Image and Vision Computing 145 (2024) 104985

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Flexible multi-objective particle swarm optimization clustering with game theory to address human activity discovery fully unsupervised

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

School of Digital Science, Universiti Brunei Darussalam, Brunei

ARTICLE INFO

Keywords:
Human activity discovery
Unsupervised learning
Clustering
Feature extraction
Incremental manner
Multi-objectives optimization
Dimension reduction
Skeleton sequence

ABSTRACT

Human activity recognition is a crucial field of study, but current approaches often require ground truth labels, which are not always available. We propose a new method called the Flexible Multi-Objective Particle swarm optimization clustering method based on Game theory (FMOPG), which can identify human activities without any supervision. Unlike traditional clustering methods that require an estimate of the number of clusters and are often inaccurate, FMOPG handles varying cluster numbers with an incremental technique, selecting clusters with good connectivity and separation. We enhance Particle Swarm Optimization (PSO) with mean-shift vectors for faster convergence and better handling of non-spherical clusters. Employing multi-objective optimization and Gaussian mutation, FMOPG provides a range of optimal solutions. We map the optimization problem to game theory to select the best solution based on different criteria. A smart grid-based method is proposed for population initialization, reducing variance and improving reliability. FMOPG outperforms state-of-the-art methods, improving clustering accuracy by 3.65%. Moreover, the incremental technique has improved clustering time by 71.18%.

1. Introduction

HUMAN activity recognition (HAR) has a tremendous impact in computer vision and is considered by many researchers due to its many applications including security surveillance, health care,intelligent transportation systems [1], game control, and robot vision [2]. HAR aims to analyze ongoing activities automatically so that it can correctly categorize the activities performed.

Many works presented in this domain use RGB-videos [3–5] and sensor-based [6–9] data. However, these data have problems such as a large amount of information to process, high amount of noise, difficulty in recording data, and high cost [10]. However, these challenges are mitigated to some extent by the use of 3D skeleton-based data, which has become increasingly popular in recent research [11]. Despite the difficulties in obtaining accurate skeletal information from real-life scenarios, 3D skeleton data offer several advantages. They often contain less redundant information compared to RGB videos, making them easier to process and analyze. Additionally, they can be more robust to noise compared to sensor-based data, which can be affected by environmental factors. Moreover, while the process of obtaining accurate skeletal information can be challenging, it may still be more cost-effective than using expensive sensor equipment for data collection. Overall, despite the challenges involved, the use of 3D skeleton-based data in HAR can offer several advantages that make it a valuable and viable option for many real-life applications [11–15].

As shown in Fig. 1, a HAR system has five major steps. In the first step (Fig. 1(a)), people's activities are recorded. Recent researches have shown that 3D-skeleton data is reliable and can be easily recorded with low-cost depth sensors [11–15]. Further, depth sensor have the advantage of not capturing personal identity images. Each frame of 3D-skeleton data has three-dimensional coordinates of the body's joints and is appropriate for displaying human activities [16]. In the feature extraction step (Fig. 1(b)), the system aims to find the most relevant set of compact and descriptive information that displays the distinct patterns of each activity. Good features help the HAR system to distinguish and identify human activities well. Feature extraction methods can be divided into three categories: displacement, statistical, and orientation. This paper proposes a combination of all approaches presented in [17] to describe the human body's posture and movement. Human Activity Discovery (HAD) is a crucial step within a Human Activity Recognition (HAR) system, depicted in Fig. 1(c). This step involves the identification and clustering of activities based on the similarities found in extracted

* Corresponding author.
E-mail addresses: 20h8561@ubd.edu.bn (P. Hadikhani), daphne.lai@ubd.edu.bn (D.T.C. Lai), weehong.ong@ubd.edu.bn (W.-H. Ong).

https://doi.org/10.1016/j.imavis.2024.104985
Received 22 January 2024; Received in revised form 28 February 2024; Accepted 10 March 2024
Available online 18 March 2024

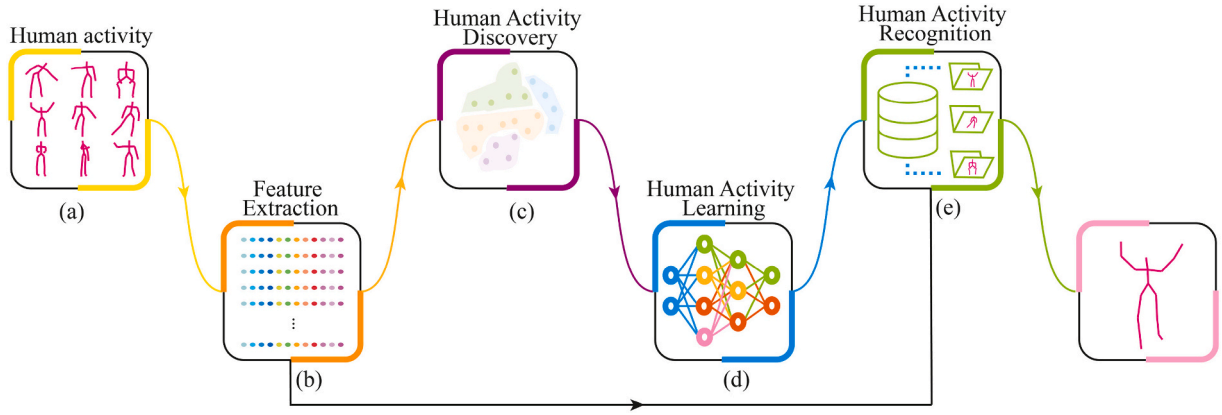

Fig. 1. Conceptual framework for human activity recognition system: (a) RGB-D sensor captures input frames and converts them to skeletal data. Then (b) features are extracted from the skeleton data to represent the unique characteristics of each activity. Subsequently, (c) extracted features are used in the Human Activity Discovery (HAD) step to cluster similar activities together based on their feature similarities. This step is crucial for identifying and categorizing activities without prior knowledge or labels. (d) Once the activities are clustered based on their features, a learning algorithm is applied to create a model for each activity. This model is trained to accurately classify and recognize the activities based on the clusters obtained in discovery step. Finally, (e) the trained model is used to recognize human activities in new input frames. The model compares the features of the input frames with the learned activity models to determine the most likely activity being performed.

features, all without prior knowledge or information about the observed activities [18]. The rationale behind introducing HAD in the context of a HAR system is to tackle the challenges arising from the absence of labeled information regarding human activities [19]. Traditional HAR systems typically rely on supervised methods that necessitate labeled datasets, where each activity is predefined and associated with a specific label [20]. However, acquiring such labeled data in real-world scenarios can be challenging and impractical [21,22]. The analogy of HAD to a child learning to understand the movements of people around them illustrates the essence of the concept. Similar to a child, HAD aims to perceive the distinctions between activities without prior knowledge of their concepts or labels [17]. Initially, it attempts to differentiate activities based on similarities and differences. The difficulty of HAD lies in the absence of labeled information to guide the learning process about human activities. While unsupervised methods have been proposed for HAR, they typically assume well-segmented videos with a single activity sample per segment and knowledge of the number of performed activities [17]. However, in reality, received videos are often unsegmented, and there is no information available about the number of performed activities. To address these challenges, we propose the utilization of the sliding window technique to handle unsegmented videos. Furthermore, we implement an incremental technique for multi-objective clustering to automatically determine the number of clusters. These techniques enhance the accuracy of HAD in HAR systems, facilitating more efficient recognition of human activities.

In the learning step of the HAR system (Fig. 1(d)), a learning algorithm is applied to the clustered activities to create a model that accurately classifies each discovered activity. This model is then used to classify new input activities (Fig. 1(e)). While supervised methods have made significant progress in learning and modeling human activities, such as dynamic time warping [23], hidden Markov models [24], and CNN [25,26] and RNN based approaches [10,27–29], these methods rely on ground truth labels and do not address the challenge of HAD. In real-world scenarios, label information is often unavailable, which significantly reduces the efficiency of supervised methods. Training supervised models requires extensive labeled data, which is time-consuming and not scalable due to the limited availability of training data. Furthermore, the model struggles to identify activities performed by different people and new activities that have not been trained on. This problem hinders the performance of HAR. However, unsupervised methods do not require understanding or labeling of input data, making them more feasible than supervised methods [30]. They recognize activities based on received patterns, and their scalability allows them to detect new activities based on received patterns, making them useful in various settings. It is worth to mention that the focus of this paper is on the discovery of human activities, which is shown in Fig. 1(a) to (c).

In this paper, a novel, Flexible Multi-Objective Particle swarm optimization clustering based on Game theory (FMOPG) is proposed to perform the HAD fully unsupervised on 3D skeleton data. The main contributions are summarized as follows: 1) A flexible multi-objective clustering algorithm is proposed as part of a framework for human activity discovery. This algorithm, based on Particle Swarm Optimization (PSO) and game theory, along with an incremental technique, is designed to estimate the number of activities and discover them accurately and efficiently, fully unsupervised. 2) To improve the algorithm's performance in discovering human activities, a new smart grid-based swarm initialization method is introduced. This method enhances the discovery process by providing better initial positions for the clusters, leading to improved convergence rates and more optimal solutions, ultimately aiding in the accurate identification of diverse human activities. 3) Additionally, a method based on mean shift clustering is introduced to update particle velocities. This approach addresses challenges such as local optima and premature convergence, specifically in the context of human activity discovery. By preventing the algorithm from getting trapped in local minima and being more adaptable to different human activity patterns, this method enhances the algorithm's effectiveness in accurately identifying and classifying human activities.

The remainder of this paper is structured as follows: in Section 2, the research background and relevant methods for human activity discovery are reviewed. The proposed approach is described in Section 3. In Section 4, the proposed method is evaluated and compared with other methods, and finally, in Section 5, the conclusion is stated.

2. Related work

2.1. Automatic clustering

The number of clusters cannot be predetermined in real-world data clustering analysis, and establishing the appropriate number of clusters for a huge and complex dataset is a tricky process. Many methods have been proposed to perform multi-objective clustering automatically. In multi-objective clustering with automatic K determination (MOCK) [31], non-convex solutions were removed from the pareto front. Then, the knee was determined as the number of clusters based on gap

statistics. In [32] the number of clusters was determined from the best solution by performing the algorithm from predefined maximum to minimum numbers of clusters. In [33] multiple clustering validity indexes were employed to find the optimal number of clusters and the best possible solution. The proposed method went a step further and instead of using validity index methods for estimating the number of clusters, which is time-consuming and problem-specific, an incremental technique was employed, which like human learning, gradually identifies and categorizes data.

2.2. Initialization methods

Initialization is one of the most influential factors in clustering and population-based algorithms. One of the common methods is Forgy's method [34]. There is no theoretical basis in this method and data points are selected as centroids randomly. In the initialization method of k-means++ [35], one data point is selected randomly as a centroid. The distance of all data points from the selected centroid is computed, and a data point based on maximum probability is selected. The further the data point is from selected centroids, the higher the chances of being selected as a centroid. However, all of the methods mentioned above perform initialization in a random point, which makes these methods prone to local minima. For this reason, Bajer al el. [36] employed a clustering approach to find the potential area and used Cauchy mutation to generate individuals from each selected area. Although they enhance the initialization, their method suffers from high complexity and is time-consuming in high-dimensional data. The main difference between our method and the previous method is that our method does not start from a random point, and its complexity is low. The proposed method divides the feature space into the $k \times k$ hyper square grid cell and selects the cells with the largest population according to the number of clusters. Centroids are randomly selected from those cells.

2.3. Human activity recognition and discovery

HAR is one of the most challenging topics in computer vision. Most methods introduced for HAR are supervised focusing in Fig. 1 part (d) and (e), training from labeled data. [37] described the relation between skeleton joints in groups activity recognition, they used deep reinforcement learning and formulate joints as a Markov model to select informative ones [38], continuously learned from skeleton activity using a brain-inspired elastic network [39], and a hybrid CNN-LSTM to extract spatial and temporal features [40]. Miao et al. [41] introduced a novel graph convolution operator that collected dynamic gradient information linked to local motion between the central joint and its associated adjacent joints for feature aggregation. Furthermore, because adjacency matrices are typically time-consuming, features were aggregated nearby the central joint using a simple graph shift operation and point-wise convolutions. In [42] a method was proposed for encoding the spatio-temporal information of a 3D skeleton. The joints of each sequence were first coded into three-dimensional colored dots and then projected onto three orthogonal planes. Three CNNs were used to extract features from each image. The final classification of the presented sequence was obtained by combining the CNN scores. Zhou et al. [43] proposed a method for learning a visual pose model and a pose lexicon model. Their method is made up of two-level hidden Markov models. On one level, the alignment between the visual poses and semantic poses was represented, while on the other level, a visual pose sequence was represented as a Gaussian mixture. Then, activities were classified by formulating the classification problem based on the acquired lexicon. Li et al. [44] introduced a scheme to represent relationship between skeleton joints. They used sub-network to represent spatial and temporal skeleton joint connectivity graph with a frame attention model and LSTM network respectively. Gao et al. [45] presented a graph regression to extract spatial and temporal features. They combined graph regression with graph convolutional network to deal with data that is not arranged in normal graph. Koniusz et al. [46] proposed a novel feature representation to capture importance interaction between visual information. They designed dynamic compatibility kernels to build spatial and temporal relationship among features. The performance of these methods was strongly dependent on training data labeled with ground truth. In our proposed method, no labeled data is used to categorize activities. It automatically categorizes activities based on differences and similarities.

A branch of works tries to explore HAR in an unsupervised manner. Mohammadzade et al. [47] projected temporal and spatial features into the low dimensional unwarped space by using a joint learning strategy. Yang et al. [14] designed a skeleton cloud colorization scheme that encodes spatial, temporal, and person-level information. Su et al. [21] used an autoencoder to extract the features. To classify activities, encoded features were given to the KNN classifier. Zheng et al. [48] extracted deep features for classifying actions. They introduced a GAN autoencoder and extracted the dynamic motions from skeleton frames. Tang et al. [49] developed a new graph convolutional network to transfer the knowledge between data with different distributions. They fed the network with the source domain and target domain to extract features. The extracted feature from the source domain was used to train the label predictor under the supervision of source labels. After reducing the domain shift between the two domains, two weight matrices were fed into the relevant domain classifier. The issue with these approaches is that unsupervised learning was only used for feature extraction and ground truth was needed to categorize activities with supervised methods. Guo et al. [50] proposed a semi-supervised method that can recognize activities with a few examples to deal with scarce training data. They used an interactive graph structure to generate the representation of activities and match them with a few frames. Liu et al. [51] proposed a framework for one shot activity recognition. They examined the semantic significance between the activities and each body component based on their descriptions. Then, the framework stresses the important body parts for each class of activities for representation. The difference between these methods and the method of this paper is that they use a small amount of labeled data to categorize activities, but our method does not require any labeled data. Generally, the task of our proposed method is to do HAR fully unsupervised without using ground truth in any part of the HAR system.

Moreover, despite the impressive accuracy demonstrated by self-supervised models [52–54], there are compelling reasons to explore unsupervised approaches in the realm of HAR. The primary motivation arises from the inherent challenges associated with obtaining and utilizing labeled data for training self-supervised models. In the context of HAR, collecting labeled data for human activities can be a time-consuming, expensive, and intricate process, especially in real-world scenarios. Unsupervised methods, in contrast, present a notable advantage by not relying on labeled data, offering adaptability and feasibility when labeled data is limited or entirely unavailable. The embrace of unsupervised approaches in HAR, exemplified by the FMOPG method, is rooted in a recognition of the hurdles associated with obtaining labeled data and the imperative for flexibility in effectively uncovering and clustering human activities within diverse and dynamic environments. Despite the considerable strides made in the accuracy of Human Activity Recognition through self-supervised models, FMOPG takes on the distinct challenges tied to labeled data acquisition. This method introduces a fully unsupervised paradigm for HAD, particularly tailored for 3D skeleton data. What sets FMOPG apart is its capacity to cluster human activities without reliance on prior knowledge, positioning it as a competitive and superior alternative to existing algorithms. Notably, FMOPG stands out by its autonomous operation, obviating the need for human intervention in the training process.

A limited number of research attempted to address HAD using a fully unsupervised approach. An early HAD literature used incremental k-means clustering and a hand-crafted feature method to discover activities [55]. They also built a model for activities by using the mixture of

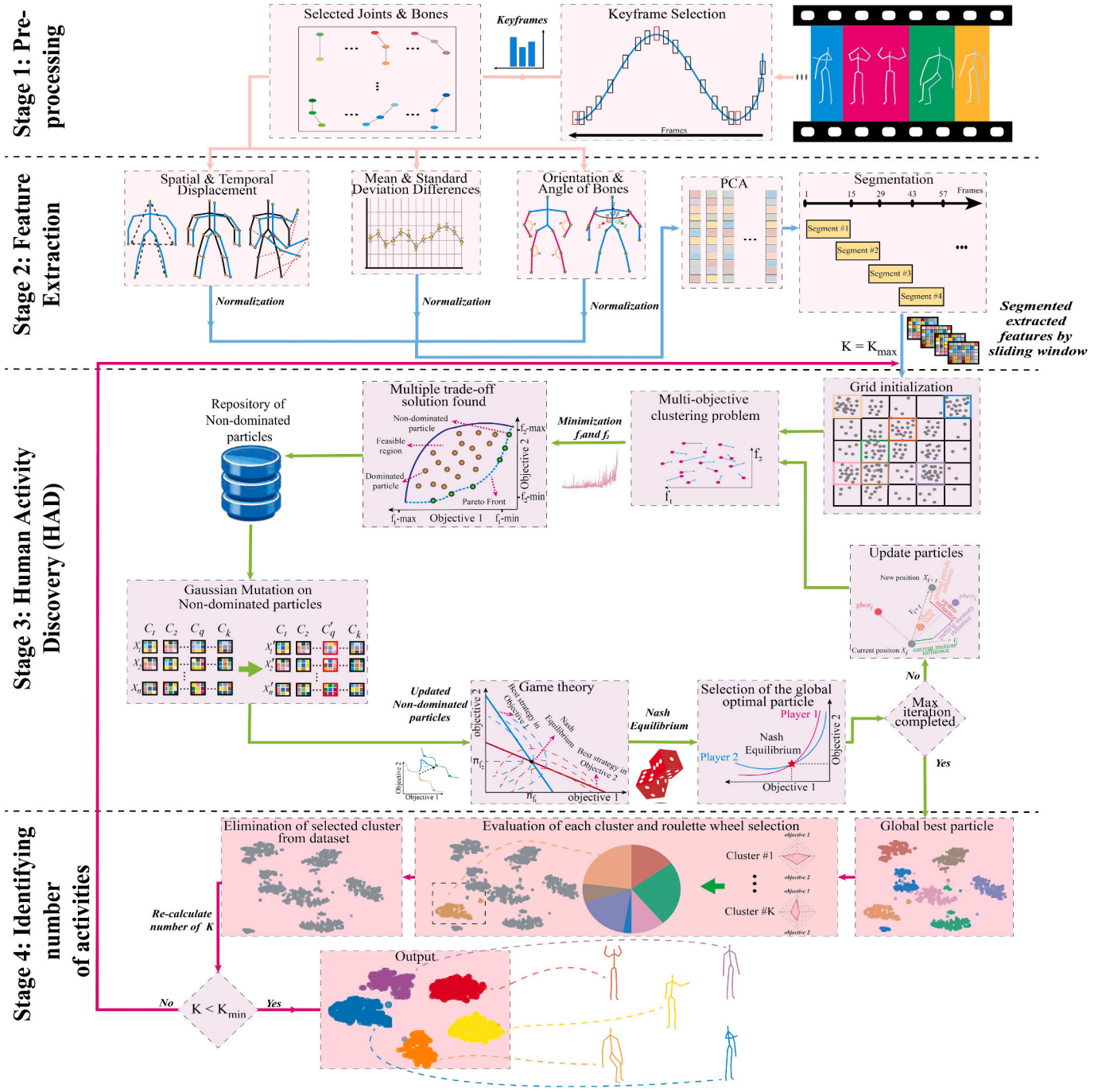

Fig. 2. This system consists of four main stages. In stage1, human activity videos are processed to select keyframes and informative joints and bones. In stage2, three methods are used to extract relevant information and features from the activity data. Dimensionality reduction is achieved through PCA to enhance computational efficiency, and the frames are partitioned into overlapping fixed-size activity instances. The HAD stage involves clustering the instances. Starting with a maximum number of clusters (k_{max}) and iterates to a minimum number of clusters (k_{min}). A smart grid-based strategy is used to generate diverse solutions for initial centroids. After initialization, non-dominated solutions are obtained based on the assessment of each solution using the objective functions. To prevent getting trapped in local optima, Gaussian mutation is applied to non-dominated solutions, and Nash equilibrium is used to identify the global best solution. The mean-shift vector is used to update particles' positions and velocities to address premature convergence and non-linear clusters. The process of finding non-dominated solutions and updating particles continues until the maximum iteration limit is reached. To identify the number of clusters in last stage, the clusters of the global best solution are evaluated using objective functions, and the best cluster is selected through roulette wheel selection, provided it meets the minimum cluster size requirement. This selected cluster is considered a final cluster, and its members are removed from the dataset. The number of clusters is recalculated based on the remaining dataset, and the HAD process is repeated.

the Gaussian Hidden Markov Model. These methods used all joints from all frames to extract the feature. Noisy data and outliers were not considered. Paoletti et al. [56] used subspace clustering to discover human activities. To reduce the number of redundant frames, they introduced a trim method. Even though their approach has achieved promising results, they cannot handle complex scenarios because their inputs were well-segmented videos, and they required prior knowledge about the number of performed activities.

Hadikhani et al. [17] presented a clustering method based on PSO that received unsegmented frames as inputs and clustered them. They presented feature extraction from informative skeleton joints by combining different techniques. They employed a kinetic energy-based method to solve the redundant data problem and extract effective frames. Most of these works used a single objective to cluster activities and did not achieve good results for all datasets. A single objective does not operate equally well for all datasets due to differences in data features. In contrast, our approach of optimizing multiple objectives, such as symmetry and connectivity, simultaneously enables us to better capture the unique properties of each dataset. By doing so, we can more effectively distinguish between different human activities and elucidate the poses between them.

To solve HAD in a fully unsupervised way, a multi-objective clustering approach with a cluster validity index to estimate the number of activities was presented in [57]. In addition, a game theory method was applied to select the best solution in the multi-objective problem. Different from [57], in this paper, an efficient incremental technique is applied to multi-objective clustering to automatically detect the appropriate clusters instead of identifying the number of clusters through an exhaustive iterative process from maximum to minimum k values. In other words, for each number of clusters in [57], all data points are clustered together regardless of the quality of each cluster. However, in our proposed method, the best clusters are gradually selected during the clustering process by examining the quality of each cluster and considering the conditions of the data points in the selected cluster. Moreover, our method introduces a novel smart grid-based swarm initialization method, which enhances the discovery process by providing better initial positions for clusters. This improvement leads to enhanced convergence rates and more optimal solutions, ultimately aiding in the accurate identification of diverse human activities. Previous methods may have used more traditional initialization approaches, such as random or uniform initialization, which may not be as effective in achieving optimal solutions. Additionally, we propose a method based on mean shift clustering to update particle velocities. This approach specifically addresses challenges such as local optima and premature convergence, which are common in the context of human activity discovery. By applying mean shift clustering to update particle velocities in the context of human activity discovery, the algorithm focuses on adjusting the representative poses (cluster centers) of each activity. This adjustment is based on the differences between each set of poses (data points) belonging to that activity and its representative pose. Mean shift clustering essentially corrects or adjusts the cluster centers to better represent the poses associated with each activity. This adjustment process helps prevent the algorithm from getting stuck in local optima and allows it to adapt more effectively to different human activity patterns. By continuously updating the cluster centers based on the poses of the data points, mean shift clustering helps the algorithm converge towards better representations of the underlying activities, thus enhancing its ability to distinguish between different human activities. Previous methods have used normal velocity update strategies, which may not have been as effective in overcoming these challenges.

3. Proposed human activity discovery

The proposed Flexible Multi-Objective PSO clustering with the integration of game theory (FMOPG) is a method to discover human activities and is illustrated in Fig. 2. It involves three key steps: preprocessing input, clustering activity samples, and detecting the number of activities (clusters).

In the prepossessing step (a full description is provided in supplementary), input frames containing data of human skeletal joints in 3D space are received. Keyframes are selected based on the kinetic energy of the joints. This selection process helps to reduce the amount of data that needs to be processed and analyzed, as keyframes represent the most informative frames in the sequence. By selecting keyframes, the computational resources required for the analysis are reduced, which can improve the speed and efficiency of the algorithm. Additionally, keyframes can provide more meaningful information about the activity being performed, as they capture the most important and distinctive poses and movements of the human body. In order to extract relevant features from the human skeletal joints, informative joints are selected from important parts of the body. These joints are selected based on their importance in capturing the motion and posture of the body during various activities. For example, joints such as the shoulders, elbows, and hips are often more informative than other joints like the fingers or toes. By selecting the informative joints, the feature extraction process can be more efficient and effective, as it focuses on the most relevant joints for the task at hand. To improve the performance of activity discovery, it is essential to extract features that can represent all aspects of each activity. In this study, we utilize a feature extraction method based on [17], which includes spatial and temporal displacement, statistical, and orientation features. Displacement-based representations are used to provide spatio-temporal human representations that are view-invariant and independent of the position and orientation of the person in relation to the camera. Orientation-based representations are employed to obtain features that are invariant to human scale changes and can find relative information between human joints. Statistical features are used to describe how an activity evolves over time and can distinguish between the actions of the arms and legs. We apply PCA to reduce the dimensionality of the extracted features, while still maintaining the high-importance features. PCA have chosen due to its effectiveness in capturing essential information while discarding redundant features, which is crucial for enhancing computational efficiency and avoiding overfitting.

At the end of the preprocessing step, a sliding window approach is used over the stream of skeleton sequences to facilitate activity discovery. Overlapping sliding windows are employed to increase the number of samples and avoid pruning important events, such as transitions between activities. This approach allows for a more comprehensive representation of the activity instances and can lead to more accurate clustering results.

Our clustering methodology utilized the PSO algorithm as its core and is further refined through several modifications aimed at optimizing its performance. PSO is advantageous for clustering because it can optimize non-convex objectives that traditional techniques struggle with. PSO is efficient at searching the solution space due to its population-based approach, with each particle representing a potential clustering solution. The swarm updates particles based on their own and the best-performing particle's experience, enabling efficient searching of the solution space. The clustering process starts with a smart grid-based strategy to generate diverse solutions. This is important because a diverse set of solutions helps to avoid getting stuck in local optima and improves the overall performance of the clustering algorithm. The smart grid-based strategy generates initial centroids for the clustering algorithm in a way that maximizes the coverage of the search space. By doing so, it increases the chances of finding better solutions and makes the clustering process more efficient. To assess each solution, multi-objectives based on Eq.(5 and 6) are utilized to obtain non-dominated solutions. Using multiple objective functions in a clustering algorithm provides a more comprehensive evaluation of the quality of the solutions compared to a single objective. In this proposed method, the objective functions are used to measure both the compactness and diversity of the clusters. The compactness objective function aims to minimize the intra-

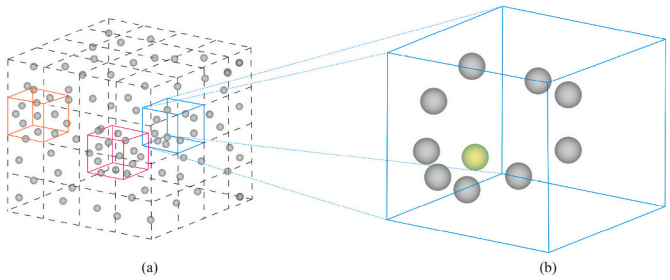

Fig. 3. Population initialization using smart grid-based. Feature space is divided into the $k \times k$ hyper square grid cell where k is the current number of clusters. (a) k grid cells with large population are selected and (b) a data point is randomly selected as a cluster center from each selected cell.

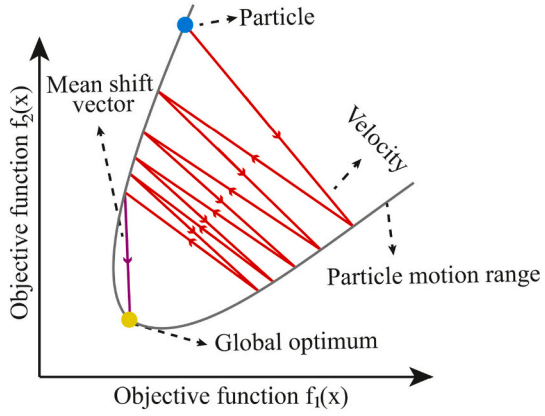

Fig. 4. Illustration of the motion of a particle that suffers from slowing down convergence as it approaches the global best solution and the effect of mean-shift vector in mitigating this problem.

cluster distance, ensuring that the members of a cluster are similar to each other. The diversity objective function aims to maximize the inter-cluster distance, ensuring that the clusters are dissimilar to each other. By optimizing both objectives simultaneously, the proposed algorithm can produce clusters that are both compact and diverse, which is desirable for discovering human activities. Gaussian mutation is then applied to the non-dominated solutions to avoid the local optimum trap and promote exploration of the solution space. To determine the global best solution, Nash equilibrium is employed. This ensures that the selected solution is the best possible trade-off between the multiple objectives and is not dominated by any other solution (the results of selecting the global best solution based on game theory is reported in Fig. 2 in supplementary). The mean-shift vector is used to update the position and velocity of particles in the FMOPG clustering method. This technique is used to deal with premature convergence and non-linear clusters, which can be problematic in traditional clustering methods. The mean-shift vector calculates the direction and magnitude of the shift needed for each particle to move closer to the global optimum. In the last step, the incremental technique is utilized for determining the number of clusters. It involves repeating the steps from finding non-dominated solutions to updating particles until the maximum iterations are achieved. Global best clusters are evaluated using both objective functions, and the best one is selected using roulette wheel selection. Roulette wheel selection provides a probabilistic approach to selecting the best cluster. This means that the selection is based on a probability proportional to the fitness of each cluster. In other words, the better the cluster's fitness, the higher the probability of it being selected. If the selected cluster meets the minimum member condition, it is considered a final cluster and its members are removed from the dataset. The dataset is then trimmed, and the number of clusters is recalculated based on its size. The clustering process is repeated based on the new number of clusters.The pseudocode of FMOPG is presented in Algorithm 1. In the subsequent subsections, we provide a detailed description of each component of the system.

Algorithm 1. FMOPG algorithm.

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

Output: A set of clusters O=$\{c_1,c_2,\ldots,c_k\}$, k detected clusters

$K_{min} \leftarrow 2$
$K_{max} = \sqrt{size(D)}$
$K = K_{max}$
MinPt=$\frac{size(D)}{K_{max}}$
while $K > K_{min}$ **do**
  Initilize the particle with Smart grid-based initialization algorithm (Algorithm 1 in supplementary)
  Calculate the fitness of each particle using Eq. (5) and (6)
  Set the initial personal best of each particle
  Repository ←non-dominated solutions
  Global best ←none
  iter ← 0, MaxIteration ← 50
  while *iter* < *MaxIteration* **do**
    for *each solution in Repository* **do**
      Calculate the $NashE_j$ for non-dominated solutions
    Global best ← find a particle with the best $NashE_j$
    Update particle by the mean-shift vector (Algorithm 2 in supplementary)
    if *Size (repository Size)* > $\frac{Numberof Particles}{2}$ **then**
      Until Size (repository Size) > $\frac{Numberof Particles}{2}$, delete one of the repository members based on "roulette wheel selection"
    for *T times* **do**
      Mutate solutions in repository according to Eq.(11) and (12) in supplementary
      Compare mutated version with previous one and choose non-dominated one
    iter ← iter+1
  Detect one cluster from Global best solution with incremental technique(Algorithm 3 in supplementary)
Choose the final the remaining unclustered datapoints as last centroid and include it in the set of clusters that named C.
Return C

3.1. Smart grid-based initialization

Initialization of particles has a significant task that influences on diversity and convergence. For this reason, a new smart grid-based initialization method is proposed. This method can detect density

areas that are suitable area for initialization and producing diverse solutions. First, the lower and upper bounds of the dataset are computed. This is done by finding the minimum and maximum values in each dimension of the dataset. This step is crucial for determining the range of values in the dataset, which is used to divide the feature space into a grid of $k \times k$ hyper square grid cells. Each data point is assigned to a grid cell according to its coordinates. If a data point is on the edge of grid cells, it is assigned to the cell on the top right. After assigning the data points to relevant grid cells, k grid cells with the most number of data points are selected. From each of the k grid cells, a data point is randomly selected as a cluster center. Finally, all the cluster centers are placed in one particle. The process of initialization is shown in Fig. 3 (see Algorithm 1 in the supplementary).

3.2. Updating particles by adopting the mean-shift vector

Although PSO has a good ability to find a global area in the search space, it slows down as it approaches the global solution. As shown in Fig. 4, PSO in its early stages moves with large steps towards the global optimum. However, over time the search process slows down dramatically as it becomes closer to the final solution. On the other hand, Mean-shift clustering has a rapid convergence in finding the local optimal. It is also able to find clusters of any size and shape due to its kernel function [58]. Therefore, we adopt Mean-shift clustering in PSO and update each particle based on the concept of mean-shift vector. In this way, before updating the particles, the radius of the clusters (σ) in the particles is calculated and considered as a bandwidth parameter in the mean-shift vector. To compute shifted centroid $M(x)$, the weight of nearby data points is determined by kernel function $kf(x)$ in Eq. (1). The $M(x)$ in the obtained radius is determined by Eq.(2).

$$kf(x) = \frac{1}{N}\sum_{i=1}^{N} exp\left(\frac{centroid - d_i}{\sigma}\right) \tag{1}$$

$$M(x) = \sum_{i=1}^{N} \frac{kf(x)}{\sum_{i=1}^{N} kf(x)} \times x_i \tag{2}$$

where N is the number of data points in the cluster with the *centroid* and d_i is the i_{th} data point in cluster. After obtaining $M(x)$ for each centroid, particle velocity update equation in PSO (Eq.(3) in supplementary) is modified as follow:

$$\begin{aligned} v_i'(t+1) &= w \times v_i + c_1 \times rand_1 \times (pbest_i(t) - x_i(t)) \\ &+ c_2 \times rand_2 \times (gbest(t) - x_i(t)) + (M(x) - x_i(t)) \end{aligned} \tag{3}$$

In Eq.(3), the shifted vector $(M(x) - x_i(t))$ is calculated, which specifies in which direction and to what extent the particle should move towards the global solution. The pesudo-code of the particle update process is given in Algorithm 2 in the supplementary.

3.3. Detecting clusters automatically with incremental approach

In the real scenario, the number of activities is not known in the input videos. To tackle this, we present an incremental technique for multi-objective clustering. In this technique, unlike conventional methods that use cluster validity to estimate the number of clusters and are time consuming, not all data are clustered at once like in Fig. 5(b). Instead, clusters are detected over time as shown in Fig. 5(a). As summarized in Algorithm 3 in the supplementary, for each value of k, clustering is done on the dataset. After obtaining the global best particle, each cluster is evaluated based on both objective functions. The plane is mapped into a grid with equal units and the centroids are placed in the cells based on their fitness values. Using the roulette method, a cell is selected. If more than one centroid is in the selected grid, one centroid is randomly selected. To guarantee that the clusters have enough observations to represent the activities, the selected cluster should have a minimum membership parameter (MinPt) which is set up based on $\frac{n}{k_{max}}$. If the number of members in the selected cluster meets the MinPt constraint, the cluster is chosen as one of the final clusters, and its members are removed from the dataset. A new value for k is calculated from new population n^* of the trimmed dataset based on $\sqrt{n^*}$. Otherwise, it is ignored and the value of k is decremented. This process continues until the value of k reaches $k = 2$. Finally, the selected clusters are aggregated during the clustering process and are considered as the final result.

4. Experiments

Seven challenging datasets including Cornell Activity Dataset (CAD 60) [59], Kinect Activity Recognition Dataset (Kard) [60], MSR DailyActivity3D (MSR) [61], UTKinect-Action3D (UTK) [62], Florence3D (f3D) [63], NTU RGB + D Dataset (NTU-60) [64] and NTU RGB + D 120

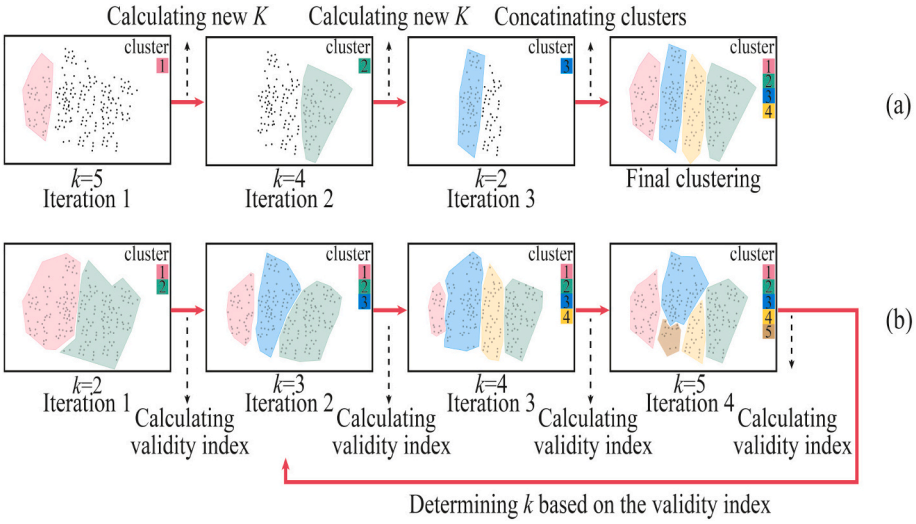

Fig. 5. Illustration of (a) incremental technique in comparison with (b) cluster validity index approach. The long red line in (b) represents the lowest value of the validity index corresponds to the optimal number of clusters, which is then selected as the final number of clusters (k) for the data. However, in the incremental technique shown in (a), clusters are detected and formed gradually over time, without the need for explicitly computing and comparing cluster validity indexes for different k values. This approach avoids the extra computation required in the cluster validity index approach. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Details of seven datasets used for evaluation.

Dataset	CAD 60	KARD	MSR	UTK	F3D	NTU-60	NTU-120
Activity	14	18	16	10	9	60	120
Subjects	4	10	10	10	10	40	106
Videos	60	2160	320	200	215	56,880	114,480

(NTU-120) [51] were used to evaluate the proposed method. These datasets are challenging as they contain large-scale performed activities, high overlap between activities and many disruptions in capturing activities. We evaluated the performance of the proposed method based on these challenges. The details of each dataset are given in Table 1. To show the effectiveness, we compared the proposed method with non-automatic methods (having prior knowledge of the number of clusters) including KM (k-means), SC (spectral clustering) and SSC (Sparse Subspace Clustering) [56] and automatic methods including DBSCAN, MS (Mean-shift clustering), PSO [65], HPGMK [17], and MOPGMGT [57]. We have selected KM, SC, DBSCAN and GMM as conventional clustering methods. To compare with the State-Of-The-Art (SOTA) methodologies in HAD, SSC and HPGMK as single objective and MOPGMGT as multiple objective clustering have been chosen. Additionally, due to the absence of recent methodologies in HAD, we leverage state-of-the-art (SOTA) clustering methods in the image domain, including Not Too Deep Clustering (N2D) [66], Deep Clustering Network (DCN) [67], Structural Deep Clustering Network (SDCN) [68] and incomplete multi-view clustering via contrastive prediction (Completer) [69] have been compared to validate the performance of FMOPG. Due to the use of PSO and MS in the structure of the proposed algorithm, we have chosen them to show that the proposed method has improved their performance. Moreover, we have compared FMOPG with prior related supervised methods applied to CAD-60 dataset including Dynamic Bayesian Network [59], Spatio-Temporal Interest Point [70], SVM + Hidden Markov Model [60], Atomic motion+naive Bayes+nearest neighbor [71], Bags of visual words+Fisher vectors+SVM [72] and Convolutional Neural Networks+SVM [73] to demonstrate the performance of our method against methods that have been trained and have prior knowledge about activities. To evaluate the performance of the incremental technique, which we proposed for determining the number of clusters, we compared it with cluster validity index methods including Silhouette Index (Sil) [74], Calinski-Harabasz (Ch) [75],Davies-Bouldin (Db) index [76], Dunn index [77], Gap statistic [78], Elbow [79], Hartigan (Ha) index [80], Krzanowski-Lai (Kl) [81], Slope [82] and Jump [83].

4.1. Evaluation measures

We compared clustering algorithms using accuracy based on [4] to investigate their performance across 30 runs. F-score and confusion matrix have been used to show the performance of each method in categorizing the activities and the confusion between activities. Moreover, the overall error (OE) of estimating the number of clusters was computed for each automatic method as given in Eq.(4).

$$Overallerror = \sqrt{\left(\sum_{i=1}^{n} k^i - k_p^i\right)} \tag{4}$$

Where k^i is the actual k, for subject i, k_p^i is the predicted k for subject i, and n is all subjects in the dataset.

4.2. Set parameters and implementation

The experiment was repeated 30 times. The number of iterations and the size of the population (particles) in methods PSO, HPGMK and MOPGMGT, FMOPG were equal to 20 and 50, respectively. The number of components in the GMM Algorithm was set to the number of clusters. Bandwidth in MS clustering was equal to 2. The minimum number of samples and maximum distance between two samples in DBSCAN were equal to 0.5 and 2, respectively. The initial value of parameters $c_{1_{max}}$ and $c_{2_{max}}$ in the proposed method were set to 2.5, $c_{1_{min}}$ and $c_{2_{min}}$ set to 0, and w_{max} set to 0.9. Stop criteria for all algorithms was based on the number of iteration. Since the videos have not been segmented in the proposed method, all compared methods in this paper received unsegmented videos for the same evaluation. All methods were applied on segmented videos produced by using the sliding window technique used in [17]. All automatic methods used the Jump method [83] to estimate the number of clusters. They were performed for different number of clusters in the range of $k=2$ to K_{max}. K_{max} was chosen based on $\sqrt{n}$ where n is the number of data points. For each value of k, the Jump value was calculated for them. Finally, the estimated number of clusters was determined based on the minimum value of Jump. The objective function SSE (sum square error) in Eq.(5) was used for all single and multi objectives clustering algorithms which should be minimized to achieve proper clustering.

$$SSE = \sum_{k=1}^{K} \sum_{\forall x \in c_k} \| x_i - \mu_k \|^2 \tag{5}$$

x_i is a data point belonging to the cluster c_k and μ_k is the mean of the cluster c_k. The second objective function for multi-objectives clustering algorithms was Conn-index [32] which should be minimized. it is calculated as follow:

$$Conn = \frac{\sum_{i=1}^{k} min \sum_{j=1}^{n} d\left(p_j^i, m_j\right)}{n\left(min_{i,j=1,i\neq j}^{k} d\left(m_i, m_j\right)\right)} \tag{6}$$

$$m_i = min_t^n \left(\sum_{i=1}^{k} d\left(p_t^i, p_j^i\right) \Big/ n \right) \tag{7}$$

where n is number of objects in cluster c_i and p_j^i is the jth object of cluster i.

4.3. Discussion and results

Fig. 6 compares the best, worst, and average accuracy obtained by the different methods (numerical results are reported in Table 1 in supplementary). As it is shown, except for the F3D dataset that SSC (knows the number of clusters) has achieved a better performance in terms of average accuracy, in other datasets, the proposed method has had a significant advantage over other methods. The overall average accuracy of the FMOPG was 80.99% for CAD-60, 54.70% for UTK, 55.76% for F3D, 40.34% for MSR, and 43.76% for KARD, 36.70% for NTU-60 and 19.82% for NTU-120. FMOPG has performed better among all automatic and non-automatic algorithms. Because other methods do not have a suitable technique for initialization of clusters to produce various solutions and improve exploration that makes the distance between the minimum and maximum accuracy in FMOPG is the lowest compared to other methods. While some methods may exhibit a narrower range between their maximum and minimum accuracy compared to FMOPG, their overall accuracy tends to be lower. This is often a result of these methods getting trapped in local optima, which restricts their capacity to explore the solution space effectively. Consequently, the maximum, minimum, and mean accuracy values of these methods tend to converge, indicating a struggle to achieve consistently high performance across different datasets. In contrast, FMOPG's ability to strike a balance between exploration and exploitation allows it to achieve higher overall accuracy, even if its range of accuracy values is broader. FMOPG struggle to achieve high accuracy in the NTU 120 dataset, as evidenced by its performance metrics. One of the primary reasons for its suboptimal performance could be attributed to the limitations of handcrafted features. These features may not be able to capture the complex and nuanced patterns present in the NTU 120 dataset, which includes a large

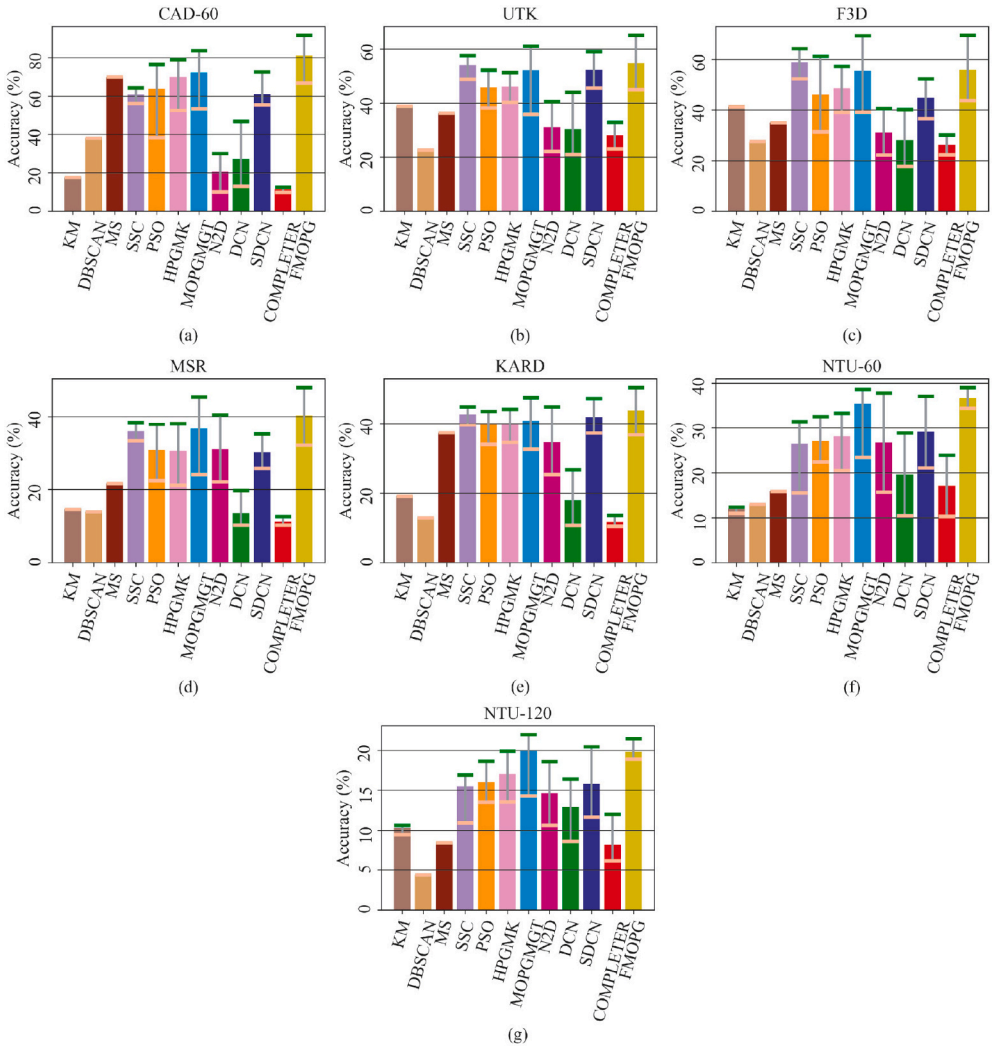

Fig. 6. The average accuracy of methods for all subjects in (a) CAD-60, (b) UTk, (c) F3D, (d) MSR (e) KARD, (f) NTU-60, (g) NTU-120.

Table 2
Comparison the results of state-of-the-art models proposed for HAR on CAD-60 with our method.

Types	Methods	Precision
Supervised	Sung 2012 [59]	67.9
	Zhu 2014 [70]	93.2
	Gaglio 2014 [60]	77.3
	Cippitelli 2016 [71]	93.9
	Seddik 2017 [72]	92.4
	Boualia 2021 [73]	95.4
Unsupervised	HPGMK [17]	73.32
	MOPGNGT [57]	77.28
	FMOPG (Ours)	80.55

number of activities, subjects, and videos. Handcrafted features are designed based on human expertise and may not fully exploit the rich information available in the dataset. To improve the accuracy of activity recognition in such a challenging dataset, it may be necessary to explore more advanced feature extraction methods, such as deep learning. Deep learning approaches have shown promising results in various computer vision tasks, including activity recognition, by automatically learning relevant features from the data. By leveraging the power of deep learning, it may be possible to extract more discriminative features that can better represent the activities in the NTU 120 dataset, ultimately leading to improved accuracy.

FMOPG also used the concept of MS clustering in refining the particle velocity update to deal with early convergence [58]. It enabled PSO's ability to have a better search around discovered solutions and prevented particles from slowing down to reach the final solution. That is why better results have been obtained compared to other algorithms. The integration of the incremental technique to find the clusters dynamically has caused, instead of estimating the number of clusters, FMOPG finds suitable clusters with good connectedness and separation. This has improved the accuracy results compared to methods that use the cluster validity method to estimate the number of clusters.

In deep clustering methods including N2D, DCN, SDCN and COMPLETER, despite using deep learning to better represent the activities, they could not obtain good results because the used networks in these methods are not able to extract spatial and temporal features that are very effective to discover activities. They also use shallow clustering methods, such as k-means that easily get stuck in the local optimization. In terms of subspace clustering including SC, ENSC, and SSC algorithms, knowing the number of clusters, they did not achieve good results. These methods are in lack of appropriate strategy to find clusters and make a good balance between exploration and exploitation [84,85]. Density-based algorithms also did not achieve good results due to their strong dependency on adjusted parameters. These algorithms easily get stuck in local optimization because they do not have an alternative strategy for exploring the search space. For the rest of the comparisons, PSO, HPGMK, MOPGMGT and FMOPG have been used because clustering is

Table 3
Comparison of the minimum clustering time by different approaches for each dataset.

Method	Dataset	CAD 60	UTK	F3D	MSR	KARD	NTU-60	NTU-120
PSO		161.77	6.28	10.618	58.14	66.78	608.154	1416.67
HPGMK		174.99	7.30	13.026	62.94	74.12	622.94	1478.048
MOPGMGT		701.869	13.43	17.71	109.06	170.22	1062.25	2523.60
FMOPG		168.26	10.30	7.90	66.31	40.60	498.58	995.16

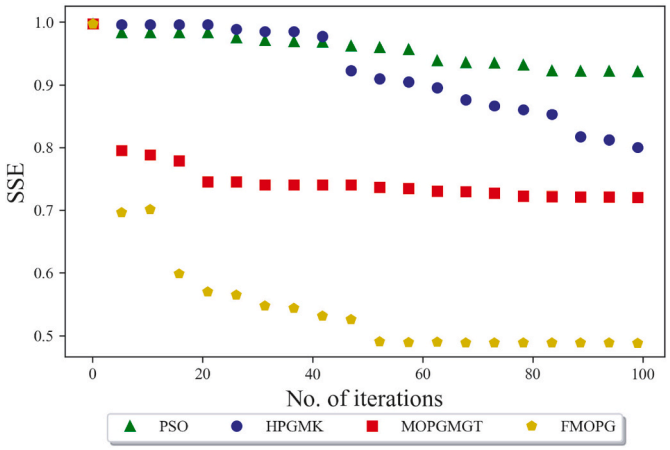

Fig. 7. Comparison of the convergence of the four methods PSO, HPGMK, MOPGMGT, and FMOPG.

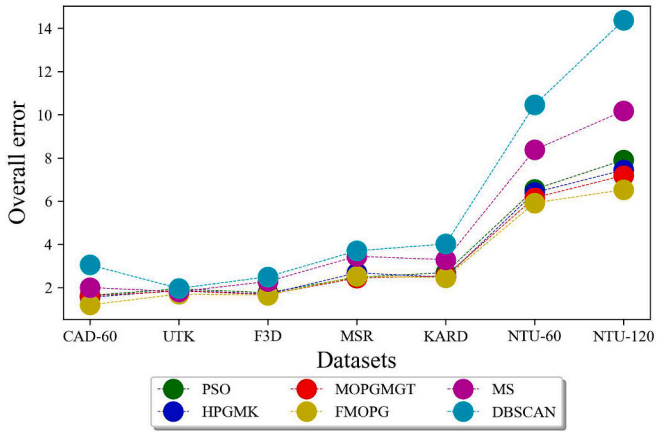

Fig. 8. Comparison of average overall error for detecting the number of clusters by methods MS, DBSCAN, PSO, HPGMK, MOPGMGT, and FMOPG for all datasets.

done automatically in these methods. The numerical results of Fig. 6 have been presented in the supplementary.

Table 2 shows a comparison of FMOPG performance with supervised and unsupervised methods based on precision for CAD-60. As can be seen, FMOPG is superior to unsupervised methods. However, compared to supervised methods, in spite of outperforming some supervised method and and obtaining acceptable result by FMOPG, it was not able to obtain the best results. Because, supervised methods use ground truth and are trained using a large number of labeled activities. In other words, these methods are already aware of the number of activities and the class of activities. But in our method, ground truth is not used and there is no information about the number of activities.

Table 3 shows a comparison of clustering time in seconds (s) between methods PSO, HPGMK, MOPGMGT, and FMOPG. This demonstrates that using the incremental technique is very effective for reducing clustering time. The time spent for clustering by FMOPG in five datasets CAD-60, UTK, F3D, MSR, KARD, NTU-60 and NTU-120 were equal to 168.26 *s*, 10.30*s*, 7.90*s*, 66.31 *s*, 40.60*s*, 498.58 *s* and 995.16 *s*, respectively. Although FMOPG has to optimize two functions at the same time, in the F3D, Kard, NTU-60 and NTU-120 it took less clustering time than single-objective PSO and HPGMK. Moreover, FMOPG's higher complexity compared to PSO and HPGMK sometimes results in these methods having higher clustering times in certain datasets. However, the use of the incremental technique enables FMOPG to achieve clustering times comparable to or even better than these single-objective methods. This indicates that not only FMOPG has solved the problem of finding clusters, but also demonstrates the effectiveness of the incremental technique in reducing clustering time and making FMOPG computationally competitive with single-objective methods.

The convergence of the four methods PSO, HPGMK, MOPGMGT, and FMOPG has been investigated in Fig. 7. Apart from the fact that the use of smart grid-based initialization has led to better convergence compared to other methods in the early stages, the proposed method has avoided premature convergence in the final stages. Because in mean-

Table 5
Impact of PCA and Keyframes on the performance of HAD and time of discovery in MSR.

Method	Performance	
	AC	Time (second)
FMOPG without PCA and Keyframes	41.70	371.39
PCA	46.23	203.48
Keyframes	44.07	187.59
PAC + Keyframes	48.02	90.35

Table 4
Comparison of total error for the obtained number of clusters by different approaches for each dataset by the different approaches.

Method	Dataset	CAD 60	UTK	F3D	MSR	KARD	NTU-60	NTU-120
Sil		2.7	4.61	4.32	4.93	6.87	7.23	7.63
Ch		2.46	4.72	4.47	5.27	6.68	6.99	8.32
DB		2.42	4.26	2.75	4.75	6.46	7.12	8.02
Dunn		4.75	3.79	3.36	4.39	6.38	9.60	9.68
Gap		4.89	5.02	4.70	5.24	7.11	9.42	10.60
Elbow		3.76	3.63	3.42	3.17	5.60	7.68	9.62
Ha		2.02	2.46	3.59	2.25	3.75	5.93	7.73
Kl		2.42	3.10	2.48	1.94	3.76	6.95	7.35
Slope		2.54	3.94	3.80	4.21	6.62	7.24	8.25
Jump		2.30	2.46	2.86	2.77	4.11	7.14	7.22
Incremental		1.27	1.70	1.65	2.52	2.46	5.92	6.53

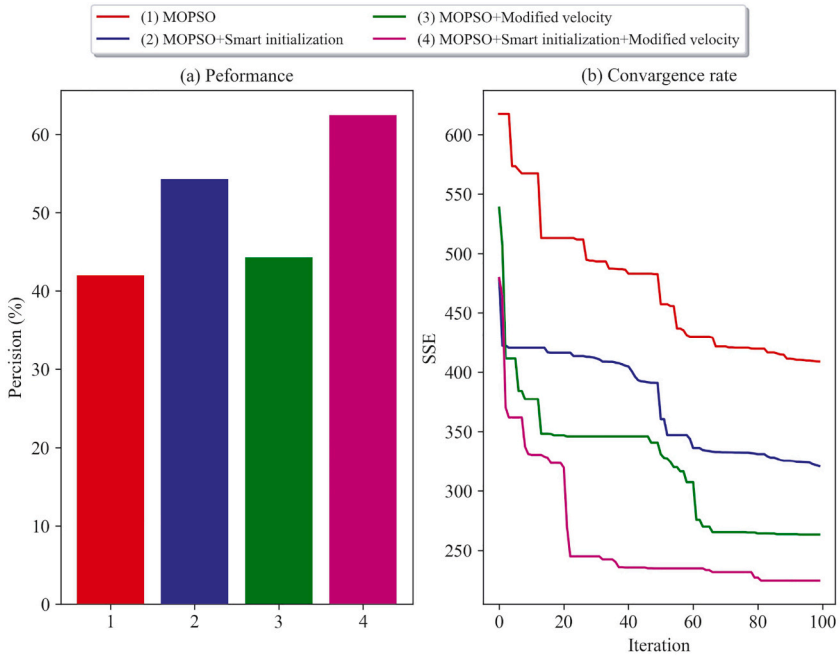

Fig. 9. Effect of Meanshift and Smart initialization on the performance of the FMOPG on subject one in F3D.

shift vector, the kernel function transforms the data into a higher-dimensional space in which they are separable and enables the proposed method to deal with complicated clusters with non-convex shapes. However, PSO and MOPGMGT did not achieve the best possible solution due to the lack of a suitable strategy to deal with non-linear cluster. Regarding HPGMK, although the combination of PSO with KM has prevented the occurrence of premature convergence in the final stages, the converged value of HPGMK is not as good as the FMOPG because of its single objective approach.

Table 4 shows the total error for obtained number of clusters by the different approaches that applied in FMOPG. In all datasets, the incremental approach outperformed the other techniques in determining the number of clusters. In this technique, unlike cluster validity index methods that find all cluster in each iteration, clusters are found gradually based on how much the results have improved. However, cluster validity index methods not only have not performed well in detecting the number of clusters and have high overall error, but the method that works better than others, in one dataset, works worse other datasets. This indicates that their results are not stable.

Fig. 8 shows the average OE of the clusters found in the proposed method compared with the number of clusters estimated in MS, DBSCAN, PSO, HPGMK, MOPGMGT. From the results, it is obvious that the proposed method was able to find and cluster the activities better than the others. For this reason, it has the lowest OE compared to density-based clustering and other methods that have used the Jump method to estimate the number of clusters. It is also important to note that due to the use of PSO as the core of their method, PSO, HPGMK, MOPGMGT and FMOPG obtained approximate average errors. But in general, FMOPG has less error than these methods because it has modified the search process of PSO algorithm using the concept of Mean-Shift (MS) clustering. In the other three methods, the random method is used to generate the population, while in FMOPG, it starts to search from a better position using smart grid-based initialization.

Table 5 indicates the effect of PCA and keyframes on HAD performance (the effect of selecting the keyframe based on kinetic energy is reported in Fig. 3 in supplementary). We can observe that combining both methods not only reduces the clustering time but also has a positive effect on the HAD performance due to the selection of frames that show the most differentiation and reducing the number of features to those most useful to HAD by keyframes and PCA, respectively.

Fig. 9 indicates the effect of mean-shift vector and the initialization on FMOPG based on best precision (Fig. 9(a)) and best convergence rate (Fig. 9(b)) in subject 1 of F3D dataset. From the results, the performance of MOPSO (red bar) without the use of smart initialization and mean-shift vector has obtained the lowest precision (42.04%) compared to other combinations (54.32% for MOPSO + Smart initialization, 44.31% for MOPSO + Modified velocity, and 62.50% for MOPSO + Smart initialization + Modified velocity). But with the combination of both smart initialization and mean-shift vector to MOPSO (purple bar), the best precision has been achieved. Moreover, looking at Fig. 9(b), when the smart initialization is used (blue line), the search starts from a better position and lower value of SSE than when the particles are randomly initialized (red and green lines). Also, using the mean-shift vector (green line) to modify the velocity of particles has improved the convergence rate compared to MOSPO (red line), especially in the final iterations. The best convergence rate has been obtained by integrating both methods into MOPSO (purple line). Therefore, adding the smart initialization enables the proposed method to start from a stable point, and by updating the particle velocities based on the mean-shift vector, early convergence is prevented.

Fig. 10 shows the confusion matrices of HPGMK (a), MOPGMGT (b), and FMOPG (c) on subject 8 in the MSR dataset. The larger darker the color of the squares on the diagonal of the matrix, the better the clustering performance. The proposed method has achieved the best performance in reducing the confusion between activities compared to the other two methods. The other two, use Euclidean distance for clustering activities and will have difficulty to find non-linear clusters.

5. Conclusion

In this paper, we have proposed a novel flexible multi-objective clustering based on game theory to address human activity discovery. A new smart grid-based initialization method was introduced to improve the initialization of the centroids. We modified particles' velocity update by using a mean-shift vector to cope with premature convergence and non-linear clusters. Gaussian mutation was employed to generate diverse solutions. Besides, an incremental technique was presented to perform clustering dynamically and flexibly without prior information

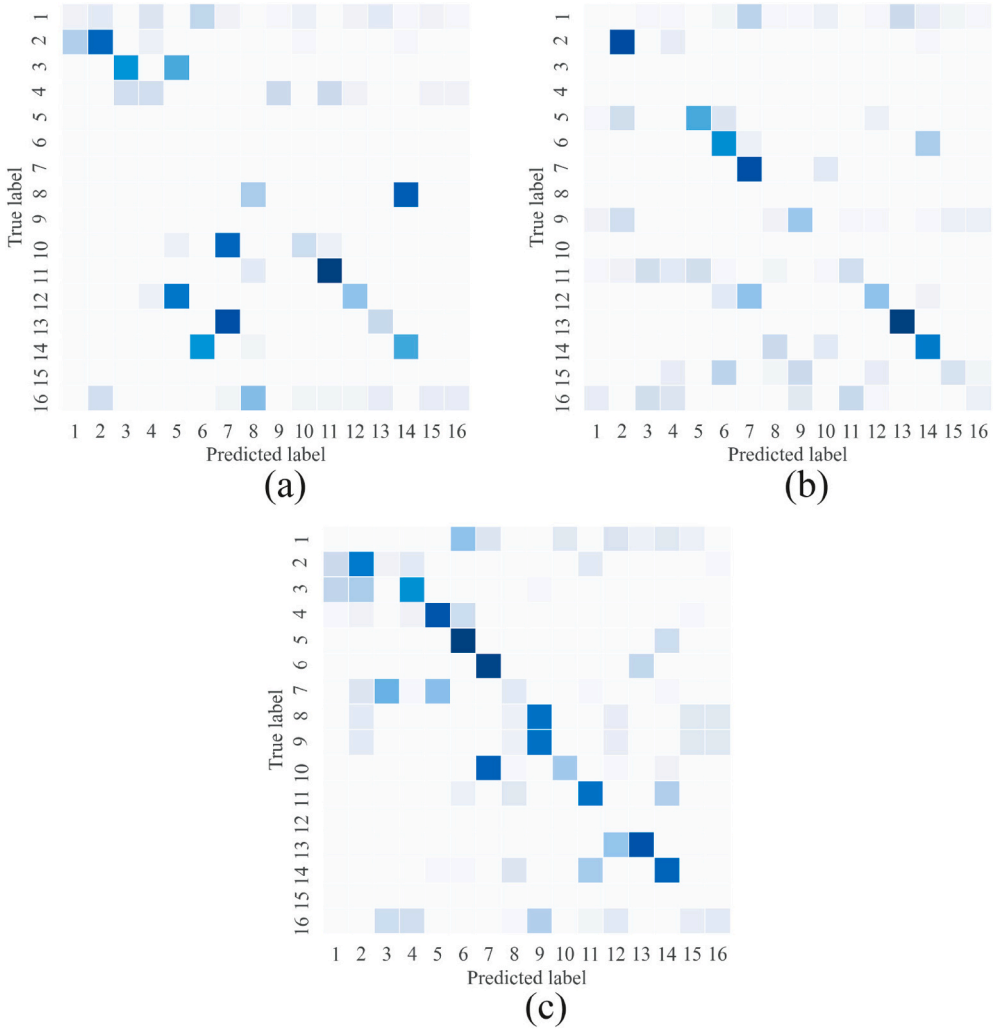

Fig. 10. Comparison of confusion matrices between the three methods HPGMK (a), MOPGMGT (b) and FMOPG (c) on subject 8 in MSR dataset. Activity list: (1) Drink; (2) Eat; (3) Read book; (4) Call cellphone; (5) Write on a paper; (6) Use laptop; (7) Use vacuum cleaner; (8) Cheer up; (9) Sit still; (10) Toss paper; (11) Play game; (12) Lie down on sofa; (13) Walk; (14) Play guitar; (15) Stand up; and (16) Sit down. AVG is the average F-score for all activities.

about the number of clusters. The effectiveness of the proposed method was shown on seven challenging datasets. The proposed method outperformed the state-of-the-art methods in all evaluation parameters and has improved the accuracy on the CAD-60, UTK, MSR, and NTU-60 by 8.66%, 1.75%, 3.56%, and 1.27%, respectively.

Although FMOPG has shown great performance in HAD, hand-crafted features are very sensitive to parameter tuning and are not reliable for a real scenario. As future work, the hand-crafted method used can be replaced by a feature learning approach. Sliding window segmentation method can be enhanced using a learning approach to detect the length of the sliding window dynamically based on input streams of skeleton sequences. Lastly, the applications of FMOPGM to address real-world problems such as analyzing the driver behavior, finding the community in social networks, and face recognition are other possible future works.

CRediT authorship contribution statement

Parham Hadikhani: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daphne Teck Ching Lai:** Writing – review & editing, Supervision. **Wee-Hong Ong:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Parham Hadikhani reports financial support was provided by Universiti Brunei Darussalam School of Digital Science. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imavis.2024.104985.

References

[1] P. Hadikhani, M. Eslaminejad, M. Yari, E. Ashoor Mahani, An energy-aware and load balanced distributed geographic routing algorithm for wireless sensor networks with dynamic hole, Wirel. Netw 26 (1) (2020) 507–519.

[2] S.K. Yadav, K. Tiwari, H.M. Pandey, S.A. Akbar, A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions, Knowl.-Based Syst. 223 (2021) 106970.

[3] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vis. 79 (3) (2008) 299–318.

[4] B. Peng, J. Lei, H. Fu, L. Shao, Q. Huang, A recursive constrained framework for unsupervised video action clustering, IEEE Trans. Industr. Inform. 16 (1) (2019) 555–565.

[5] T. Wang, W.W. Ng, J. Li, Q. Wu, S. Zhang, C. Nugent, C. Shewell, A deep clustering via automatic feature embedded learning for human activity recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (1) (2021) 210–223.

[6] F. Leotta, M. Mecella, D. Sora, Visual process maps: a visualization tool for discovering habits in smart homes, J. Ambient. Intell. Humaniz. Comput. 11 (5) (2020) 1997–2025.

[7] W. Huang, L. Zhang, H. Wu, F. Min, A. Song, Channel-equalization-har: a light-weight convolutional neural network for wearable sensor based human activity recognition, IEEE Trans. Mob. Comput. 22 (9) (2022) 5064–5077.

[8] D. Cheng, L. Zhang, C. Bu, X. Wang, H. Wu, A. Song, Protohar: prototype guided personalized federated learning for human activity recognition, IEEE J. Biomed. Health Inform. (2023) 3900–3911.

[9] S. Xu, L. Zhang, Y. Tang, C. Han, H. Wu, A. Song, Channel attention for sensor-based activity recognition: embedding features into all frequencies in dct domain, IEEE Trans. Knowl. Data Eng. 35 (12) (2023) 12497–12512.

[10] 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.

[11] M. Liu, Q. He, H. Liu, Fusing shape and motion matrices for view invariant action recognition using 3d skeletons, in: In 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3670–3674.

[12] X. Shu, L. Zhang, G.-J. Qi, W. Liu, J. Tang, Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction, IEEE Trans. Pattern Anal. Mach. 44 (6) (2021) 3300–3315.

[13] T. Hossain, S. Sarker, S. Rahman, M.A.R. Ahad, Skeleton-based human action recognition on large-scale datasets, in: Vision, Sensing and Analytics: Integrative Approaches, Springer, 2021, pp. 125–146.

[14] S. Yang, J. Liu, S. Lu, M.H. Er, A.C. Kot, Skeleton cloud colorization for unsupervised 3d action representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13 423–13 433.

[15] S. Liao, T. Lyons, W. Yang, K. Schlegel, H. Ni, Logsig-rnn: a novel network for robust and efficient skeleton-based action recognition, *arXiv preprint arXiv:2110.13008*, 2021.

[16] 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.

[17] P. Hadikhani, D.T.C. Lai, W.-H. Ong, A novel skeleton-based human activity discovery technique using particle swarm optimization with gaussian mutation, arXiv preprint arXiv:2201.05314, 2022.

[18] M. Zhang, T. Zhu, M. Nie, Z. Liu, More reliable neighborhood contrastive learning for novel class discovery in sensor-based human activity recognition, Sensors 23 (23) (2023) 9529.

[19] 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.

[20] M. Eldib, W. Philips, H. Aghajan, Discovering human activities from binary data in smart homes, Sensors 20 (9) (2020) 2513.

[21] 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.

[22] P. Hadikhani, D.T.C. Lai, W.-H. Ong, M.H. Nadimi-Shahraki, Automatic deep sparse multi-trial vector-based differential evolution clustering with manifold learning and incremental technique, Image Vis, Comput. 136 (2023) 104712.

[23] S. Sempena, N.U. Maulidevi, P.R. Aryan, Human action recognition using dynamic time warping, in: Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, IEEE, 2011, pp. 1–5.

[24] D. Wu, L. Shao, Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 724–731.

[25] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recogn. 68 (2017) 346–362.

[26] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, S. Xia, Graph cnns with motif and variable temporal block for skeleton-based action recognition, Proc. AAAI Conf. Artificial Intel. 33 (01) (2019) 8989–8996.

[27] G. Liu, J. Qian, F. Wen, X. Zhu, R. Ying, P. Liu, Action recognition based on 3d skeleton and rgb frame fusion, in: In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2019, pp. 258–264.

[28] I. Lee, D. Kim, S. Lee, 3-d human behavior understanding using generalized ts-lstm networks, IEEE Trans. Multimed. 23 (2020) 415–428.

[29] X. Shen, Y. Ding, Human skeleton representation for 3d action recognition based on complex network coding and lstm, J. Vis. Commun. Image Represent. 82 (2021) 103386.

[30] P. Hadikhani, D.T.C. Lai, W.-H. Ong, M.H. Nadimi-Shahraki, Improved data clustering using multi-trial vector-based differential evolution with gaussian crossover, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, 2022, pp. 487–490.

[31] N. Matake, T. Hiroyasu, M. Miki, T. Senda, Multiobjective clustering with automatic k-determination for large-scale data, in: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, 2007, pp. 861–868.

[32] A. Abubaker, A. Baharum, M. Alrefaei, Automatic clustering using multi-objective particle swarm and simulated annealing, PLoS One 10 (7) (2015) e0130995.

[33] H. Qu, L. Yin, X. Tang, Multi-objective automatic clustering with gene rearrangement and cluster merging, in: Advances in Intelligent Systems Research and Innovation, Springer, 2022, pp. 87–127.

[34] E.W. Forgy, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, biometrics 21 (1965) 768–769.

[35] S. Vassilvitskii, D. Arthur, k-means++: The advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2006, pp. 1027–1035.

[36] D. Bajer, G. Martinović, J. Brest, A population initialization method for evolutionary algorithms based on clustering and cauchy deviates, Expert Syst. Appl. 60 (2016) 294–310.

[37] M. Perez, J. Liu, A.C. Kot, Skeleton-based relational reasoning for group activity analysis, Pattern Recogn. 122 (2022) 108360.

[38] B. Nikpour, N. Armanfard, Joint Selection Using Deep Reinforcement Learning for Skeleton-Based Activity Recognition, 2021.

[39] T. Li, Q. Ke, H. Rahmani, R.E. Ho, H. Ding, J. Liu, Else-net: Elastic semantic network for continual action recognition from skeleton data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13 434–13 443.

[40] G. Ercolano, S. Rossi, Combining cnn and lstm for activity of daily living recognition with a 3d matrix skeleton representation, Intell. Serv. Robot. 14 (2) (2021) 175–185.

[41] S. Miao, Y. Hou, Z. Gao, M. Xu, W. Li, A central difference graph convolutional operator for skeleton-based action recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (7) (2021) 4893–4899.

[42] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra-based action recognition using convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. 28 (3) (2016) 807–811.

[43] L. Zhou, W. Li, P. Ogunbona, Z. Zhang, Jointly learning visual poses and pose lexicon for semantic action recognition, IEEE Trans. Circuits Syst. Video Technol. 30 (2) (2019) 457–467.

[44] B. Li, X. Li, Z. Zhang, F. Wu, Spatio-temporal graph routing for skeleton-based action recognition, Proc. AAAI Conf. Artificial Intel. 33 (01) (2019) 8561–8568.

[45] X. Gao, W. Hu, J. Tang, J. Liu, Z. Guo, Optimized skeleton-based action recognition via sparsified graph regression, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 601–610.

[46] P. Koniusz, L. Wang, A. Cherian, Tensor representations for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2) (2021) 648–665.

[47] H. Mohammadzade, M. Tabejamaat, Sparseness embedding in bending of space and time; a case study on unsupervised 3d action recognition, J. Vis. Commun. Image Represent. 66 (2020) 102691.

[48] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, Z. Gong, Unsupervised representation learning with long-term dynamics for skeleton based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence 32, 2018 no. 1.

[49] Y. Tang, Y. Wei, X. Yu, J. Lu, J. Zhou, Graph interaction networks for relation transfer in human activity videos, IEEE Trans. Circuits Syst. Video Technol. 30 (9) (2020) 2872–2886.

[50] M. Guo, E. Chou, D.-A. Huang, S. Song, S. Yeung, L. Fei-Fei, Neural graph matching networks for fewshot 3d action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 653–669.

[51] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, Ntu rgb+ d 120: a large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2019) 2684–2701.

[52] L. Lin, S. Song, W. Yang, J. Liu, Ms2l: Multi-task self-supervised learning for skeleton based action recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2490–2498.

[53] P. Wang, J. Wen, C. Si, Y. Qian, L. Wang, Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition, IEEE Trans. Image Process. 31 (2022) 6224–6238.

[54] Y. Zhu, H. Han, Z. Yu, G. Liu, Modeling the relative visual tempo for self-supervised skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13 913–13 922.

[55] W.-H. Ong, L. Palafox, T. Koseki, Autonomous learning and recognition of human action based on an incremental approach of clustering, IEEJ Trans. Elect. Inform. Syst. 135 (9) (2015) 1136–1141.

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

[57] P. Hadikhani, D.T.C. Lai, W.-H. Ong, Human Activity Discovery with Automatic Multi-Objective Particle Swarm Optimization Clustering with Gaussian Mutation and Game Theory, 2022.

[58] K. Huang, X. Fu, N. Sidiropoulos, On convergence of epanechnikov mean shift, in: Proceedings of the AAAI Conference on Artificial Intelligence 32, 2018 no. 1.

[59] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from rgbd images, in: IEEE International Conference on Robotics and Automation 2012, IEEE, 2012, pp. 842–849.

[60] S. Gaglio, G.L. Re, M. Morana, Human activity recognition process using 3-d posture data, IEEE Trans. Human-Machine Syst. 45 (5) (2014) 586–597.
[61] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: In 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1290–1297.
[62] 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.
[63] 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.
[64] 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.
[65] D. Van der Merwe, A.P. Engelbrecht, Data clustering using particle swarm optimization, in: The 2003 Congress on Evolutionary Computation, IEEE, 2003, pp. 215–220, 2003. CEC'03., vol. 1.
[66] R. McConville, R. Santos-Rodriguez, R.J. Piechocki, I. Craddock, N2d:(not too) deep clustering via clustering the local manifold of an autoencoded embedding, in: In 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 5145–5152.
[67] B. Yang, X. Fu, N.D. Sidiropoulos, M. Hong, Towards k-means-friendly spaces: Simultaneous deep learning and clustering, in: International Conference on Machine Learning, PMLR, 2017, pp. 3861–3870.
[68] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, L. Van Gool, Scan: Learning to classify images without labels, in: European Conference on Computer Vision, Springer, 2020, pp. 268–285.
[69] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, X. Peng, Completer: Incomplete multi-view clustering via contrastive prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11 174–11 183.
[70] Y. Zhu, W. Chen, G. Guo, Evaluating spatiotemporal interest point features for depth-based action recognition, Image Vis. Comput. 32 (8) (2014) 453–464.
[71] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, A human activity recognition system using skeleton data from rgbd sensors, Comput. Intell. Neurosci. 2016 (2016).
[72] B. Seddik, S. Gazzah, N. Essoukri Ben Amara, Human-action recognition using a multi-layered fusion scheme of kinect modalities, IET Comp. Vision 11 (7) (2017) 530–540.
[73] S. Neili Boualia, N. Essoukri Ben Amara, Deep full-body hpe for activity recognition from rgb frames only, in: Informatics vol. 8, MDPI, 2021, p. 2, no. 1.
[74] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[75] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. Theory Methods 3 (1) (1974) 1–27.
[76] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1979) 224–227.
[77] J.C. Dunn, A Fuzzy Relative of the Isodata Process and its Use in Detecting Compact Well-Separated Clusters, 1973.
[78] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. Royal Stat. Soc. Series B (Stat. Methodol.) 63 (2) (2001) 411–423.
[79] R.L. Thorndike, Who belongs in the family, in: *Psychometrika*. Citeseer, 1953.
[80] J.A. Hartigan, Clustering Algorithms, John Wiley & Sons, Inc, 1975.
[81] W.J. Krzanowski, Y. Lai, A criterion for determining the number of groups in a data set using sum-of-squares clustering, Biometrics (1988) 23–34.
[82] A. Fujita, D.Y. Takahashi, A.G. Patriota, A non-parametric method to estimate the number of clusters, Comp. Stat. Data Anal. 73 (2014) 27–39.
[83] C.A. Sugar, G.M. James, Finding the number of clusters in a dataset: an information-theoretic approach, J. Am. Stat. Assoc. 98 (463) (2003) 750–763.
[84] P. Agarwal, S. Mehta, A. Abraham, A meta-heuristic density-based subspace clustering algorithm for high-dimensional data, Soft. Comput. (2021) 1–20.
[85] 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.