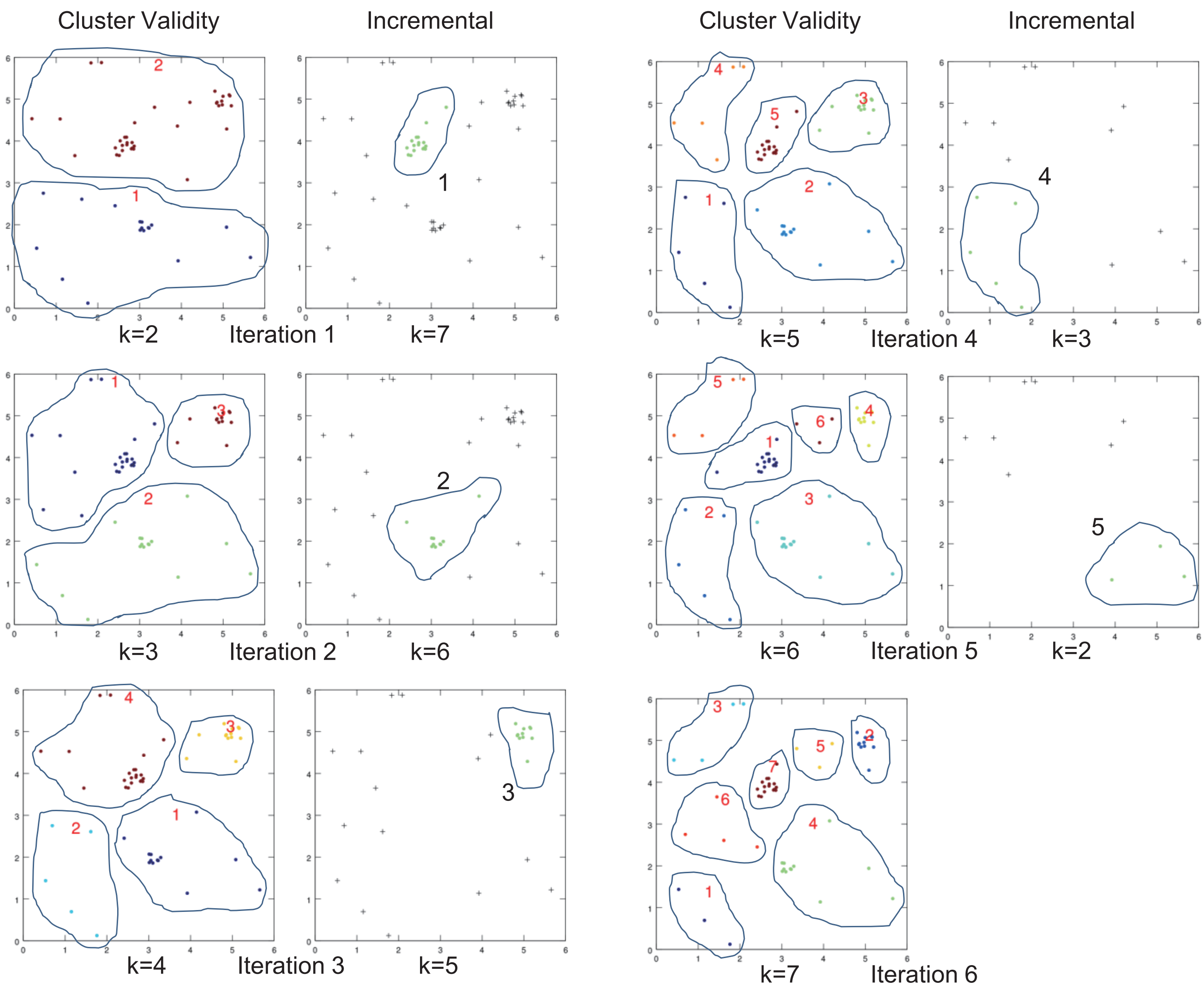


Fig. 3. Illustration of Effectiveness of Incremental Approach in Comparison with Non-Incremental Approach Using Cluster Validity Index. The dataset is shown in Fig. 2.

A factitious dataset with 53 data points was used. There are apparently three clusters surrounded with noise. Given $n = 53$, k_{max} is approximately 7. The incremental approach started with highest $k = k_{max}$, while the non-incremental approach started with lowest $k = 2$. The non-incremental approach iterated through all values of k from 2 to 7 and computed the cluster validity index in each iteration. An "optimal" k-value was determined based on the cluster validity index value and the resulting clusters from the "optimal" k-value were taken as the outcome. All data points were assigned to a cluster, i.e. there was no noise rejection. This approach took six iterations to complete.

On the other hand, with the incremental approach, the clusters are obtained from the "best" cluster of each run of the clustering algorithm. It found one cluster in each iteration. The discovered cluster is trimmed from the dataset after each run and the distribution of the data is perturbed. This trimming can potentially lead to the discovery of clusters not found in the large dataset. By fifth iteration, it has found five clusters and rejected some noise points. The objective of the approach is not to estimate an optimal number of clusters, but instead it attempts to find good clusters from the data. This is useful when we have noisy data.

The trimming in each iteration also reduces the overall computation time when compared to the non-incremental approach. In the non-incremental approach using cluster validity index, the algorithm search through the range of k-value, from $k = 2$ to $k = k_{max}$, with the same number of data points, n, at all k-values. On the other hand, the incremental approach keeps reducing the number of data points, n, in each iteration. It also lowers the k-value based on the reduced number of data points. It does not necessary go through all values of k within the range of $k = 2$ to $k = k_{max}$. The algorithm starts from the highest k-value allowing it to find a cluster right from the first iteration.

4.2 K-means Clustering For the purpose of evaluating the effectiveness of the incremental approach, we use K-means [19] clustering. It is one of the simplest unsupervised learning algorithms. Given the required number of clusters, k, K-means group the points (observations) in the dataset by minimizing the distance from each data point to a cluster center (centroid). The goal of K-means is to minimize the sum of the squared error over all k clusters as given in Eq. (3).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \cdots\cdots\cdots\cdots (3)$$

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 minimize the effect of random initialization in K-means, it's common practice to perform multiple runs of K-means for a given k-value and select the result from the run that gives the lowest cost function value.

4.3 Cluster Homogeneity Measure To assess the homogeneity, i.e., cohesiveness and compactness, of individual cluster and rank them accordingly, we define two measures: the intra-cluster mean variance ($\bar{\sigma}^2$) and mean joint probability density function ($\overline{P}_f$).

Low value of variance $\bar{\sigma}^2$ indicates compactness of the cluster.

$$mean\ variance\ \overline{\sigma}^2\left(C_j\right) = mean\left(var\left(C_j\right)\right) \cdots\cdots (4)$$

$$var\left(C_j\right) = \frac{1}{n_j - 1} \sum_{x_i \in C_j} \left(x_i^{(j)} - \overline{x}^{(j)}\right)^{\wedge 2} \cdots\cdots\cdots (5)$$

$$centroid\ \overline{x}^{(j)} = mean\left(x_i \in C_j\right) \cdots\cdots\cdots\cdots (6)$$

where $x_i^{(j)} = [x_{i1} \cdots x_{ip}]$ is a data point in Cluster C_j with dimension p, n_j is the number of points in Cluster C_j, $\bar{x}^{(j)}$ (dimension p) is the mean of all points in Cluster C_j, $\wedge 2$ is element-wise square.

The main difference between the two criteria is that the variance evaluates the distance of the data points to the centroid, while the probability density function evaluates if the data points are distributed in the form of normal distribution.

The joint probability density function assumes that observations of a non random activity should be normally distributed within its cluster around the cluster centroid with the standard deviation of the cluster. High value of mean joint probability density function $\overline{P}_f$ indicates good cohesion of the cluster based on the assumption of normal distribution. Logarithm in Eq. (8) is used to compress the range of the values.

In both cases, the "mean" is taken across the dimensions of the data. Since the data points are multidimensional, the average (i.e. mean) across all dimensions gives a single value measure.

$$mean\ joint\ probability \cdots\cdots\cdots\cdots (7)$$
$$\overline{P}_f\left(C_j\right) = mean\left(P\left(C_j\right)\right)$$

$$P\left(C_j\right) = \sum_{f=1}^{p} ln\left(pdf\left(x_{if}^{(j)}\right)\right) \cdots\cdots\cdots\cdots (8)$$

$$pdf\left(x_{if}^{(j)}\right) = \frac{1}{\sigma_f^{(j)}\sqrt{2\pi}} e^{-\frac{\left(x_{if}^{(j)} - \mu_f^{(j)}\right)^2}{\left(2\sigma_f^{(j)}\right)^2}} \cdots\cdots\cdots\cdots (9)$$

where $x_{if}^{(j)}$ is the fth dimension of $x_i^{(j)}$, $\mu_f^{(j)}$ is the mean value of the fth dimension of all points in Cluster C_j, $\sigma_f^{(j)}$ is the standard deviation value of the fth dimension of all points in Cluster C_j.

5. Data and Experiment

5.1 Data The input data to our algorithm are features extracted from the 3-D coordinates of the joint position of the human skeleton as shown in Fig. 4. The coordinates are obtained from the OpenNI SDK [20] of the commercial RGB-D (RGB-Depth) sensor, Microsoft Kinect [21]. In this paper, we evaluated the proposed algorithm on our dataset and the activities from a third party dataset Cornell Activity Dataset CAD-60 [22].

There are five datasets being considered. Each of the dataset is a set of activities performed by a single subject. Table 1 gives the list of activities by four of the subjects, Person 1 to Person 4. The nine activities (A1 to A9) are from the CAD-60 dataset. Table 2 gives the list of activities by the fifth subject, Person 5. The sixteen activities (B1 to B16) are from our dataset. In all the dataset, random movements (RA) are included. Each dataset contains 56 instances

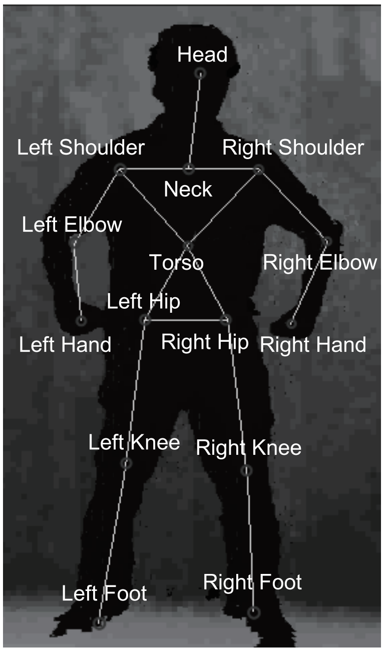


Fig. 4. Human skeleton composed from fifteen (15) joints.

Table 1. List of activities by Person 1 to 4

1.	A1	Brushing teeth
2.	A2	Cooking (chopping)
3.	A3	Cooking (stirring)
4.	A4	Relaxing on couch
5.	A5	Still (standing)
6.	A6	Talking on couch (sitting)
7.	A7	Talking on the phone
8.	A8	Working on computer
9.	A9	Writing on whiteboard
10.	RA	Random

Table 2. List of activities by Person 5

1.	B1	Bowing	10.	B10	Walking
2.	B2	Drinking (left)	11.	B11	Wave bye (left)
3.	B3	Drinking (right)	12.	B12	Wave bye (right)
4.	B4	Sit	13.	B13	Wave come (left)
5.	B5	Sit down	14.	B14	Wave come (right)
6.	B6	Stand	15.	B15	Wave go (left)
7.	B7	Stand up	16.	B16	Wave go (right)
8.	B8	Talking on phone (left)	17.	RA	Random
9.	B9	Talking on phone (right)			

of each activity and 112 instances of random movements. Therefore, each dataset for Person 1 to Person 4 comprises of $9 \times 56 + 112 = 616$ activity instances, whereas the dataset for Person 5 comprises of $16 \times 56 + 112 = 1008$ activity instances.

Each activity instance or observation has a fix duration of two seconds with reduced frame rate giving a total of fifteen frames in two seconds. In each frame, there are forty-two features obtained from local vectors formed from the 3-D coordinates of the fifteen joints. The total number of features is therefore $15 \times 42 = 630$ features per activity instance. The details of the dataset and feature extraction are given in our earlier paper [23], and the paper by the authors of CAD-60 [22].

5.2 Experiment We carried out five experiments to evaluate the activity discovery on the dataset of each subject.

- H: clustering using Hartigan index (this is the same experiment conducted in our earlier work [12]),
- Var: clustering using the proposed incremental approach using mean variance as the homogeneity measure without setting the minimum point parameter, i.e. $MinPt = 0$,
- Pdf: clustering using the proposed incremental approach using mean joint probability as the homogeneity measure without setting the minimum point parameter, i.e. $MinPt = 0$,
- Var-25: clustering using the proposed incremental approach using mean variance as the homogeneity measure with the minimum point parameter set to 25, i.e. $MinPt = 25$,
- Pdf-25: clustering using the proposed incremental approach using mean joint probability as the homogeneity measure with the minimum point parameter set to 25, i.e. $MinPt = 25$.

The algorithm was given all the data points, all activities mixed up, without label from the dataset of each subject. In each experiment, we evaluated how well the clustering discovered the activities and reject the random movements (RA) in each dataset (Person 1 to Person 5). For each clustering result, we calculated the precision and recall. Given that the result of K-means is sensitive to the random initialization, we performed five runs of each experiment to obtain an average evaluation of the clustering performance.

6. Results and Discussions

Fig. 5 shows the average precision of the five experiments, H, Var, Pdf, Var-25 and Pdf-25 as described in Section 5, on the dataset of each subject. The right most set of columns are the overall average across five subjects. For each subject, the values given are the average across all activities, i.e. nine activities for Person 1 to Person 4, and sixteen activities for Person 5.

We observe a few things from Fig. 5:

(1) For all subjects, the precision achieved using Hartigan index (H), i.e. the non-incremental approach, is the lowest and it is significantly lower than the precision achieved by the incremental approaches (Var, Pdf, Var-25 and Pdf-25).

(2) For all subjects, the highest precision is that achieved by the incremental approach using mean variance as the cluster homogeneity measure and with minimum point parameter $MinPt = 25$ (Var-25).

(3) For all subjects, the precision achieved by the incremental approach using mean joint probability (Pdf and Pdf-25) as the cluster homogeneity measure is lower than that achieved using mean variance (Var and Var-25) as the homogeneity measure given the same minimum point parameter. The use of probability density function makes assumption on the distribution of the data points in the cluster. The use of variance does not make such assumption and only evaluates the closeness of the data points to their cluster centroid.

(4) Setting a value for the minimum point parameter improves precision in both cases of cluster homogeneity measures.

(5) The average precision across all five subjects for the

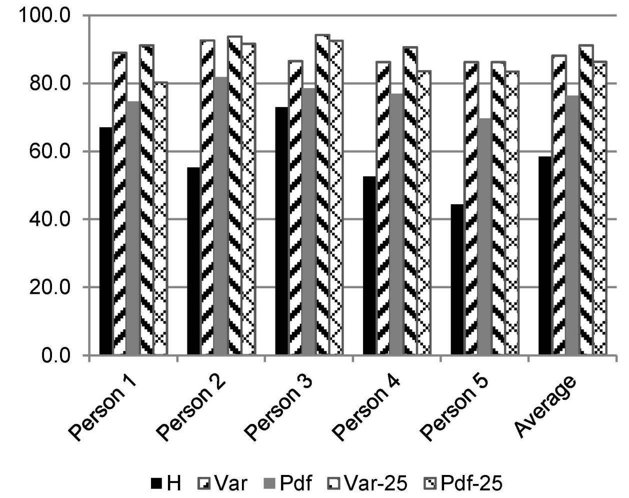


Fig. 5. Average precision of different approaches for the data of each subject.

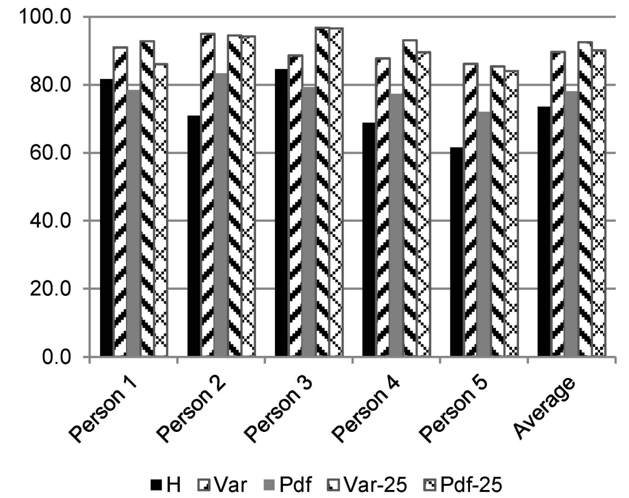


Fig. 6. Average recall of different approaches for the data of each subject.

result using Hartigan index (H) is 58.4%.

(6) The average precision across all five subjects for the result using mean variance and $MinPt = 25$ is 91.2%.

(7) The incremental approach with mean variance and $MinPt = 25$ achieved an average 32.8% improvement in the precision over the non-incremental approach using Hartigan index.

Fig. 6 shows the average recall of the five experiments. We observe similar trend as in the case of precision. The overall average recall for the result using Hartigan index (H) is 73.5%, while the result using mean variance and $MinPt = 25$ achieved an average 92.5% recall. The incremental approach improves the recall by an average of 19%.

The incremental approach also showed great potential to reject random movements. Table 3 shows the complete result in one run of the incremental approach using mean variance and $MinPt = 25$ for the data of Person 4. The first row, i, is the iteration of the incremental approach as given in Fig. 1. The second row is the k-value used in each iteration. The remaining rows are the activities as listed in Table 1. In the first iteration, $i = 1$, the $k = \sqrt{616} \simeq 24$. In this iteration, one

Table 3. One complete run of the incremental discovery algorithm using mean variance and $MinPt = 25$ for Person 4

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
k	24	23	22	21	20	19	18	17	16	15	13	11	7	5	3	U
A1	0	0	38	0	1	0	0	0	0	0	0	17	0	0	3	0
A2	0	0	0	1	0	0	0	0	0	55	0	0	0	0	0	0
A3	0	0	0	36	0	0	0	0	0	0	0	0	0	0	20	0
A4	0	30	0	0	0	0	26	0	0	0	0	0	0	0	20	0
A5	0	0	0	0	0	0	0	56	0	0	0	0	0	0	0	0
A6	0	0	0	0	0	0	0	0	33	0	0	0	0	20	3	0
A7	0	0	0	0	56	0	0	0	0	0	0	0	0	0	3	0
A8	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A9	0	0	0	0	0	39	0	0	0	0	17	0	0	0	0	0
RA	0	0	0	1	0	0	0	16	0	0	8	9	25	15	13	25

cluster with lowest mean variance was found. This cluster contains all the 56 instances of Activity 8 (Working on computer). These 56 instances were trimmed from the dataset, leaving $616 - 56 = 560$ instances. In the next iteration, $i = 2$, $k = \sqrt{560} \simeq 23$. One cluster was found and it contains 30 instances of Activity 4 (Relaxing on couch). The incremental process continued until no further cluster can be obtained. A total of fifteen clusters were found, i.e. up to $i = 15$. Twenty five instances of the random movements (RA) were not assigned to any cluster as collected in the right most column, U. We observe a few things from Table 3:

(1) The clusters obtained in early iterations are homogeneous, i.e. contains only one activity.

(2) The instances of random movements (RA) are only clustered towards the end of the incremental process.

(3) One exception to the above observation is at eighth iteration, $i = 8$. At eighth iteration, a significant number of random movements (RA) were clustered with Activity 5 (Still). Further investigation into the dataset revealed that these random movements (RA) happened to be sampled from instances when the subject was in still state.

Given the above observations, the incremental approach can theoretically discover all the nine activities performed by Person 4 by the time it reached tenth iteration, $i = 10$. For Activity 4 (Relaxing on couch) and Activity 6 (Talking on couch), there would be twenty instances of each of the activity being left out. However, there are clusters with significant number of instances to model these two activities: the clusters at $i = 7$ and $i = 9$.

There is a reason we choose to elaborate the detailed results for Person 4. The results using Hartigan index (H) on dataset of Person 4, among the four subjects in CAD-60 dataset, are the worst among all subjects, as indicated by the low precision and recall in Fig. 5 and Fig. 6. In our earlier work [12], the results show that the non-incremental approach using Hartigan index could only discover one activity for Person 4, whereas our proposed approach could discover all activities.

Using the Hartigan Index, or any other cluster validity index, requires that the clustering be performed on all the training data and evaluation is performed on all clusters. This implies the selected value of k did not go through the process of evaluating each cluster. On the other hand, in the incremental algorithm, the k value is irrelevant. The algorithm evaluates each cluster and selects good cluster in each iteration. For this reason, the incremental algorithm has showed superior performance when compared with the non-incremental approach using cluster validity.

7. Conclusions

We have proposed an incremental approach of clustering to address the problem of human activity discovery. The approach was implemented with simple clustering algorithm, i.e. K-means. The approach was evaluated on two different datasets comprising of twenty five activities and five subjects in total. The results show that the approach achieved an overall average precision of 91.2% and recall of 92.5% across the dataset of the five subjects. This represents an improvement of over 30% in precision and 19% in recall when compared to the non-incremental approach using cluster validity index.

The results suggest the use of mean variance as the cluster homogeneity measure. The results show that the proposed approach can potentially reject random movements or noise. The incremental approach also reduces computation time by trimming the dataset in each iteration and recalculating k-value based on the population of the trimmed dataset.

While the development of the proposed approach was motivated by the purpose of human activity discovery, it is a generalized clustering approach. It can be used in other applications requiring an unsupervised learning without prior knowledge of the data and with significantly noisy data.

References

(1) J.K. Aggarwal and M.S. Ryoo: Human activity analysis A review, ACM Computing Surveys (2011)
(2) P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea: "Machine recognition of human activities A survey", IEEE Trans. on Circuits and Systems for Video Technology (2008)
(3) R. Poppe: "A survey on vision-based human action recognition", Image and vision computing (2010)
(4) T. Huynh, M. Fritz, and B. Schiele: "Discovery of activity patterns using topic models", 10th International Conference on Ubiquitous computing (2008)
(5) M. Stikic, D. Larlus, S. Ebert, and B. Schele: "Weakly supervised recognition of daily life activities with wearable sensors", IEEE Trans. on Pattern Analysis and Machine Intelligence (2011)
(6) R. Hammid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C.L. Isbell: "Unsupervised activity discovery and characterization from event-streams", arXiv preprint arXiv:1207.1381 (2012)
(7) B. Chikhaoui, S. Wang, and H. Pigot: "ADR-SPLDA Activity discovery and recognition by combining sequential patterns and latent Dirichlet allocation", Pervasive and Mobile Computing (2012)
(8) J.C. Niebles, H. Wang, and L. Fei-Fei: "Unsupervised learning of human action categories using spatial-temporal words", International Journal of Computer Vision (2008)
(9) P. Cui, F. Wang, L.F. Sun, and J.W. Zhang: "A Matrix-Based Approach to Unsupervised Human Action Categorization", IEEE Trans. on Multimedia (2012)
(10) A.K. Jain: "Data clustering: 50 years beyond K-means", Pattern Recognition Letters (2010)
(11) U. Maulik and S. Bandyopadhyay: "Performance evaluation of some clustering algorithms and validity indices", IEEE Trans. on Pattern Analysis and Machine Intelligence (2002)
(12) W.H. Ong, T. Koseki, and L. Palafox: "Investigation of Cluster Validity Indices for Unsupervised Human Activity Discovery", 2013 International Conference on Artificial Intelligence (2013)
(13) T. Caliński and J. Harabasz: "Communications in Statistics-theory and Methods", Taylor & Francis (1974)
(14) D.L. Davies and D.W. Bouldin: "A cluster separation measure", IEEE Trans. on Pattern Analysis and Machine Intelligence (1979)
(15) J.A. Hartigan: Clustering algorithms, John Wiley & Sons, Inc. (1975)
(16) W.J. Krzanowski and Y. Lai: A criterion for determining the number of groups in a data set using sum-of-squares clustering, Biometrics (1988)
(17) P.J. Rousseeuw: Silhouettes a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics (1987)
(18) K.V. Mardia, J.T. Kent, and J.M. Bibby: Multivariate analysis, Academic press (1980)
(19) J. MacQueen: Some methods for classification and analysis of multivariate observations, Fifth Berkeley Symposium on Mathematical Statistics and Probability (1967)
(20) OpenNI: OpenNI—The standard framework for 3D sensing, OpenNI (2010)http://openni.org/ Accessed: 2012-04-30
(21) Z. Zhang: Microsoft Kinect Sensor and Its Effect, IEEE MultiMedia (2012)
(22) J. Sung, C. Ponce, B. Selman, and A. Saxena: Human Activity Detection from RGBD Images, AAAI workshop on Plan Activity and Intent Recognition (2011)
(23) W.H. Ong, T. Koseki, and L. Palafox: Unsupervised Human Activity Detection with Skeleton Data From RGB-D Sensor, 2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks (2013)

Wee-Hong Ong (Non-member) received the B.Eng. in Communication and Control Engineering from the University of Manchester, Institute of Science and Technology in 1997. He received the M.Sc. in Computing Science from the Imperial College London in 2004. He is currently a PhD candidate in the Department of Electrical Engineering and Information Systems in the University of Tokyo, Japan. His research interests are in personal robotic and ambient intelligence. His PhD work is focused on realizing human activity recognition in natural human living environment.

Leon Palafox (Non-member) received the B.Sc. degree in Electronic Engineering from the National Autonomous University of Mexico (UNAM). He received the M.Sc. degree and PhD degree in Electronic Engineering from the Graduate School of Engineering, University of Tokyo, Tokyo, Japan. He is currently a Postdoctoral Fellow in the University of California, Los Angeles. His research interests include neuroscience, machine learning and bioinformatics.

Takafumi Koseki (Member) received the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan in 1992. He is currently a Professor in the Department of Electrical Engineering and Information Systems, School of Engineering, The University of Tokyo. His current research interests include applicati on s of electrical engineering to public transport systems, especially to linear drives, and analysis and control of traction systems. Dr. Koseki is a member of the Institute of Electrical Engineers of Japan, the Institute of Electric and Electronic Engineers, the Japan Society of Mechanical Engineering, the Japan Society of Applied Electromagnetics and Mechanics, the Japan Society of Precision Engineering, and Japan Railway Electrical Engineering Association.