A Novel Skeleton-Based Human Activity Discovery Technique Using Particle Swarm Optimization with Gaussian Mutation

Parham Hadikhani\textsuperscript{a,\ast}, Daphne Teck Ching Lai\textsuperscript{a,\textbullet}, Wee-Hong Ong\textsuperscript{a,\textbullet}

\textsuperscript{a}School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link, Brunei
\textsuperscript{b}Institute of Applied Data Analytics, Universiti Brunei Darussalam, Jalan Tungku Link, Brunei

Abstract

Human activity discovery aims to distinguish the activities performed by humans, without any prior information of what defines each activity. Most methods presented in human activity recognition are supervised, where there are labeled inputs to train the system. In reality, it is difficult to label data because of its huge volume and the variety of activities performed by humans. In this paper, a novel unsupervised approach is proposed to perform human activity discovery in 3D skeleton sequences. First, important frames are selected based on kinetic energy. Next, the displacement of joints, set of statistical, angles, and orientation features are extracted to represent the activities information. Since not all extracted features have useful information, the dimension of features is reduced using PCA. Most human activity discovery proposed are not fully unsupervised. They use pre-segmented videos before categorizing activities. To deal with this, we used the fragmented sliding time window method to segment the time series of activities with some overlapping. Then, activities are discovered by a novel hybrid particle swarm optimization with a Gaussian mutation algorithm to avoid getting stuck in the local optimum. Finally, k-means is applied to the outcome centroids to overcome the slow rate of PSO. Experiments on three datasets have been presented and the results show the proposed method has superior perfor-

\textsuperscript{\ast}Corresponding author.

Email addresses: 20h8561@ubd.edu.bn (Parham Hadikhani), daphne.lai@ubd.edu.bn (Daphne Teck Ching Lai), weehong.ong@ubd.edu.bn (Wee-Hong Ong)
mance in discovering activities in all evaluation parameters compared to the other state-of-the-art methods and has increased accuracy of at least 4% on average. The code is available here: [Human-Activity-Discovery-HPGMK](#)

**Keywords:** Human activity discovery, Unsupervised learning, Feature extraction, Skeleton sequence, Dimension reduction, Particle swarm optimization

### 1. Introduction

Research in Human Activity Recognition (HAR) has attracted much attention due to its applications in fields such as human-computer interaction, intelligent transportation systems, and monitoring applications [1]. Despite its spectacular progress, obtaining high accuracy is a serious challenge. The purpose of activity recognition is to automatically identify actions and activities humans perform in different environments. The input to a HAR system is a sequence of frames of a person performing different movements. The output is a set of labels that represent the actions taken or activities in those movements. Many existing works use visual data as input. But such data have considerable complexity detrimental to the HAR systems performance. These complexities include cluttered background, changes in brightness and points of view [2]. Using 3D skeleton data partially overcome these complexities [3, 4] and protect people’s privacy [5]. Each frame represented by 3D coordinates of the main body joints is appropriate for representing human actions [5] and can be easily obtained in real-time with low-cost depth sensors [6]. As shown in Fig. 1, there are at least 7 steps in vision-based HAR systems. Activities performed by a person are captured by vision sensors. The skeletal information comprising of joints coordinates are then extracted from captured videos, containing image sequences called frames. Meaningful features are then extracted for more accurate activity discovery. The system discovers activities by clustering them without prior knowledge of the activities. It starts the learning process to model each cluster of activity. The learned model is used to recognize future activities. Significant progress has been made in the supervised learning of activity models [7, 8], illustrated in block (f) and (g) of Fig. 1. Unlike learning and recognition steps, which rely on human-labelled training data to categorize activities and do not perform block (e). Human activity discovery (see block (e) of Fig. 1) is a HAR process where activities are categorized based on their similarities without
any knowledge of activity labels or any information that characterise an activity, making this step particularly challenging. In other words, activity discovery is like a child learning to move where there is no prior information to define a specific sequence of movements to mean a specific activity such as crawling or waving and so forth to the child learner. Using the ability to differentiate, they learn from unlabeled data and based on that training, form a model that can post-label new data. In human activity discovery, there is no known information or knowledge that the flow of a particular movement including its start to end is an activity such as picking up something. This means the input is a series of movements where there is no prior knowledge of which start and end points indicate an activity. In some existing work, the input data are segmented by activity [5]. Thus, the start and end point of the activities are already known, though the techniques to elucidate these activities may be unsupervised. In this paper, we focus on the less developed activity discovery steps, illustrated in block (d) to (e) of Fig. 1. To reduce computational load and increase accuracy [6], keyframe selection and PCA are used to remove redundant frames and features respectively. A new feature extraction methodology is applied to extract features from the most informative joints and bones which include joint displacement, joint orientation, and statistical time domain. A hybrid Particle Swarm Optimization (PSO) clustering technique is proposed to find activities. PSO can reach the optimal solution independent of initial population states. Although PSO is faster than evolutionary algorithms, swarm particles can get trapped in local optima. One reason is the particles converge to a specific point between the best global position and the best personal position. To address this, a new hybrid PSO with Gaussian mutation is proposed. Then, K-means is applied to the centroids obtained by PSO to refine their location and get the best possible solution. Our methodology performs activity discovery using unsegmented input data and the proposed techniques used are purely unsupervised, with no prior knowledge of what comprise the different activities except the number of clusters. The aim of this paper is to present a new and efficient methodology to cluster different activities using un-segmented skeleton data with no prior information about the activities. The input data holds no information regarding the start and end of each of the activities. The main contributions are: (1) A method based on kinetic energy is applied to select the most important frames and remove redundant information (with the assistance of kinetic energy); (2) a sampling method is employed with assistance of overlapping sliding windows to segment frames
Figure 1: Overview of HAR system: (a) performed activities are (b) captured by a Kinect sensor. (c) After that, pose of humans are estimated by extracting joints. (d) To make raw data more usable, their salient and defining features are identified. (e) Based on the similarities and differences, activities are discovered. (f) Afterwards, the system begins to learn from the discovered activities and (g) finally human activities are recognized.
at specified intervals; (3) a novel feature extraction technique for extracting three types of features comprising spatial and temporal displacement, mean and standard deviation differences, and orientation and angle features of selected joints; and (4) a new hybrid algorithm to discover and obtain the most optimal clusters of human activities, using the Gaussian Mutation to enhance the exploration ability of the PSO and using K-means to refine the obtained cluster centers from the PSO. In this paper, the background and related methods are discussed in section 2. The methodology is described in section 3. We present the evaluation of the proposed approach, comparing with state-of-the-art techniques in Section 4, and finally, the conclusion in Section 5.

2. Related works

Many studies in HAR used supervised approaches. They focus on feature representation and activities classification [10, 11, 12, 13, 14, 15, 16]. These approaches require activity labels to learn. These labels are given by humans during data preparation. On the other hand, human activity discovery automatically recognizes human activity in a fully unsupervised way and the challenge is dealing with unlabeled data. Majority of existing methods were developed for sensor-based [17, 18] and RGB video data [19, 20, 21, 22, 23]. The challenges of sensor-based approach are difficult to implement in the environment and take long time to install [24]. Furthermore, it is impractical for people to wear sensors everywhere. With RGB videos, the problems faced are millions of pixel values, illumination variations, viewpoint changes, and cluttered backgrounds [6]. It is not the objective of this work to investigate in such data.

In this work, we use 3D skeleton-based data as it does not have the problems of the other two data types. Such data show three-dimensional coordinates of main body joints obtained in each frame. Skeletal information is valuable to show human movements and activities [25]. Here, we review the 3D skeleton-based approaches for human activity discovery. The first work proposed used k-means to cluster unlabeled activities into groups and estimate the number of groups using cluster validity indices [26]. A feature extraction approach using a range of human movements as features was presented in [27]. An incremental clustering approach to deal with undefined number of activities and noisy data was proposed in [28]. In [24], an autonomous learning technique was developed based on mixture of
Gaussian hidden Markov model with assistance of incremental k-means clustering. State-of-the-art approaches include Spectral Clustering (SC), Elastic Net Subspace Clustering (ENSC) and Sparse Subspace Clustering (SSC), which use covariance descriptor to solve HAR without labels [5]. They used an affinity matrix to find similarities and then applied spectral clustering. A time stamp pruning approach was used to remove redundant data. Although they achieved impressive results, the data used were already segmented by activity before applying clustering.

3. Proposed Human Activity Discovery

In this section, the proposed method is presented, as shown in Fig. 2. Keyframes are first extracted from input video based on the kinetic energy of frames, excluding frames that do not have useful information. This increases accuracy and clustering speed [9]. Three types of features are then extracted from the keyframes: spatial and temporal displacement, mean and standard deviation differences, and orientation and angle features. Next, PCA is applied to reduce the feature vectors, improving computation and ultimately the efficiency of the model. This is followed by having the frames sampled in specific time periods to segment the activity stream. Finally, the proposed clustering technique assigns each sample to an appropriate category of activities. Details of each part of the algorithm are described in the following subsections.

3.1. Keyframe Selection

Keyframe selection is a process to select frames reflecting the main activities in the video. The kinetic energy $E(f_i)$ of each frame $f_i$ is calculated [9] using Eq. (1), based on the displacement of joints over time. In this way, the movement of joint $j$ in frame $i$ and $i-1$ is calculated for all joints $j$. The sum of these movements for all joints is the current frames energy. Frames with local maxima and minima amount of kinetic energy compared to neighboring frames are considered as keyframes (see Fig. 3). These frames have good distinguishing features compared to other frames [29].

$$E(f_i) = \sum_{j=1}^{J} E(f_{ij}) = 1/2 \sum_{j=1}^{J} (f_{ij} - f_{ij-1})^2$$

(1)
Figure 2: Methodology of the proposed approach: First, (A) keyframes are selected from the video sequence by computing kinetic energy. Then, (B) features based on different aspects of skeleton including displacement, orientation, and statistical are extracted. (C) Principal components are then chosen by applying PCA on the features. Next, (D) overlapping time windows is used to segment a series of keyframes as activity samples. (E) Hybrid PSO clustering with Gaussian mutation operator is used to discover the groups of activities. (F) Eventually, K-means clustering is applied to the resultant cluster centers to refine the centroids.
3.2. Feature extraction

A new method is introduced to extract a set of statistical displacements, angles, and orientation features for encoding key aspects of activities. These important features are extracted from selected (informative) joints in the data to fully describe the shape and movement of humans. Selected joints include both left and right hand, foot, hip, shoulder, elbow and knee. In this way, we use information related to the position and movement of joints, the orientation and angle between a pair of bones and activity variations over time. The normalization procedure [30] is performed on all features.

3.2.1. Displacement features

Joint displacement-based features encode information on the position and motion of body joints, particularly displacement between joints of a pose and 3D position differences of skeleton joints across time periods [30].

- Spatial joint displacement is computed using pairwise Euclidean distances between joints \( P_i \) and \( P_j \) (\( i \neq j \)) in 3D space in the same frame, Eq. (2). The joint pairs used are both hands, hands and head, and hip and feet on both sides, giving 5 feature vectors per frame, (see Fig. 4a)).

\[
\text{PairwiseDistances} = \sqrt{\sum_{x,y,z} (P_i - P_j)^2} \tag{2}
\]
Temporal joint displacement is calculated based on two modes. $T_{cp}$ is the difference between each selected joint $P_i$ in current frame $P_i^c$ and previous frame $P_i^{c-1}$ (see Fig. 4(b)) to determine the small changes in joint movements over time (Eq. (3)). $T_{cn}$ is the difference between each selected joint of current frame and the frame of neutral pose $P_i^n$, illustrated in Fig. 4(c), to find general changes in joint movements as given in Eq. (4).

$$T_{cp} = P_i^c - P_i^{c-1}$$

$$T_{cn} = P_i^c - P_i^n$$

### 3.2.2. Statistical Features

The mean and standard deviation of time-domain features are used to express the transitions characteristics within an activity. These features describe how an activity changes over time, particularly in making good distinction between activities related to the arms and legs. Thus, statistical time-domain features encode information of variations across a collection of activity poses within the time-domain. These features are calculated by the difference of selected joint $P_i$ in current position from mean $P_{(i,mean)}$ and standard deviation $P_{(i, std)}$ of the selected joint coordinates within an activity sequence as given by Eq. (5) and (8) [30].
• Joint coordinate-mean difference

\[ P_{(i,\text{mean})} = P_i - P_{\text{mean}}, \quad P_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} P_i \]  

(5)

\( N \) is the number of frames.

• Joint coordinate-standard deviation difference

\[ P_{(i,\text{std})} = P_i - \sqrt{\frac{\sum_{i=1}^{N} (P_i - P_{\text{mean}})^2}{N}} \]  

(6)

3.2.3. Orientation Features

The three-dimensional coordinate system \( \{x,y,z \in R^3\} \) represents points as joints. \( x, y, \) and \( z \) denote the 3D coordinates of joints. Joints and bones can be described by the orthonormal vectors [31] as follow:

\[ J^f_i = x^f_i e_1 + y^f_i e_2 + z^f_i e_3 \]  

(7)

\[ B^f_{ij} = (x^f_i - x^f_j) e_1 + (y^f_i - y^f_j) e_2 + (z^f_i - z^f_j) e_3 \]  

(8)

where \( J^f_i \) is the \( i \)th skeleton joint at the \( f \)th frame and \( e_1, e_2, e_3 \) are orthonormal vectors. \( B^f_{ij} \) is the bone between two adjacent joints \( J^f_i \) and \( J^f_j \). Moreover, magnitude and direction of bones are represented by geometric product. This product for two bones \( a \) and \( b \) is the sum of internal (.) and external (\( \wedge \)) product:

\[ ab = a.b + a \wedge b \]  

(9)

Where the inner product is used to compute the length or angle between two bones \( a \) and \( b \). The outer product of two bones can be regarded as an oriented plane containing \( a \) and \( b \). The orientation and angles between bones features are obtained in the process described as follows.

• The rotation matrix is a transformation matrix that describes the rotation from a bone to another. Three angles are required to define the rotation matrix between two bones. The rotation angles are considered as orientation features. The elements of these features are the rotation of bones relative to the x, y and z axes (see Fig. 5(a)).

• The angle features consist of the angles between the bones of elbow-
Figure 5: (a) Illustration of the rotation between two bones A and B. $\alpha$, $\beta$, and $\gamma$ are the orientation of angles. (b) The angles of the selected body bones. The angles of elbow-wrist and shoulder-elbow in both sides and angles between the bones of hip-knee and knee-ankle in both sides are used to calculate angle feature.

wrist and shoulder-elbow in both sides and the angles between the bones of hip-knee and knee-ankle in both sides. These angles are highlighted in Fig. 5(b). The angle between the selected bones is calculated as

$$\theta = 180 \times \frac{\arctan^2 \left( \frac{\text{bone}_i \times \text{bone}_j}{\text{bone}_i \times \text{bone}_j} \right)}{\pi} + 180$$

where $\text{bone}_i$ and $\text{bone}_j$ are determined by Eq.(8).

3.3. Feature selection and Sampling

For fast clustering and complexity reduction, key features are extracted by PCA. Then, sliding windows are used to segment frames into time windows. Each window should have enough frames to represent an instance of the relevant activity. Each window comprises of 15 frames. The overlap of sliding windows increases performance because it makes sure that transition between activities are not missed [32]. The first 15 frames do not overlap while in other samples, their first frame starts from the last frame of the previous sample (see Fig. 6).
Figure 6: Illustration of sampling based on overlapping sliding windows. Each sample \((S_1, S_2, \ldots, S_n)\), except the first sample, starts with the last frame of the previous sample.

3.4. Proposed Clustering

PSO is a population-based optimization algorithm \[33\]. A population is made up of a number of particles and each particle represents a solution and moves according to its speed. The changes in velocity and position of the particles are calculated based on the following formula:

\[
x_i(t + 1) = x_i(t) + v_i(t)
\]

\[
v_i(t + 1) = w \times v_i + c_1 \times \text{rand}_1 \times (pbest_i(t) - x_i(t)) + c_2 \times \text{rand}_2 \times (gbest(t) - x_i(t))
\]

\[
w = w_{\text{max}} + t \times (w_{\text{max}} - w_{\text{min}}) / t_{\text{max}}
\]

\[
c_1(t + 1) = (c_{1\text{max}} - c_{1\text{min}}) \times t / t_{\text{max}} + c_{1\text{max}}
\]

\[
c_2(t + 1) = (c_{2\text{max}} - c_{2\text{min}}) \times t / t_{\text{max}} + c_{2\text{max}}
\]

In Eq. (11) and (12) \(x_i(t)\) and \(v_i(t)\) are the position and velocity of the particle \(i\) at time \(t\) respectively. \(pbest_i\) is the best position found by particle \(i\). \(gbest\) is the best position found in the population. \(w\) is the inertial weight defined by Eq. (13) and starts to decrease from \(w_{\text{max}}\). \(c_1\) and \(c_2\) are accelera-
tion coefficients expressed by Eq. (14) and (15). The $c_{1\text{max}}$, $c_{2\text{max}}$ and $c_{1\text{min}}$, $c_{2\text{min}}$ are initial and final values, respectively, $t$ is the number of iterations and $t_{\text{max}}$ is the maximum number of iterations [34]. $\text{rand}_1$ and $\text{rand}_2$ are random variables between 0 and 1. Each solution is evaluated by Eq. (16) which should be minimized to achieve proper clustering.

$$SSE = \sum_{k=1}^{K} \sum_{S_x \in c_k} \|x_i + \mu_k\|^2$$  \hspace{1cm} (16)

$x_i$ is a data point belonging to the cluster $c_k$ and $\mu_k$ is the mean of the cluster $C_k$. To avoid in local optimum, a proposed Gaussian mutation operator based on [35] and [36] is applied to the global particle as follows:

$$v'_{\text{gbest}}(d) = v_{\text{gbest}}(d) \times G(0, h) \times (x_{\text{max}}(d) - x_{\text{min}}(d))$$  \hspace{1cm} (17)

$$x'_{\text{gbest}}(d) = x_{\text{gbest}}(d) + G(0, h) \times v'_{\text{gbest}}(d)$$  \hspace{1cm} (18)

Where $X_{\text{gbest}}$ and $V_{\text{gbest}}$ represent the position and velocity of global best particle. $X_{\text{max}}$ and $X_{\text{min}}$ are the maximum and minimum value in $d^{\text{th}}$ dimension. Gaussian $(0, h)$ is Gaussian distribution with the mean 0 and the variance $h$. $h$ is linearly decreased during each iteration according to Eq. (19) to ensure that the capability of exploration and exploitation are strong at the initial and later stages respectively.

$$h(t + 1) = h(t) - (1/t_{\text{max}})$$  \hspace{1cm} (19)

Where $t_{\text{max}}$ is the maximum number of iterations. Fig. 7 is an illustration of the Gaussian mutation.

After obtaining the centroids from PSO, K-means is applied to refine the obtained centroids. The routine of the proposed clustering algorithm is shown
Figure 7: Visualization of the Gaussian Mutation Operator. In each iteration of hybrid PSO, one centroid \((C_q)\) is chosen from \(X_{\text{gbest}}\) randomly. Then, Gaussian distribution is applied on position and velocity of selected centroid based on Eq. (17) and (18) to create a new offspring \(C'_q\). The new global best \((X'_{\text{gbest}})\) is then compared to \(X_{\text{gbest}}\). If \(X'_{\text{gbest}}\) has better fitness value than \(X_{\text{gbest}}\), \(X'_{\text{gbest}}\) is replaced with current global best.

Algorithm 1: Hybrid PSO with Gaussian Mutation and K-means (HPGMK)

\[
\text{Input: } D=\{d_1, d_2, \ldots, d_n\} \quad //\text{Set of data points} \\
\quad k \quad //\text{Number of desired activities (clusters)} \\
\text{Output: } \text{Set of } k \text{ clusters} \\
1 \text{ Initialize a population of particles with random positions and velocities in the search space} \\
2 \text{ for } \text{iter}=1 \text{ to the maximum number of iteration do} \\
3 \quad \text{ for each particle } i \text{ do} \\
4 \quad \quad \text{ Update position and velocity of particle } i \text{ according to Eq. (11) and Eq. (12)} \\
5 \quad \quad \text{ Evaluate Fitness value of particle } i \text{ according to the fitness function in Eq. (16)} \\
6 \quad \quad \text{ Update } pbest_i(t) \text{ and } gbest(t) \text{ if necessary} \\
7 \quad \quad \text{ for } T \text{ times do} \\
8 \quad \quad \quad \text{ Mutate } gbest(t) \text{ according to Eq. (18) and (17)} \\
9 \quad \quad \quad \text{ Compare mutated } gbest(t) \text{ with previous and choose the best as new } gbest(t) \\
10 \text{ Use } gbest(t) \text{ as the initial centroids} \\
11 \text{ while until no change do} \\
\text{ // Refining the centroids} \\
12 \text{ Calculate distances of data points to centroids} \\
13 \text{ Assign data points to the closest cluster} \\
14 \text{ Centroids are updated using following equation} \\
\text{ centroid}_i = \frac{1}{n_i} \sum_{d_i \in C_i} d_i, \quad \text{where } n_i \text{ is the number of data points in the cluster } i
4. Experiments

4.1. Datasets

Three benchmark datasets were used to evaluate the effectiveness of the proposed method; Cornell Activity Dataset (CAD-60) [37], UTKinect-Action3D (UTK) [38], and Florence3D (F3D) [39]. Moreover, other two datasets including Kinect Activity Recognition Dataset (KARD) [40] and MSR Daily-Activity3D (MSR) [41] are also used to evaluate the proposed method that their results are provided as supplementary material to this paper. These datasets have different dimensions, features, and activities. Table 1 shows the statistical information of these datasets. They are discussed as follows.

| Frequency/Dataset | CAD-60 | UTK  | F3D |
|-------------------|--------|------|-----|
| Activity          | 14     | 10   | 9   |
| Subjects          | 4      | 10   | 10  |
| Videos            | 60     | 200  | 215 |

Table 1: Number of activities, subjects and videos in the three datasets used

**CAD-60**: This dataset includes 14 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, still (standing), random, and working on computer. Each activity was performed by 4 subjects including one left-handed person. They were performed in 5 different environments: bathroom, bedroom, kitchen, living room, and office. It contains activities of cyclic nature such as brushing teeth and similar activities such as drinking water and talking on the phone.

**UTK**: There are 10 activities in this dataset: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands. These activities were performed by 10 subjects and repeated twice by each subject. The significant intra-class and viewpoint variations are the main challenges of this dataset.

**F3D**: This dataset includes 9 activities: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, and bow. These activities are repeated twice or thrice by 10 subjects. The challenge with this dataset is that activities are performed at high speed. This provides a small number of frames for the algorithm to sample and learn from.
4.2. Method

Through preliminary experiments, the best values for swarm size and number of iterations were 20 and 50 respectively. The experiment was repeated 30 times and the average was obtained. The performance of our proposed method (HPGMK) was evaluated by comparing with K-means Clustering (KM), SC, ENSC, SSC, and PSO methods. This comparison was evaluated using Accuracy [22]. Moreover, F-score was used to show the performance of each method in categorizing each activity and the confusion between them was shown in confusion matrix. Convergence test and clustering time of HPGMK are shown in supplementre to acknowledge the benefits of each component used in the HPGMK on its performance.

4.3. Results and discussion

Fig. 8 shows the accuracy of HPGMK with the state-of-the-art techniques for all subjects of each dataset based on the maximum, minimum and average accuracies. The average overall accuracy of the HPGMK was 77.53 % for CAD-60, 56.54 % for UTK, and 66.84 % for F3D. As seen in Fig. 8, HPGMK has the best performance in terms of maximum and average accuracy in all datasets. This proves the robustness and effectiveness of the HPGMK for human activity discovery. By utilizing the Gaussian mutation and KM along with PSO, our approach brings significant performance improvement compared to the other methods. As PSO sometimes gets stuck in local optimum, Gaussian mutation mitigate this. Additionally, applying KM enhances the exploration ability of PSO. Moreover, ENSC and SSC, which are subspace clustering algorithms, do not achieve good cluster quality compared to PSO and HPGMK because these algorithms do not use an efficient search strategy [42]. These approaches did not maintain the balance between exploitation and exploration well and sometimes get stuck in the local solutions. Parameters are required to be set and finding the right values for them is tricky and complex such as size of subspace [43, 44]. In contrast, HPGMK uses the power of PSO in exploration, KM in exploitation, and proposed Gaussian mutation to prevent premature convergence which improve the exploration and exploitation capability as a whole.

To show activity discovery performance, we show the confusion matrix for different methods applied on CAD-60 in this paper in Fig. 9. This figure demonstrate that HPGMK produced distinctively meaningful clusters. Because keyframes have the most important information to differentiate between activities compared to other frames and led to reduced overlapping
among the activities in the results of the proposed method. Moreover, instead of using all the joints to extract the relevant features which can increase redundancy and overlapping between activities, this method uses features from informative joints to differentiate the activities. In addition, the Gaussian mutation, with exploration capability has improved solution pool in the swarm that help to differentiate the different similar activities. On the other hand, cluster overlapping appears relatively high in the other methods especially KM and SC in all datasets. In CAD-60 the largest error in the activity assignment occured between brushing teeth and random due to similar body gestures. There was also activity overlapping between brushing teeth, and wearing contact lenses, because of similar hand movement near the head.

Fig. 10 to 12 show the average F-score for each activity for all subjects. By examining the average F-scores in all datasets, it shows HPGMK outperforming other methods in all datasets used. HPGMK achieved slightly under 75% on CAD-60, just over 45% on UTk, and almost 60% on F3D. In Fig. 10 for HPGMK although F-score is high in most of the activities compared with other methods, Cooking (chopping) and Talking on the phone were discovered with low F-scores. This was due to the high similarity between Cooking (chopping) with Cooking (stirring) and Talking on the phone with Wearing contact lenses. In Fig. 11 a few actions were discovered with low F-scores. For example, throw samples were easily mistakenly regarded as push ones due to similar hand movements, resulting in low F-scores in both activities. In Fig. 12 we observe HPGMK outperformed other methods except for activities wave, clap and tight lace.
Figure 9: Comparison of confusion matrix of CAD-60 on subject 1. Activity list: (1) Brushing teeth; (2) Cooking (chopping); (3) Rinsing mouth with water; (4) Still(standing); (5) Taking on the couch; (6) Talking on the phone; (7) Wearing contact lenses; (8) Working on computer; (9) Writing on whiteboard; (10) Drinking water; (11) Cooking (stirring); (12) Opening pill container; (13) Random; and (14) Relaxing on couch.
Figure 10: The average F-score for all subjects in CAD-60. Activity list: (1) Brushing teeth; (2) Cooking (chopping); (3) Rinsing mouth with water; (4) Still(standing); (5) Taking on the couch; (6) Talking on the phone; (7) Wearing contact lenses; (8) Working on computer; (9) Writing on whiteboard; (10) Drinking water; (11) Cooking (stirring); (12) Opening pill container; (13) Random; and (14) Relaxing on couch. AVG is the average F-score for all activities.

Figure 11: The average F-score for all subjects in UTK. Activity list: (1) Walk; (2) Sit down; (3) Stand up; (4) Pick up; (5) Carry; (6) Throw; (7) Push; (8) Pull; (9) Wave hands; and (10) Clap hands. AVG is the average F-score for all activities.
5. Conclusion

We have proposed a novel Hybrid Particle Swarm Optimization with Gaussian Mutation and k-means (HPGMK) approach to solve human activity discovery on skeleton-based data with no prior knowledge of the activities in the data, apart from the number of activities. A method based on kinetic energy was used to select important frames to overcome the data redundancy problem. A novel feature extraction method was introduced to extract joint displacement, joint orientation, joint angle, and statistical time domain features from 3D skeleton data. In this method, instead of using all joints to extract the features, informative joints were used to prevent the creation of redundant features that can cause errors in the identification of activities. Using such features, activities with high intra-class variables were well clustered using HPGMK. A segmentation method of sampling overlapping frames was employed to divide the entire incoming video into instances of the activities and retain important characteristics of them. A hybrid evolutionary clustering was proposed to discover human activity. It exploited the advantages of PSO, Gaussian mutation, and KM algorithm. This clustering has improved the convergence speed of PSO and prevented it from falling
into the local optimum. Three different datasets were used to assess the performance of our method. The results obtained show that HPGMK achieved an average overall accuracy of 77.53%, 56.54%, and 66.84% in datasets CAD-60, UTK, and F3D respectively and validate the significant superiority of HPGMK over other methods. In activities with high intra-class variables and corrupted data with sitting and standing positions, HPGMK has been effective in activity discovery compared to other state-of-the-art methods. Because Gaussian mutation in HPGMK maintain good, diverse and meaningful particles to evolve, without losing the meaningful hidden structure during exploration. Moreover, the advantage of KM in exploitation has not only increased the efficiency of discovery, but also improved the convergence rate and reduced clustering time. This work paves the way towards the implementation of fully unsupervised human activity discovery in practical applications using skeleton-based data. There are various factors in the proposed methods can impact the efficacy of the final clusters. One factor is the number of clusters that fits performed activities. This indicates the proposed HPGMK can be further extended to address the task of human activity discovery automatically by estimating the number of activities. Another factor is detecting outlier or noisy data. Outliers shift the cluster centers towards itself, thus affecting optimal cluster formation. It can be beneficial to use outlier detection methods in HPGMK to reject outliers.

Acknowledgment

This work was supported by Grant UBD/RSCH/1.11/ FICBF(b)/2019/001 from Universiti Brunei Darussalam.

References

[1] H. V. Chandrashekhar, et al., Human activity representation, analysis, and recognition (2006).

[2] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: European conference on computer vision, Springer, 2016, pp. 816–833.

[3] M. Wang, B. Ni, X. Yang, Learning multi-view interactional skeleton graph for action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).
[4] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Constructing stronger and faster baselines for skeleton-based action recognition, arXiv preprint arXiv:2106.15125 (2021).

[5] G. Paoletti, J. Cavazza, C. Beyan, A. Del Bue, Subspace clustering for action recognition with covariance representations and temporal pruning, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 6035–6042.

[6] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: A review, IEEE transactions on cybernetics 43 (5) (2013) 1318–1334.

[7] C. Wu, X.-J. Wu, J. Kittler, Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[8] K. Su, X. Liu, E. Shlizerman, Predict & cluster: Unsupervised skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9631–9640.

[9] J. Shan, S. Akella, 3d human action segmentation and recognition using pose kinetic energy, in: 2014 IEEE international workshop on advanced robotics and its social impacts, IEEE, 2014, pp. 69–75.

[10] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.

[11] J. Liu, G. Wang, P. Hu, L.-Y. Duan, A. C. Kot, Global context-aware attention lstm networks for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647–1656.

[12] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12026–12035.
[13] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, P. Ogunbona, Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 595–604.

[14] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1112–1121.

[15] J. Cai, N. Jiang, X. Han, K. Jia, J. Lu, Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2735–2744.

[16] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen, J. Chen, Memory attention networks for skeleton-based action recognition, IEEE Transactions on Neural Networks and Learning Systems (2021).

[17] F. Leotta, M. Mecella, D. Sora, Visual process maps: A visualization tool for discovering habits in smart homes, Journal of Ambient Intelligence and Humanized Computing 11 (5) (2020) 1997–2025.

[18] P. Gupta, R. McClatchey, P. Caleb-Solly, Tracking changes in user activity from unlabelled smart home sensor data using unsupervised learning methods, Neural Computing and Applications 32 (16) (2020) 12351–12362.

[19] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, International journal of computer vision 79 (3) (2008) 299–318.

[20] K. Soomro, M. Shah, Unsupervised action discovery and localization in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 696–705.

[21] X. Yan, Y. Ye, X. Qiu, Unsupervised human action categorization with consensus information bottleneck method., in: IJCAI, 2016, pp. 2245–2251.
[22] B. Peng, J. Lei, H. Fu, L. Shao, Q. Huang, A recursive constrained framework for unsupervised video action clustering, IEEE Transactions on Industrial Informatics 16 (1) (2019) 555–565.

[23] X. Yan, Y. Ye, X. Qiu, H. Yu, Synergetic information bottleneck for joint multi-view and ensemble clustering, Information Fusion 56 (2020) 15–27.

[24] W.-H. Ong, L. Palafox, T. Koseki, Autonomous learning and recognition of human action based on an incremental approach of clustering, IEEE Transactions on Electronics, Information and Systems 135 (9) (2015) 1136–1141.

[25] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive recurrent neural networks for high performance human action recognition from skeleton data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2117–2126.

[26] W.-H. Ong, T. Koseki, L. Palafox, Investigation of cluster validity indices for unsupervised human activity discovery, in: Proceedings on the International Conference on Artificial Intelligence (ICAI), The Steering Committee of The World Congress in Computer Science, Computer . . . , 2013, p. 1.

[27] W.-H. Ong, L. Palafox, T. Koseki, Investigation of feature extraction for unsupervised learning in human activity detection, Bulletin of Networking, Computing, Systems, and Software 2 (1) (2013) 30–35.

[28] W.-H. Ong, L. Palafox, T. Koseki, An incremental approach of clustering for human activity discovery, IEEJ Transactions on Electronics, Information and Systems 134 (11) (2014) 1724–1730.

[29] M. M. Arzani, M. Fathy, A. A. Azirani, E. Adeli, Switching structured prediction for simple and complex human activity recognition, IEEE transactions on cybernetics (2020).

[30] D. A. Adama, A. Lotfi, C. Langensiepen, K. Lee, P. Trindade, Human activity learning for assistive robotics using a classifier ensemble, Soft Computing 22 (21) (2018) 7027–7039.
[31] X. Liu, Y. Li, R. Xia, Rotation-based spatial–temporal feature learning from skeleton sequences for action recognition, Signal, Image and Video Processing 14 (6) (2020) 1227–1234.

[32] L. L. Presti, M. La Cascia, 3d skeleton-based human action classification: A survey, Pattern Recognition 53 (2016) 130–147.

[33] P. Hadikhani, P. Hadikhani, An adaptive hybrid algorithm for social networks to choose groups with independent members, Evolutionary Intelligence 13 (4) (2020) 695–703.

[34] J. Cai, H. Wei, H. Yang, X. Zhao, A novel clustering algorithm based on dpc and pso, IEEE Access 8 (2020) 88200–88214.

[35] B. Jana, S. Mitra, S. Acharyya, Repository and mutation based particle swarm optimization (rmpso): A new pso variant applied to reconstruction of gene regulatory network, Applied Soft Computing 74 (2019) 330–355.

[36] C. Li, S. Yang, I. Korejo, An adaptive mutation operator for particle swarm optimization (2008).

[37] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from rgbd images, in: 2012 IEEE international conference on robotics and automation, IEEE, 2012, pp. 842–849.

[38] L. Xia, C.-C. Chen, J. K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE, 2012, pp. 20–27.

[39] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 479–485.

[40] S. Gaglio, G. L. Re, M. Morana, Human activity recognition process using 3-d posture data, IEEE Transactions on Human-Machine Systems 45 (5) (2014) 586–597.
[41] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1290–1297.

[42] Y. Lu, S. Wang, S. Li, C. Zhou, Particle swarm optimizer for variable weighting in clustering high-dimensional data, Machine learning 82 (1) (2011) 43–70.

[43] P. Agarwal, S. Mehta, A. Abraham, A meta-heuristic density-based subspace clustering algorithm for high-dimensional data, Soft Computing (2021) 1–20.

[44] P. Agarwal, S. Mehta, Analyzing subspace clustering approaches for high dimensional data, in: Artificial Intelligence for a Sustainable Industry 4.0, Springer, 2021, pp. 169–195.