Human Activity Discovery with Automatic Multi-objective Particle Swarm Optimization Clustering with Gaussian Mutation and Game Theory

Parham Hadikhani 1,1, Daphne Teck Ching Lai 2, and Wee-Hong Ong 2

1Universiti Brunei Darussalam
2Affiliation not available

October 31, 2023

Abstract

Despite many advances in human activity recognition, most existing works are conducted with supervision. Supervised methods rely on labeled training data. However, obtaining labeled data is difficult, costly, and time-consuming. In this paper, we introduce an automatic multi-objective particle swarm optimization clustering based on Gaussian mutation and game theory (MOPGMGT) to tackle the problem of human activity discovery fully unsupervised and map the multi-objective clustering problem to game theory to get the best optimal solution. The proposed algorithm does not require any prior knowledge of the number of activities to be discovered and can find the optimal number. Multi-objective optimization problems typically cannot have a single optimal solution. To solve this issue, Nash Equilibrium (NE) is applied to the pareto front to choose the best solution. NE does not just look for the best solution, but tries to optimize the final solution by considering the effect of choosing each of the solutions as the best solution on the other solutions and one with the best impact is chosen. Moreover, a Gaussian mutation is applied on the pareto front to avoid premature convergence. Experiments on five challenging datasets demonstrate that the proposed approach is the most efficient to achieve better accuracy in human activity discovery and also can determine the optimal number of clusters.
Human Activity Discovery with Automatic Multi-Objective Particle Swarm Optimization Clustering with Gaussian Mutation and Game Theory

Parham Hadikhani*, Daphne Teck Ching Lai, and Wee-Hong Ong

Abstract—Despite many advances in human activity recognition, most existing works are conducted with supervision. Supervised methods rely on labeled training data. However, obtaining labeled data is difficult, costly, and time-consuming. In this paper, we introduce an automatic multi-objective particle swarm optimization clustering based on Gaussian mutation and game theory (MOPGMGT) to tackle the problem of human activity discovery fully unsupervised and map the multi-objective clustering problem to game theory to get the best optimal solution. The proposed algorithm does not require any prior knowledge of the number of activities to be discovered and can find the optimal number. Multi-objective optimization problems typically cannot have a single optimal solution. To solve this issue, Nash Equilibrium (NE) is applied on the pareto front to choose the best solution. NE does not just look for the best solution, but tries to optimize the final solution by considering the effect of choosing each of the solutions as the best solution on the other solutions and one with the best impact is chosen. Moreover, a Gaussian mutation is applied on the pareto front to avoid premature convergence. Experiments on six challenging datasets demonstrate that the proposed approach is the most efficient to achieve better accuracy in human activity discovery and also can determine the optimal number of clusters.

Index Terms—Human activity discovery, unsupervised learning, clustering, multi-objectives, evolutionary algorithm, game theory, dimension reduction, skeleton sequence.

I. INTRODUCTION

The building of intelligent systems that can interact with humans in tasks such as security surveillance, health care, game control, and robot vision using massive datasets of performed human activities has led to active research in HAR [1]. As shown in Fig 1, the process of a HAR system has five steps [2]. The first step is to receive input data (Fig 1(a)). There are mainly three types of data used for categorizing activities, and they are sensor-based, RGB-videos, and 3D-skeleton data [2]. In this work, we work with 3D-skeleton data. Recent research has shown that using the 3D-skeleton data partially overcomes the issues of noise, data complexities, and human privacy [3]. In 3D skeletal data, each frame is represented by 3D coordinates of the body’s main joints that describe a static activity motion. This data type is highly appropriate for representing human actions because human activities take place in a three-dimensional space [4]. Also, such data can be easily obtained in real-time with low-cost depth sensors [5]. But some frames contain noise, and many frames are similar. Such occurrences increase the computational load and reduce the accuracy in recognizing activities [6], [7]. To overcome these issues, a keyframe selector based on kinetic energy is used to select the frames [7]. After capturing the activities and converting them into the desired form (3D skeleton-based), they are mapped into informative and non-redundant features. These features facilitate the HAR system to have reliable recognition by describing unknown activities (Fig 1(b)). In this paper, we extract features based on the method in [2]. These features are selected because unlike most feature extraction methods, which examine only one aspect of the characteristics of activities and use all joints, different aspects of the characteristics of activities, including temporal and spatial of informative joints and angle and orientation between some bones that have more effect in representation of activities are extracted. After extracting features, to reduce the time complexity, features are selected by PCA that contain valuable information and can make a good distinction between activities. Furthermore, frames containing extracted features are segmented by time windows of equal size and fed to the next step of HAR.

The Discovery step (Fig. 1(c)) is one of the crucial tasks in HAR for identifying activities without prior knowledge, which is called Human Activity Discovery (HAD). The task of HAD is to categorize human activities without labels or any previous knowledge. Thus, HAD is challenging due to the lack of labeled information to identify human activities. This paper focuses on this less-developed step compared to other parts of the HAR system to cluster the human activities. Most of the unsupervised approaches [8] presented for HAD are based on well-segmented videos, and each segment of the video represents exactly one activity sample. Moreover, they know the number of performed activities in the received videos. In reality, the received videos are unsegmented, and each video contains an unknown number of activities. In this paper, we cluster the performed activities without prior knowledge from an unsegmented video that contains multiple activity instances. The next step in the HAR workflow is to build a model from clustered activities to recognize the performed activities. In this step, the HAR system is trained by a learning algorithm.

P. Hadikhani, D.T.C. Lai, and W.H. Ong are with the School of Digital Science, Universiti Brunei Darussalam, Brunei. E-mail: {20b8561, daphne.lai, weehong.ong}@ubd.edu.bn.
D.T.C. Lai, and W.H. Ong are with Institute of Applied Data Analytics, Universiti Brunei Darussalam, Brunei
* Corresponding author
(Fig. 1(d)) to create the correct classification of human activities. Finally, the learned model is utilized to recognize new activities (Fig. 1(e)). Many supervised learning approaches have been proposed to recognize the human activities using training data with ground truth labels e.g., dynamic time warping [9], hidden Markov model [10], CNN based [11], [12] and RNN based approaches [13], [14]. Although these approaches have achieved spectacular progress, the discovery step is ignored in their model. In other words, these approaches are dependent on labeled data. In reality, labeled data are not available, and manual annotation of human activities is time-consuming due to the large volume of activities performed. Moreover, the possibility of human error is high.

Intelligent optimization algorithms used to solve clustering problems are divided into two general categories; single-objective and multi-objective. The goal of clustering is to create appropriate groupings of input data points by optimizing some objective functions. Each objective function represents the different properties of the clusters, such as compaction, separation, and connection. Since it is known that none of the clustering evaluation criteria can work well for all types of datasets, optimizing only one objective function cannot capture various characteristics of datasets. Therefore, single-objective optimization algorithms cannot be effective in clustering high-dimensional data or data with complex structure [15]. This work proposes an automatic Multi-Objective Particle Swarm optimization (PSO) clustering with Gaussian mutation and Game Theory (MOPGMGT) to discover the human activities fully unsupervised. In this approach, the search space of PSO clustering for solving the multi-objective problem and finding the best solution among all potential solutions is mapped to the game theory space. It simultaneously estimates the number of clusters to cluster activities automatically. Moreover, to produce diverse solutions to avoid premature convergence, Gaussian mutation is applied to non-dominated solutions. In multi-objective optimization, a set of optimal solutions is obtained, which makes finding the best solution challenging. To solve the multi-objective problem and find the global best, particles and objective functions are mapped to game theory (GT) space. Using Nash equilibrium (NE), the effect of each solution on other solutions is evaluated to find the global best solution from the pareto front. In addition, it affects the final result by considering the effect of the solutions on each other while choosing the best answer. To improve the result of discovery and produce diverse solutions, a Gaussian mutation is applied to non-dominated solutions. Moreover, a validity index is used to find the optimal number of activities.

The main purpose of this paper is to present a fully unsupervised and automatic approach to discover human activities using unsegmented skeleton data with no prior information about the activities. The main contributions of this paper are summarized as follows: 1) An automatic multi-objective clustering approach to discover human activities and find the optimal number of performed activities. We investigate the application of multi-objective clustering in HAD 3D skeleton video data streams to categorize activities without any prior knowledge. To our knowledge, no previous research has looked into the issue of HAD with multi-objective clustering.

2) The application of Gaussian mutation on clustering to escape from local optimal. 3) The application of NE to find the best optimal solution in a multi-objective problem. 4) Achieving competitive performance on five challenging 3D skeleton-based datasets.

The rest of the paper is structured as follows: in section II, the research background and the methods introduced for human activity discovery are reviewed. The proposed approach is described in section III. In section IV, the proposed method is evaluated and compared with other methods, and finally, in Section V, the conclusion is stated.

II. RELATED WORK

A. Multi-objective game theory

In the multi-objective problem, choosing the best solution is one of the most critical challenges. Li et al.[16] mapped the multi-objective problem into game theory for scheduling. They used Nash equilibrium to find the best solution and used that solution in a hybrid genetic and tabu search algorithm. In our work, game theory is used to find the best solution during clustering for the clustering problem. To the best of our knowledge, this is the first time that game theory is used to solve multi-objective problems by PSO. Heloulou et al. [17] proposed a new multi-objective clustering based on game-theoretic. They predicted the number of clusters by addressing the conflict between clusters based on a sequential two-player game. Although they achieved promising results, their approach was computationally expensive. The difference between our work and this method is that we apply game theory on PSO and map particles and objectives to game theory space to find the global best, but they used game theory to predict the number of clusters.

B. Skeleton-based Human activity recognition and discovery

The existing methods proposed for HAR can be categorized based on three data input types including sensor-based [18], RGBD based [19], and 3D skeleton-based. We review approaches that have been proposed for HAR on 3D skeleton-based data.

1) Supervised approach: A large number of methods in HAR are supervised, one of the the-state-of-the-art supervised methods has been proposed by [20]. They proposed a new attention mechanism based on residual learning to represent temporal features better. Also, a convolutional network was used to model spatial features. Then, a softmax classifier was used to categorize the activities. Alsarhan et al. [21] introduce a graph convolutional network to extract discriminative features. They integrate the squeeze and excitation module into their model to suppress unimportant features. The obtained features are fed into a Softmax classifier to acquire the final prediction. The problem with these models is that despite the brilliant results, they need a lot of guidance from the user to work. In contrast, our work performs all of the processes without using labeled data. Moreover, these supervised methods are aware of the number of activities and the outcome of prediction during the activity recognition
Fig. 1. Conceptual framework for human activity recognition system: (a) input frames are recorded by RGB-D sensor and (b) converted to skeleton data. After (c) extracting features from skeleton data, (d) activities are clustered based on the similarities and differences of features. Next, the system (e) learns the model for each activity based on clusters obtained in the discovery step, and finally (f) human activities can be recognized.

process, but in our proposed method, there is no prior knowledge of the number of activities performed and the outcome.

2) Partially supervised approach: In the case of unsupervised skeleton-based methods, Su et al. [22] presented an auto-encoder to represent the activities unsupervised. They used a recurrent neural network in the structure of the auto-encoder. After achieving the features, they applied the KNN classifier to group the activities. In [23], in addition to using auto-encoder, they normalized motion sequence based on spatial-temporal asynchronous method during the process of auto-encoder. They pruned the temporal information that had less effect on the learning process. They generated the motion sequences that were less affected by the form of human bodies. Finally, the KNN classifier was used to assign activities to the relevant group. In [24] auto-encoder was formulated into an Expectation–maximization algorithm to improve the features. First, encoded features were used to create action prototypes by k-means clustering. Then, in addition to giving encoded features to the decoder, similar encoded features were approaching nearer to the action prototypes based on contrastive learning. The loss function was calculated based on the decoder and contrastive learning results. Finally, a linear fully connected layer along with softmax classifier were used to classify the activities. The difference between our method and unsupervised methods is that they first receive the activities in fully segmented videos, which means they are already aware of the type of activities. Secondly, they are aware of the number of activities. Also, the whole process of HAR is done without any knowledge.

3) Fully unsupervised approach: Given that recognizing human activities without the need for guidance from the user during the training phase from the skeleton-based data is challenging, there are few approaches to discover skeletal-based activity. One of the seminal works to discover human activity was presented by Ong et al. [25]. They proposed an incremental clustering that was able to find the number of clusters. Contrary to conventional approaches that estimate the number of clusters. In their approach, the number of clusters was obtained during the clustering. Moreover, a mixture of Gaussians hidden Markov model was employed to model the activities. The difference between our method and this method is that we examined problem HAD in a multi-objective manner, but their method was based on improving a single clustering goodness metric. Using a single validity measure cannot detect overlap between activities correctly and causes these methods to perform poorly in discovering activities. In one of the state-of-the-art methods for HAD, Paoletti et al. [3] introduced a novel subspace clustering combined with covariance descriptors method to cluster activities. First, they trimmed the data to distinguish between activity sequences. After that, several methods of subspace clustering were applied to discover activities. Although this method has achieved good results and was very competitive compared with the supervised methods, their approach needed to receive the activities in thoroughly segmented videos and predefined the number of clusters. Furthermore, their technique necessitates knowledge of the subspace dimensions, making it sensitive to initialization. Hadikhani et al. [2] proposed a new mechanism based on an evolutionary algorithm which is the latest the-state-of-the-art for HAD. They first provided a hybrid feature extraction method to consider most aspects of human activity. Then, PSO clustering along with k-means was presented to discover the activities. They introduced a new Gaussian mutation to
improve the activity discovery to generate diverse solutions. The difference between this method and ours is that they use only one criterion to discover the activities that lead to not performing well in handling a large number of activities with a complex structure. Also, another problem with their method is being aware of the number of activities. In contrast, our method performs discovery based on several criteria and is unaware of the number of activities. In this paper, the proposed method seeks to solve HAD using a multi-objective manner to distinguish activities with high complexity and perform all processes fully unsupervised. Based on our knowledge, this is the first time HAD has been solved using multi-objective clustering.

III. PROPOSED HUMAN ACTIVITY DISCOVERY

In this section, the proposed method is presented, as shown in Fig. 2. The discovery of human activities consists of two main parts: feature extraction and clustering. In the feature extraction stage, keyframes are first selected by computing the kinetic energy of frames. Next, different features of body movements are extracted based on different characteristics, and important ones are then selected. Afterward, frames are segmented into samples of activity instances. In terms of discovery, activities are discovered using a new approach multi-objective method based on PSO with the integration of game theory. Details of each component of the system are described in the following subsections.

A. Keyframe

The incoming frames have a high degree of correlation and redundancy [7]. The kinetic energy \( E(f_i) \) of each frame \( f_i \) is estimated using Eq. (1) to keep frames that reflect the major contents of a video and improve the computational efficiency of the procedure. The movement of the joint \( j \) in the frame \( i \) is estimated in this way, applied on every joint in the body. The present frame’s energy is the total of these motions for all joints. As seen in Fig. 3, keyframes are frames that have the lowest and maximum kinetic energy in comparison to nearby frames.

\[
E(f_i) = \sum_{j=1}^{J} E(f_i^j) = 1/2 \sum_{j=1}^{J} (f_i^j - f_i^{j-1})^2 \quad (1)
\]

B. Feature extraction

Based on the method in [2], a set of features comprising of displacement-based, statistical [26], angles [27], and orientation features [28] are extracted to provide a good description of the activities performed in a video. In terms of displacement-based features, distances between joints of both hands, hands and head, and hip and feet at both sides in the same frame are computed as temporal displacement. Statistical features are obtained by computing the difference of each of the selected joints \( P^j_k \) in current frame \( k \) from the mean \( J_{mean} \) and standard deviation \( J_{var} \) of the same joint within an activity sequence to make a good distinction between activities related to the arms and legs. To describe human body posture, the orientation and angle features are extracted. For obtaining the orientation between the bones, a rotation matrix is used. This matrix is calculated using the bones rotation angle relative to each other. These angles are computed with the assistance of internal and external products between the bones. The rotation angles relative to the x, y, and z axes are considered as orientation features which are shown in Fig. 4 (d). Moreover, the angle features are computed based on the angles between the bones of elbow-wrist and shoulder-elbow at both sides and the angles between the bones of hip-knee and knee-ankle at both sides. These angles are highlighted in Fig. 4 (e).

C. Feature selection and Sampling

After obtaining the features, PCA is employed to reduce the dimension to reduce the time complexity. Frames are segmented into specific time windows by the sliding windows method. The number of frames in each window should be large enough to reflect a subset of an activity. In this paper, the length of each window was 15 frames. The overlap of sliding windows increases clustering performance because it makes sure that transitions between activities are not missed [29]. 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. 5).

D. Mutation-based Multi-objective PSO clustering with integrated game theory

PSO uses a swarm of particles which each of them represents a solution. Each particle has a velocity and tries to find the global optimization by moving their position. The velocity and position of the particles are updated based on the following formula [30]:

\[
x_i(t+1) = x_i(t) + v_i(t) \quad (2)
\]

\[
v_i(t+1) = w \times v_i + c_1 \times rand_1 \times (pbest_i(t) - x_i(t)) + c_2 \times rand_2 \times (gbest(t) - x_i(t)) \quad (3)
\]

\[
w = \frac{w_{max}}{t_{max}} + t \times (w_{max} - w_{min}) \quad (4)
\]

\[
c_1(t+1) = (c_{1max} - c_{1min}) \times \frac{t}{t_{max}} + c_{1max} \quad (5)
\]

\[
c_2(t+1) = (c_{2max} - c_{2min}) \times \frac{t}{t_{max}} + c_{2max} \quad (6)
\]

where in Eq. (2) and (7) \( x_i \) is position and \( v_i \) is velocity of the particle \( i \) at time \( t \), \( pbest_i \) and \( gbest \) are the best particle position and the best group position, respectively. \( rand_1 \) and \( rand_2 \) are random value parameters with range [0,1]. The parameter \( w \) is a constant weight of inertia which is defined by the Eq. (4). \( w_{max} \) is initial weight. \( c_1 \) and \( c_2 \) are acceleration coefficients expressed by Eq. (5) and (6) [31].
Fig. 2. Illustration of the proposed system. After receiving input frames and selecting keyframes by computing kinetic energy, features are extracted based on joint displacement, statistical time domain, and angle and orientation of bones. Next, extracted features are normalized. Important features are selected by using PCA. Before clustering the activities, the sliding window technique is used to segment frames into the fix sized time overlapping windows. In the discovery step which is executed for different values of $k$ in the range of $k_{\text{min}}$ to $k_{\text{max}}$, centroids are first selected randomly from samples. After initialization and evaluation of each solution based on objective functions used, non-dominated solutions are obtained. To avoid getting stuck into the local optimal trap, Gaussian mutation is applied to non-dominated solutions. Then, Nash equilibrium is used to select the global optimal solution. Finally, the Jump method is employed to find the optimal number of clusters and return the result of the estimated number of clusters.
A possible solution is obtained. Each of these solutions shows the difference between MOPSO with PSO is that more than one solution set in all objectives and that solution is strictly better than other members in at least one objective (see Fig. 6 (a)). A Gaussian mutation operator is applied to the MOPSO to reduce the risk of falling into the local optimum, commonly found in evolutionary computing problems, by using a random value from a Gaussian distribution to introduce new individuals from the current generation into the population and boost population diversity. The Gaussian mutation operator used in this paper is based on [32] and [33]. It is applied to the global particle as follows:

\[
v'_{\text{non-dominated}}(d) = v_{\text{non-dominated}}(d) \\times G(0, h) \times (x_{\max}(d) - x_{\min}(d))
\]

\[
x'_{\text{non-dominated}}(d) = x_{\text{non-dominated}}(d) \\times G(0, h) \times v_{\text{non-dominated}}(d)
\]

Where \(x'_{\text{non-dominated}}(d)\) and \(v'_{\text{non-dominated}}(d)\) represent the position and velocity of \(d\)th non-dominated particle in \(d\)th dimension. \(x_{\max}\) and \(x_{\min}\) are the maximum and minimum value in \(d\)th dimension. \(G\) is Gaussian distribution with the mean 0 and the variance \(h\). \(h\) is linearly decreased over each iteration according to Eq. (10) to guarantee that the exploration and exploitation capabilities are intensified at the beginning and end of the process. This means that initially, a large area of the problem space is searched due to little knowledge the approximate location of the optimal solution. But over time, as the approximate optimal solution is approached, the search space becomes more limited to looking for a better solution around the global best solution.

\[
h(t + 1) = h(t) - (1/t_{\max})
\]

Game Theory (GT) is applied on a pareto-optimal set to make a decision to select the global optimal solution. GT is a study of how to mathematically determine the best strategy for given conditions in order to optimize the outcome. A game space has three essential features; players of the game, player’s strategies are actions that the player can choose in different steps of the game, and payoff which represents the result of the strategies employed by each player. To build the GT model of the pareto-optimal set, as shown in Fig. 6 (b), each objective function can be described as a player in the game, particles are considered as players’ strategies, and the fitness value of each objective function as a payoff \((\pi)\). The solution of a game is obtained by reaching a strategic equilibrium.

We employed two objectives to assess different elements of clusters. One is SSE, which looks at a distance inside the cluster, and the other is Cnn-index, which looks at the distance between clusters (relevant explanations are given in Section IV-B). A balance must be struck between these two

Since single objective clustering cannot cluster different properties of datasets such as compactness and separation, multi-objective functions in PSO (MOPSO) are used to optimize solutions simultaneously in an evolving population. Mathematically, a multi-objective optimization problem can be formulated as follows:

\[\text{Min} f_i(x) : i = 1, ..., N\]

where \(f_i\) is the objective function, \(N\) is the number of fitness functions, and \(x\) is the decision vector. The most important difference between MOPSO with PSO is that more than one possible solution is obtained. Each of these solutions shows the trade-off between the different objective functions. Given a set of solutions, non-dominated ones (pareto-optimal set) in the swarm are saved to a sub-swarm called repository. The repository has a predefined size. If the number of non-dominated particles exceeds the predefined size repository, some of the non-dominated particles archived in the repository are then removed using roulette wheel selection. A solution is non-dominated solution if it is no worse than any member of the solution set in all objectives and that solution is strictly better than other members in at least one objective (see Fig. 6 (a)). A Gaussian mutation operator is applied to the MOPSO to reduce the risk of falling into the local optimum, commonly found in evolutionary computing problems, by using a random value from a Gaussian distribution to introduce new individuals from the current generation into the population and boost population diversity. The Gaussian mutation operator used in this paper is based on [32] and [33]. It is applied to the global particle as follows:

\[
v'_{\text{non-dominated}}(d) = v_{\text{non-dominated}}(d) \\times G(0, h) \times (x_{\max}(d) - x_{\min}(d))
\]

\[
x'_{\text{non-dominated}}(d) = x_{\text{non-dominated}}(d) \\times G(0, h) \times v_{\text{non-dominated}}(d)
\]

Where \(x'_{\text{non-dominated}}(d)\) and \(v'_{\text{non-dominated}}(d)\) represent the position and velocity of \(d\)th non-dominated particle in \(d\)th dimension. \(x_{\max}\) and \(x_{\min}\) are the maximum and minimum value in \(d\)th dimension. \(G\) is Gaussian distribution with the mean 0 and the variance \(h\). \(h\) is linearly decreased over each iteration according to Eq. (10) to guarantee that the exploration and exploitation capabilities are intensified at the beginning and end of the process. This means that initially, a large area of the problem space is searched due to little knowledge the approximate location of the optimal solution. But over time, as the approximate optimal solution is approached, the search space becomes more limited to looking for a better solution around the global best solution.

\[
h(t + 1) = h(t) - (1/t_{\max})
\]

Game Theory (GT) is applied on a pareto-optimal set to make a decision to select the global optimal solution. GT is a study of how to mathematically determine the best strategy for given conditions in order to optimize the outcome. A game space has three essential features; players of the game, player’s strategies are actions that the player can choose in different steps of the game, and payoff which represents the result of the strategies employed by each player. To build the GT model of the pareto-optimal set, as shown in Fig. 6 (b), each objective function can be described as a player in the game, particles are considered as players’ strategies, and the fitness value of each objective function as a payoff (\(\pi\)). The solution of a game is obtained by reaching a strategic equilibrium.

We employed two objectives to assess different elements of clusters. One is SSE, which looks at a distance inside the cluster, and the other is Cnn-index, which looks at the distance between clusters (relevant explanations are given in Section IV-B). A balance must be struck between these two
The proposed clustering algorithm is shown in Algorithm 1.

**Algorithm 1: MOPGMGT algorithm**

**Input:** $D = \{d_1, d_2, \ldots, d_n\}$ //Set of data points

- $K_{\text{max}}$ //Maximum number of cluster
  - calculated by $\sqrt{\text{size}(D)}$

**Output:** The best clustering result,

Number of clusters

1. for $K_{\text{min}} = 2 \rightarrow K_{\text{max}}$
do
2. Randomly generate initial positions and velocities of particles
3. Calculate the cost of each particle using Eq. (14) and (15)
4. Set the initial personal best of each particle
5. Repository $\leftarrow$ non-dominated solutions
6. Global best $\leftarrow$ none
7. iter $\leftarrow$ 0, MaxIteration $\leftarrow$ 50
8. while iter $\prec$ MaxIteration do
  
  for each solution in Repository do
    
  Calculate the $NashE_j$ for non-dominated solutions using Eq. (11)
  
  Global best $\leftarrow$ find solution with the best $NashE_j$
  
  for each particle do
    
  Update particle velocity and position
  
  Calculate the new cost of each particle
  
  Eq. (14) and 15)
  
  Update the particle personal best
  
  if $\text{Size (Repository size)} > \frac{\text{number of particles}}{2}$ then
    
  Until $\text{Size (Repository Size)} > \frac{\text{number of particles}}{2}$, delete one Repository members based on “Roulette Wheel Selection”
  
  for $T$ times do
    
  Mutate solutions in Repository according to Eq. (8) and (9)
  
  Compare mutated version with previous one and choose non-dominated one
  
  iter $\leftarrow$ iter+1
  
  Select the Best-Solution using Nash Equilibrium
  
  Select best value of $k$ with minimum value of Jump

**IV. EXPERIMENTS**

**A. Datasets**

We validate our method on Six datasets NTU RGB+D Dataset (NTU-60) [35], contain Cornell Activity Dataset (CAD-60) [36], Kinect Activity Recognition Dataset (Kard) [37], MSR DailyActivity3D (MSR) [38], UTKinect-Action3D (UTk) [39], and Florence3D (f3D) [40]. These datasets are very challenging due to significant data corruption, and the similarity and overlapping of many activities. Moreover, they have different numbers of activities, data size, number of joints, and subjects. Consequently, the effectiveness of the proposed model can be well
assessed under different conditions. Table I shows the datasets statistical information. Fig. 7 represent datasets and MOPGMGT clustering results on each of them. In dataset CAD-60 (Fig. 7(a)), data points are very dense due to the large volume of similarities between activities. This makes it difficult to distinguish the exact boundaries between activities. The activities in datasets KARD (Fig. 7(c)) and MSR (Fig. 7(e)) overlap a lot and make it very difficult to identify and differentiate between them. In addition, the same activities are highly dispersed in the MSR dataset. This is because of performing the activities in both standing and sitting positions. Datasets UTK (Fig. 7(g)) and F3D (Fig. 7(i)) have very short instances that cause them to be too scattered and make clustering challenging. In datasets NTU-60 (Fig. 7(k)), data points have became very dense due to the high number of activities. Also, the similarities between same activities are highly dispersed in the MSR dataset. The activities in datasets KARD (Fig. 7(c)) and MSR (Fig. 7(e)) overlap a lot and make it very difficult to identify and differentiate between them. In addition, the same activities are highly dispersed in the MSR dataset. This is because of performing the activities in both standing and sitting positions. Datasets UTK (Fig. 7(g)) and F3D (Fig. 7(i)) have very short instances that cause them to be too scattered and make clustering challenging. In datasets NTU-60 (Fig. 7(k)), data points have became very dense due to the high number of activities. Also, the similarities between the many activities have caused the activities to overlap so that it is not easy to separate them. It can be figured out from the clustering results in Fig. 7 that the MOPGMGT not only has no problem for clustering of non-overlapping activities but also performs well in complex datasets with high overlap.

![T-SNE plot of the performance of MOPGMGT in all datasets. MOPGMGT results for CAD-60, KARD, MSR, UTK and F3D (a,c,e,g,i,k), respectively, and their corresponding Ground Truth (b,d,f,h,j,l).](image)

**C. Evaluation measures**

We compare clustering algorithms using accuracy [19] and precision rate to investigate their performance across the 30 runs. F-score and confusion matrix are used to show the performance of each method in categorizing each activity and the confusion between activities. To show the performance of the proposed method in estimating the number of clusters, the overall error (OE) of estimating the number of clusters is computed for each method as given in Eq. (17).

\[
OE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (k_i^* - k_i)^2}
\]

where \(k_i^*\) is the \(k\) suggested, for subject \(i\), \(k_i\) is the predicted \(k\) for subject \(i\), and \(n\) is all subjects in the dataset.

### TABLE I

**DETAILS OF SIX DATASETS USED FOR EVALUATION**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CAD-60</th>
<th>KARD</th>
<th>MSR</th>
<th>UTK</th>
<th>F3D</th>
<th>NTU-60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>14</td>
<td>18</td>
<td>16</td>
<td>10</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Subjects</td>
<td>4</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>Videos</td>
<td>60</td>
<td>2160</td>
<td>320</td>
<td>200</td>
<td>215</td>
<td>56000</td>
</tr>
</tbody>
</table>

**B. Setup**

The experiment is repeated 30 times. For each run, swarm size and number of iterations are 20 and 50 respectively. \(c_{1\text{max}}\) and \(c_{2\text{max}}\) were set to 2.5 , \(c_{1\text{min}}\) and \(c_{2\text{min}}\) to 0, and \(w_{\text{max}}\) to 0.9. The stop criteria was based on the number of iteration. The performance of the proposed method was compared with the four automatic clustering including PSO [41], HPGMK [2], MOPSO (multi-objective PSO) [42], MOPGM (multi-objective PSO with Gaussian mutation) and four non-automatic clustering (with known number of clusters) algorithms contain KM (k-means), SC (spectral clustering), ENSC (elastic net subspace clustering), and SSC (Sparse Subspace Clustering) [3]. We have used KM, SC as the baseline approach and PSO, MOPSO and MOPGM have been picked for comparison because our proposed method is based on these algorithms. Moreover, ENSC, SSC and HPGMK have been selected as the state-of-the-art methods. The objective function SSE (sum square error) in Eq. (14) was used for all single and multi objectives clustering algorithms which should be minimized to achieve proper clustering.

\[
SSE = \sum_{k=1}^{K} \sum_{x \in c_k} \|x - \mu_k\|^2
\]

\(x_i\) is a data point belonging to the cluster \(c_k\) and \(\mu_k\) is the mean of the cluster \(c_k\). The second objective function for multi objectives clustering algorithms is Conn-index [43] which should be minimized and is calculated as follows:

\[
Conn = \frac{\sum_{i=1}^{K} \min_{j=1,i \neq j} d(m_i,m_j)}{n}\frac{1}{n(n-1)}
\]

where \(n\) is number of objects in cluster \(c_i\) and \(p_j\) is the \(j\)th object of cluster \(i\). Instead of using sampled data based on the knowledge of the beginning and end of activities, the input data is sampled based on our sampling method described in section III, done fully unsupervised. Afterward, the sampled data is given to the mentioned methods for clustering activities.
D. Discussion and results

Fig. 8 indicates the minimum, maximum, and average accuracy of the proposed algorithm and other algorithms on all subjects of each dataset. The overall average accuracy of the MOPGMGT was 72.43% for CAD-60, 47.41% for KARD, 36.78% for MSR, 52.06% for UTK, 56.43% for F3D, and 35.43% for NTU-60. MOPGMGT outperformed other algorithms in four datasets: CAD-60, MSR, F3D, and NTU-60 in terms of maximum, and average accuracy. However, MOPGM performs slightly better than MOPGMGT in KARD and UTK datasets. It is noteworthy, unlike methods KM, SC, ENSC, and SSC, MOPGMGT has no prior knowledge of the number of clusters and automatically estimates them and has been able to outperform all approaches in most datasets. NE has led to the best decision of choosing the most optimal solution in MOPGMGT. Another advantage of the proposed method over other methods is the use of Gaussian mutation to modify the solutions that allows more areas of the search space to be covered and explored. This has resulted in the proposed method performing better than other methods, as well as preventing early convergence and generating more diverse solutions. Moreover, compared to HPGMK and PSO that used a single objective function, MOPGMGT considered two aspects of good solutions in clustering including compactness and connectivity by using two different objective functions. It enabled MOPGMGT to capture diverse properties of activities and cluster them appropriately. Methods KM and SC had the worst results due to being stuck in local optimization. ENSC and SSC have had poor results compared to the MOPGMGT. These methods lack an efficient strategy to balance exploration and exploitation [44], [45]. Fig. 9 shows how to select the global best solution (continuous gray line) from other possible solutions of pareto-front with two methods, game theory (blue dotted line) and roulette while selection (pink dotted line). This experiment was performed knowing the best solution (grey line) in each iteration to show the effect of game theory on finding the best solution. The results confirm that game theory (Fig. 9a) has great performance when it applies to pareto-front to select the global best solution compared to the MOPSO that uses an adaptive grid and roulette wheel selection [42]. As can be seen, unlike roulette wheel selection in Fig. 9b, which has failed to find the global best solution in most iterations, game theory has been able to identify the best solution in almost all iterations. This is because game theory not only seeks for the global best solution but also examines the effects of choosing one of the solutions on the other solutions and chooses a solution that puts the rest of the solutions in the best possible state. In addition, it can be seen from Fig. 9 that game theory has not only been able to find the global best solution but the accuracy of the solution incrementally has improved as the algorithm iterates. In contrast, in the roulette wheel selection method, due to the lack of a specific strategy and consideration for the consequence of picking each of the solutions, it has failed in finding the global best solution.

Fig. 10 shows the comparison of best precision obtained of all automatic clustering algorithms from all subjects of the six datasets used. It has been shown that MOPGMGT surpassed other algorithms in five datasets and achieved 81.07% in CAD-60, 52.13% in KARD, 36.77% in MSR, 62.87% in F3D and 29.63% in NTU-60 except in UTK. The difference is mainly because MOPGMGT simultaneously benefits from Gaussian mutation to avoid getting stuck in the local optimization and generating a variety of solutions, as well as using NE to find the best possible solution. While in each of the methods MOPSO and MOPGM do not have one of the advantages of MOPGMGT. Moreover, PSO and HPGMK have not been able to obtain acceptable results due to single-objective clustering.

Table II shows the result of estimating the number of clusters for different methods by using the Jump index. The best results are presented in bold and when the best results were achieved in several methods, they are in italics. The best results are those that are closer to the actual number of clusters and the worst are the opposite. As can be seen from the table, methods demonstrate different performances in estimating the number of clusters on different datasets. However, the proposed algorithm has the lowest OE due to the better finding of global best solution using NE than the two methods MOPSO and MOPGM, as well as using multi-
Fig. 10. Compare the best precision obtained by different approaches for each dataset.

objective to perform clustering than the PSO and HPGMK. In NTU-60, due to the high number of activities and the similarity of many of them, the algorithms consider a large number of activities as an independent cluster. Therefore, more features need to be extracted to solve this problem.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#clusters</th>
<th>PSO</th>
<th>HPGMK</th>
<th>MOPSO</th>
<th>MOPGM</th>
<th>MOPGMGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD-60</td>
<td>14</td>
<td>19</td>
<td>16</td>
<td>17</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>KARD</td>
<td>18</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>MSR</td>
<td>16</td>
<td>12</td>
<td>22</td>
<td>20</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>UTK</td>
<td>10</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>F3D</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>NTU-60</td>
<td>17</td>
<td>17</td>
<td>19</td>
<td>19</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>OE</td>
<td>-</td>
<td>4.561</td>
<td>4.491</td>
<td>4.457</td>
<td>4.413</td>
<td>4.345</td>
</tr>
</tbody>
</table>

Fig. 11. The average F-score for all subjects in KARD. Activity list: (1) Horizontal arm wave; (2) High arm wave; (3) Two hand wave; (4) Catch cap; (5) High throw; (6) Draw X; (7) Draw tick; (8) Toss paper; (9) Forward kick; (10) side kick; (11) Take umbrella; (12) Bend; (13) Hand clap; (14) Walk; (15) Phone call; (16) Drink; (17) Sit down; and (18) Stand up. AVG is the average F-score for all activities.

Fig. 12. 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.

Fig. 13. The confusion matrix corresponding to MOPGMGT for one subject of different datasets. In MOPGMGT, due to the benefit of multi-objective optimization to consider and examine the different characteristics of each activity, challenging activities with high similarities have been well distinguished, such as in cooking (stirring) and cooking (chopping) in CAD-60 (Fig. 13(a)) and High arm wave in KARD (Fig. 13(b)). NE is also yield the final solution which diminishes the impact of current intra-class in between activities by reaching the best feasible solution. In the corresponding confusion matrix of KARD dataset (Fig. 13(b)), most error in the activity assignment occurred between draw x and draw tick, drink and high throw, forward kick and side kick, toss paper and take umbrella, and bend and hand clap due to the similar motions. In confusion matrix of MSR (Fig. 13(c)), activities overlapping were seen between call cellphone and drink, Read book and Play game, and Stand up and Sit down. One of the reasons for these overlapping is each subject performed each activity twice, once in a standing position, and once in a sitting position. Another reason is data corruption. In other words, in this dataset, some frames and activities miss their skeleton features and become meaningless. In the confusion matrix of UTK (Fig. 13(d)), A3
(stand up) - A2 (sit down), A5 (carry) - A1 (walk), and A6 (throw) - A8 (pull) - A7 (push) had the most number of errors due to the existence of intra-class. From the confusion matrix of F3D (Fig. 13(e)), most of the clustering errors occurred between the wave and tight lace, and read watch and clap) due to the similarity of them. In NTU-60(Fig. 13(f)), Although MOPGMGT in some cases has been able to cluster activities correctly, most activities overlap due to their high similarity as well as viewpoint variations. Another reason for the overlap between activities is the large number of activities, which must be addressed by deep learning to extract the features.

V. CONCLUSION

In this paper, a new approach has been introduced to address the problem of fully unsupervised HAR. We proposed an automatic multi-objective PSO clustering based on Gaussian mutation and game theory (MOPGMGT) to discover human activities from an unsegmented sequence of unlabeled 3D skeleton activities data. Gaussian mutation was applied to non-dominated solutions to avoid PSO clustering from falling into the local optima trap. Moreover, Nash equilibrium in game theory was employed to handle good optimal solution selection in the multi-objective problem. We demonstrated through Nash equilibrium, the best optimal solution is selected from pareto-front in comparison with a conventional method. The MOPGMGT does not need prior knowledge of the number of activities performed. It identifies the number of activities by using the Jump method. Our method achieves promising results by using informative joints for extracting features of activities and selecting keyframes to eliminate redundant information. The MOPGMGT can achieve competitive performance with respect to the state-of-the-art. Our experiments also showed how the balance between exploitation and exploration using multi-objectives can lead to a significant improvement in the detection of overlapping activities. We have shown the effectiveness of MOPGMGT on six challenging datasets. The average overall accuracy is 72.43% in datasets CAD-60, 47.41% in datasets KARD, 36.78% in datasets MSR, 52.06% in datasets UTK, 56.43% in datasets F3D and 35.43% in datasets NTU-60. The results showed that the significant performance superiority and the robustness of our proposed method in comparison with the existing state-of-the-art methods. To the best of our knowledge, this is the first fully unsupervised method proposed with multi-objectives for the task of human activity discovery on 3D skeleton data. Currently, the MOPGMGT considers a fixed length of sliding window to segment input streams of skeleton sequences. However, it might hinder the clustering ability if streams of video are short. To overcome this, a dynamic sliding window based on input data length should be incorporated into the workflow. Our proposed method can be improved based on clustering time through an incremental approach. Moreover, using MOPGMGT to solve other real-world problems such as discovering the emotions from full-body movements, selecting cluster heads in wireless sensor networks, and choosing independent members in the social networks is another possible future work.

ACKNOWLEDGMENT

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

REFERENCES


