International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Oct-23)

http://dx.doi.org/10.12785/ijcds/1401110

A Comparative Study of Supervised and Unsupervised Approaches in Human Activity Analysis Based on Skeleton Data

Md Amran Hossen[1] **and Pg Emeroylariffion Abas**[1]

[1]*Faculty of Integrated Technologies, Universiti Brunei Darussalam, Gadong, Brunei Darussalam*

Received 07 Jul. 2023, Revised 14 Oct. 2023, Accepted 18 Oct. 2023, Published 20 Oct. 2023

Abstract: One of the important areas of machine intelligence research today is human activity recognition (HAR), with the goal of automatically identifying human activities from various types of sensor data. Most of the existing human activity recognition methods use hand-crafted features and labelled data, but these methods fail to identify new activities not defined in the training dataset. As human activities are numerous and executed in various ways, it is challenging to obtain enough labelled data to train a model to recognize the different activities. In this paper, the performance of five different supervised learning algorithms on the human activity recognition task with skeleton-based features has been evaluated, using five publicly available datasets and an experimental dataset. Accuracies of above 90% are achievable on datasets with a limited number of samples using commonly available classification algorithms and simple skeleton-based features. Subsequently, the same feature sets are used on unsupervised learning methods for an unsupervised clustering task. Using the unsupervised learning algorithms, an average of 74% f1-score on the publicly available CAD60 dataset and 61% f1-score on the experimental dataset, are obtained. These results demonstrate the effectiveness of simple skeleton-based features, coupled with common supervised and unsupervised learning algorithms in human activity recognition tasks.

Keywords: Human Activity Recognition; Activity Discovery; Activity Classification, Machine Learning, Activity Analysis

1. INTRODUCTION

Extensive work has been carried out in the field of Human Activity Recognition (HAR) due to its broad applications, including in Active and Assisted Living (AAL), surveillance and monitoring, human-computer interaction, healthcare and more. Commonly, HAR researchers use various sensing technologies to collect activity data and identify suitable methods to accurately predict activities. Different types of sensor data are frequently used to analyse human activities, some of which have been shown to give a good performance in identifying a limited range of activities. For instance, a mobile phone kept in a person's pocket [1] can be used to efficiently identify activities, including standing, walking, sitting, lying down and standing up. Other researchers have used wearable devices [2] to monitor different types of activities. As seen in the previous survey, wearable devices are effective in collecting accurate data on the movement of the human body [3][4]; however, they are limited in identifying a handful of activities only, and the wearers may experience discomfort [4] over an extended period of time.

The emergence of RGB-D sensors on the market has stimulated the advancement of innovative methods to create cost-effective and improved solutions with vision technologies [5][6] [7]. A depth image is a relatively reliable source of data that is not influenced by changes in ambient light, and subsequently, can mitigate human identification and segmentation problems [8]. In fact, skeleton joints obtained from a depth map can be used as an accurate representation of a human body without actual sensors being attached to the human body. Additionally, privacy is a crucial aspect in AAL, that may be significantly impacted by HAR leading to consequences [9]. Depth images are more privacy-preserving than standard colour images due to their abstract representation of a humanoid figure from the depth stream and it is possible to only use the skeleton figure to depict a person [10]. This skeleton data can be used to recognise human activities.

The majority of existing works on human activity recognition focus on hand-crafted features from depth maps or RGB images, which are often aimed at improving the accuracy of an existing set of activities [11]. These approaches do not normally address the identification of activities, which have not been presented in the training dataset [12]. Consequently, there is increasing interest in the use of unsupervised approaches for human activity recognition or

E-mail address: 20h8453@ubd.edu.bn, emeroylariffion.abas@ubd.edu.bn

detection [13]. Most of the works, in this respect, have utilised colour images, depth images, body sensors and environment sensors, with a very limited number of works in semi-supervised [14] and unsupervised human activity detection based on skeleton data [12] [15] are available. Additionally, it has been observed that the majority of works in the research literature have performed extensive feature extractions and proposed complex learning models to achieve high recognition accuracy [11]. The ease of extracting skeleton data from a depth camera has provided efficient ways for investigating the effectiveness of simple skeleton-based features with both classical and deep learning algorithms for human activity recognition.

Classical machine learning and deep learning are distinct paradigms of machine intelligence, differing in their fundamental approaches to learning patterns from data. In classical machine learning, algorithms rely on mathematical models that are explicitly designed to map input features to output action labels [16]. Typically, this is represented as a function f that takes input features X and maps them to output labels Y:$Y = f(X)$. This mapping function is learned by optimizing model parameters to minimise a predefined loss function that quantifies the disparity between predicted and actual activity labels.

Conversely, deep learning employs neural networks, a hierarchy of interconnected layers of neurons or nodes. Each layer computes a weighted sum of its inputs and applies an activation function, often denoted as $h(z)$, to produce an output. The transformation from one layer to the next is represented mathematically as $h(WX+b)$, where W represents weights, X represents inputs, and b represents biases. Deep learning models consist of multiple layers (hence, the term ”deep”), enabling them to learn complex and hierarchical features from the data. The learning process involves optimising the weights and biases to minimise the loss function, typically represented as $L(Y,\hat{Y})$, where Y is the actual output and $\hat{Y}$ is the predicted output. This optimization is often achieved through backpropagation and gradient descent algorithms. In summary, classical machine learning focuses on explicit feature engineering and the optimization of predefined models, while deep learning leverages neural networks to automatically learn features and hierarchical representations from data through the iterative adjustment of network parameters. Both approaches aim to minimize a loss function but differ in their underlying architectures and learning mechanisms.

Both machine learning and deep learning methods are based on the assumption that an activity can be considered as a comprehensive series of skeleton postures, and these postures can be used to identify different activities. As seen in the literature [17][18], either a defined collection of main poses has been derived for each activity or extensive and sophisticated features have been extracted. These poses and features are often over-engineered to improve the accuracy of an existing set of activities, where the necessity of identifying new incoming activities is often overlooked. In this study, a more straightforward set of features for supervised and unsupervised activity detection has been investigated.

The organization of this paper is structured to provide readers with a clear understanding of our research approach and findings. In the subsequent sections, a thorough exploration of the relevant prior work in the field of human activity recognition, discussing key methodologies and their implications, is given. The Materials and Methods section details the datasets employed for the study, describing the features extracted from data, the classification of activities, and the performance measures used for evaluation. Moving forward, the Results section presents a comprehensive analysis of the outcomes obtained from the application of both supervised and unsupervised learning algorithms, offering insights into the strengths and limitations of each approach. This organization ensures a logical flow of information, guiding readers through the background, methodology, and outcomes of the research in a systematic manner.

2. RELATED WORK

Several studies have attempted to recognize human activities by utilizing different sensors. As seen in the previous surveys [17][18], modalities often include wearable devices, mobile phones, RGB cameras, stereo images constructed from 2D sensors, motion capture systems (MoCAP), range sensors, and 3D input sensors, such as Microsoft Kinect. Due to the affordability and potential performance benefits, depth cameras are increasingly utilized in human activity analysis research. Authors in references [19] [20] have reviewed several skeleton joint-based methods for human activity recognition. Brief literature reviews on human activity recognition work that have used skeleton data derived from skeleton images are described in the following section.

Many techniques have been developed for human activity recognition, and a number of them are geared towards extracting skeleton features from depth data, where the key concept is to evaluate subvolume descriptors of spatiotemporal depth. Yang et al. [21] proposed using a collection of hypersurface normal (polynormal), including details of geometry and local motion that have been derived from depth sequences. The polynormal elements are integrated to construct the final depth map representation, named Super Normal Vector (SNV). In reference [22], depth images have been interpreted as a series of features and these have been briefly modelled as subspaces that lie on the Grassmann manifold. Beginning with the direction of the usual vector at each surface level, this description represents the geometric structure and dynamics of the human body without using the joint location. Devanne et al. [23] suggested portraying human activities through spatio-temporal trajectories of motion in a sixty-dimensional space. Each activity involves 20 joints, with each joint represented by 3 coordinates. An elastic metric (a metric which is insensitive to speed and time of motion) has been used to describe the difference

among multiple activities in a Riemannian form of space. Subsequently, action recognition has been performed in the Riemannian domain using a K-Nearest-Neighbour (K-NN) classifier. The portrayal of APJ3D [24] is represented with a subset of 15 skeletal joints, from which the relative locations and local radial angles are determined. The activity is partitioned using an evaluated Fourier Temporal Pyramid [25] after a set of key positions and the classification is rendered by a random forest classifier (RF).

A joint representation named HOJ3D has been proposed by Xia et al. [26], in which the 3D space is segmented into n bins and a Gaussian weight feature is applied to associate the skeletal joints to each bin. The data is then reprojected into lower dimension by using Linear Discriminant Analysis (LDA), before using a k-means clustering algorithm to select k-posture visual words to represent each activity. A distinct Hidden Markov Model (HMM) has also been used to model the temporal evolution of postures to classy different activities [27]. To portray the skeleton, Taha et al. [27] took advantage of joint spherical coordinates, and a framework consisting of a multiclass support vector machine (SVM) and a discrete HMM, was used to classify human activities. Cippitelli et al. [6] used a k-means clustering algorithm to select the most insightful postures for each activity sequence, choosing a separate set of k postures that construct the feature vectors for each activity. Finally, action recognition has also been performed using a multiclass SVM. The study in reference [28] presented a comparative study of human activity recognition using 2D and 3D human postures extracted from depth images, and the findings show that the Random Forest classification model yields the highest accuracy among eight classification models; demonstrating the effectiveness of both 2D and 3D postures in achieving accurate activity classification. Most works have utilised shallow classification models such as SVM for skeleton-based activity recognition tasks. On the other hand, other shallow classification models including decision trees, random forest classifier, K-nearest neighbourhood and multi-layer perceptron classifier with only two hidden layers, were additionally used in this paper, and compared with some of the research that have used handcrafted features.

HON4D, introduced by authors in reference [29], is a global feature representation that encapsulates the geometry and motion of human actions in a 4D space consisting of spatial coordinates, depth, and time. Similarly, the HDG method proposed by Rahmani et al. [30] uses depth sequences which were first separated into smaller sections; with depth derivatives and histograms of depth calculated for each section. Another feature-based method called HOPC [31] has also been proposed. The approach model's depth images as 3D point clouds and introduces two types of support volumes: spatial support volume and spatio-temporal support volume. The HOPC descriptor is computed by extracting features from the point cloud data within the support volume of each point. Key points, known as spatio-temporal Key points (STK), are identified based on eigenvalue ratios exceeding a threshold. Likewise, the LARP-SO algorithm known as the Lie Algebra Relative Pairs via SO(3) was employed by Vemulapalli and Chellappa [32] for action recognition. The 3D action recognition algorithm in the study utilized the concept of a rolling map, which illustrates the movement of one Riemannian manifold over another along a smooth rolling curve. To represent each skeleton sequence, the algorithm analyses the relative 3D rotations between different body parts and models each action as a curve in the Lie Group. Additionally, the algorithm incorporates Fourier Temporal Pyramid (FTP) representation [33] to enhance the descriptor's resilience to noise and reduce sensitivity to temporal misalignments.

Dictionary learning and unsupervised feature learning have found applications in the domain of human activity recognition, offering valuable tools for extracting discriminative features from skeleton data. Dictionary learning methods, such as Sparse Coding [34][35], enable the representation of activity data as a linear combination of basic elements in a learned dictionary. This approach has proven effective in capturing spatiotemporal patterns in activity sequences, enhancing recognition accuracy. Likewise, unsupervised feature learning techniques, strive to discover meaningful representations directly from raw sensor data, reducing the need for handcrafted features. However, both dictionary learning and unsupervised feature learning methods are not without limitations. One significant challenge is their reliance on extensive labelled data for training, which can be impractical or costly to obtain in real-world scenarios. Additionally, these techniques may struggle to handle high-dimensional sensor data, leading to increased computational complexity.

In recent years various researchers have proposed deep learning-based methods for activity sequence learning, which is inspired by the effectiveness of deep learning in applications, including video captioning [36], audio recognition [37], neural machine translation [38], image recognition [39], and speech recognition [40][41][42]. Other deep learning approaches based on skeletal data include Recurrent Neural Network (RNN) [5][43], Convolutional Neural Network (CNN) [44], and Graph Convolutional Network (GCN) [45]. Notably, Graph Convolution Networks (GCNs) have gained prominence in this field. GCNs enable the modelling of skeletal joint connections as a graph, allowing for the propagation of information across the skeleton structure.

Researchers have harnessed the power of GCNs to capture both spatial and temporal dependencies among skeletal joints, achieving state-of-the-art results [46]. Concurrently, Convolutional Neural Networks (CNNs) have been adapted for skeleton data by treating it as a sequence of 2D or 3D heatmaps. CNNs excel at automatically learning hierarchical features from these heatmaps, enabling accurate action recognition [47]. Authors in reference [47] intro-

duced PoseConv3D, an alternative approach that relies on a 3D heatmap volume rather than a graph sequence to represent human skeletons. PoseConv3D attains state-of-the-art performance on multiple skeleton-based action recognition benchmarks and outperforms existing methods when fused with other modalities in multi-modality action recognition benchmarks. Additionally, Transformers, known for their self-attention mechanisms, have demonstrated promise in capturing long-range dependencies in skeleton sequences [48]. The attention mechanism allows Transformers to focus on relevant joints and their interactions, leading to improved recognition accuracy. These recent advancements in deep learning techniques have significantly enriched the landscape of skeleton-based human activity recognition, offering promising avenues for future research. Although deep learning-based methods have achieved unprecedented performance improvement in action/activity recognition as shown in a comprehensive review by the authors in [49], deep learning-based methods commonly require abundant labelled data.

Naturally, understanding unseen activity is more complicated than modelling existing activities. Many researchers have proposed clustering-based methods for discovering unseen activities from accelerometer-based data [50], and have utilized various metrics for performance evaluation. Accelerometer-based data generates distinctive patterns for a set of daily activities, such as climbing stairs, walking, running, jogging, lying down, and standing up, making it a popular choice for unsupervised human activity discovery, however, this approach fails to identify high-level activities such as talking on the phone and drinking. Ironically, very little work has been done on unsupervised human activity detection from skeleton data. Hoda et al. [51] proposed an unsupervised 3D action recognition method based on the sparseness embedding of time and space for unsupervised action learning. However, their approach focuses more on learning various activities from labelled data instead of grouping activities without labelled data. Ong et al. [52] used k-means clustering for unsupervised activity detection. The human ranges of motion have been used as a feature for activity detection. However, the work only used k-means clustering, where it has been shown that k-means clustering performs better when clusters are spherical. However, in actual reality, cluster shapes and distributions may vary. Reference [15] assessed the performance of various clustering algorithms, including k-means, spectral, hierarchical, and BIRCH clustering, in distinguishing different daily activities for human activity discovery from unlabelled observations, however, the investigation has not proposed any new method.

In this study, 3D joint information, obtainable from an RGB-D camera, was harnessed to extract straightforward yet informative features. While both handcrafted features and shallow classifiers have found utility in human action recognition, a comprehensive comparison of action recognition algorithms, specifically assessing the efficacy of shallow classifiers combined with simpler features, has been notably absent in the existing literature. Our study bridges this research gap by conducting a meticulous evaluation of five state-of-the-art methods for human action recognition. Particular emphasis has been placed on comparing the efficiency of handcrafted features versus skeleton-based features, recognizing the scarcity of such comprehensive comparisons in the domain of human action recognition. Furthermore, we recognize the crucial role that dataset size plays in the performance of machine learning and deep learning models. As such, our investigation extends to assessing these algorithms' performance across datasets of varying sizes. Through this endeavour, we aim to shed light on the suitability of different algorithms and features for human action recognition, considering the practical implications of dataset size and the existing literature's focus on either handcrafted features or deep learning approaches.

3. MATERIALS AND METHODS

A. *Scheme for Human Activity Recognition*

Figure 1 depicts an overview of the activity recognition scheme adopted in this paper. A sequence of depth images is illustrated in Figure 1(a), and skeleton joint coordinates were extracted for each depth image using the method proposed by the authors in reference [53], which is available in the Microsoft Kinect software development kit (SDK) package and also used by the authors in references [28][15]. Subsequently, joint orientation and pairwise Euclidean distance (PED) between the hip center and other joints were derived from the skeleton joints, as shown in Figure 1 (b). Features from multiple frames were concatenated sequentially to represent a particular activity (Figure 1(c)); these activity instances were used as input to train activity classification models. Prior to classification, each dataset was partitioned into training and testing sets using an 80:20 ratio. Furthermore, a 10-fold cross-validation technique was applied to the cross-subject scenario with a random shuffle of the training data. Final evaluation was performed on the test dataset. Algorithms that contained a random factor were evaluated 10 times, and the mean result was reported. Multiple classification algorithms such as k-nearest neighbors, decision tree, support vector machine, random forest, and multi-layer perceptron classifier with two hidden layers are evaluated in this paper.

Skeleton joint coordinates derived from the depth images are real-world joint positions. The coordinates are first translated from the real-world coordinates to camera coordinates centring the hip centre joint and then a rotation based on the left and right shoulder joint is applied to achieve view invariance. The first frame of an activity sequence is rotated facing the viewer while the rotation in the subsequent frames is relative to the first frame. The rotation method is almost similar to the one proposed in reference [5]. However, while reference [5] applied a view adaptive recurrent neural network, the proposed method automatically derived the rotation angle for each frame by calculating the angle formed by translating the line

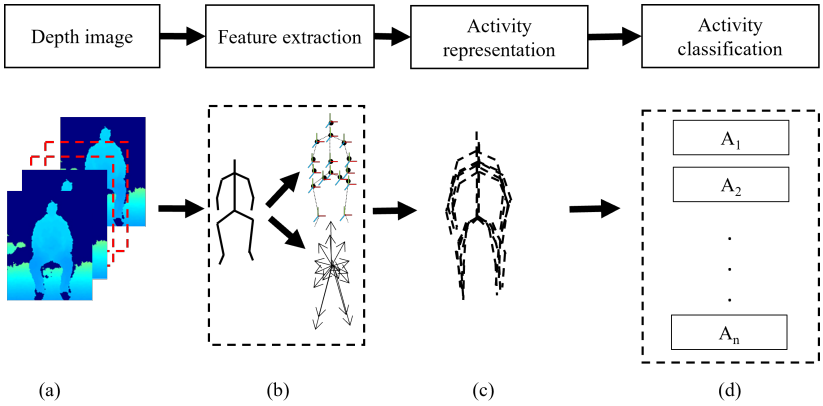


Figure 1. Overview of the activity recognition scheme. (a) multiple depth frames in sequence (b) skeleton joints positions are extracted features extracted from each frame (c) features from multiple frames are concatenated to construct an activity vector (d) a model is trained to classify activities.

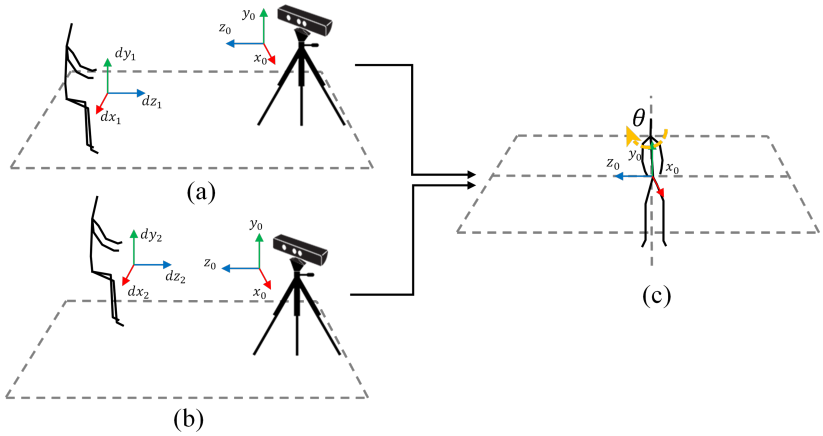


Figure 2. Skeleton joint transformations (a) and (b) represent real-world skeleton joint coordinates in two different positions. (c) Skeleton joints are translated to the camera coordinate system, centered on the hip-center joint, and rotated by an angle θ degrees.

TABLE I. Publicly Accessible Datasets And An Experimental Dataset Used For 3d Action Recognition Experiments.

Datasets	Year	Actions	Subjects	Views	Videos	Device	Sense modality	Joints	Frames
MSR Action3D [33]	2010	20	10	1	567	Kinect v1	Depth + 3DJoints	20	59
CAD60 [54]	2011	14	4	1	68	Kinect v1	RGB + Depth + 3D Joints	15	30
F3D [55]	2013	9	10	1	215	Kinect v1	RGB + Depth + 3D Joints	15	30
3D Action Pairs [29]	2013	12	10	1	360	Kinect v1	RGB + Depth + 3D Joints	20	30
UWA3D Multiview Activity II [31]	2015	30	9	4	1070	Kinect v1	RGB + Depth + 3D Joints	15	70
NTU RGB+D 120 [56]	2017	120	106	155	114480	Kinect v2	RGB + Depth + IR + 3D Joints	25	300
Experimental dataset [15]	2019	18	3	1	2295	Kinect v1	RGB + Depth + 3D Joints	20	70

formed by connecting the shoulder left, shoulder centre and shoulder right to the x-axis centring the shoulder centre joint. Skeleton joint transformation is shown in Figure 2. In the first instance, labelled activity instances are used for training the different classification models, which are then used for the supervised and unsupervised activity classification tasks.

B. *Dataset*

There are a number of publicly available datasets [40] that contain skeleton joint data which may be used for human activity recognition works. The list of datasets used for this study is shown in Table I 1, which includes the MSRAction3D [33] dataset, Florence 3D dataset (F3D) [55], 3D action pairs dataset [29], CAD-60 [54], UWA3D Multiview Activity II [32] dataset, and NTU RGB+D 120 [56] dataset.

Additionally, an experimental dataset, used in reference [15] for human activity discovery has been used. The aim of using the dataset was to evaluate recognition performance with similar activities. The dataset contains seventeen activities performed by three subjects. All of the activities have

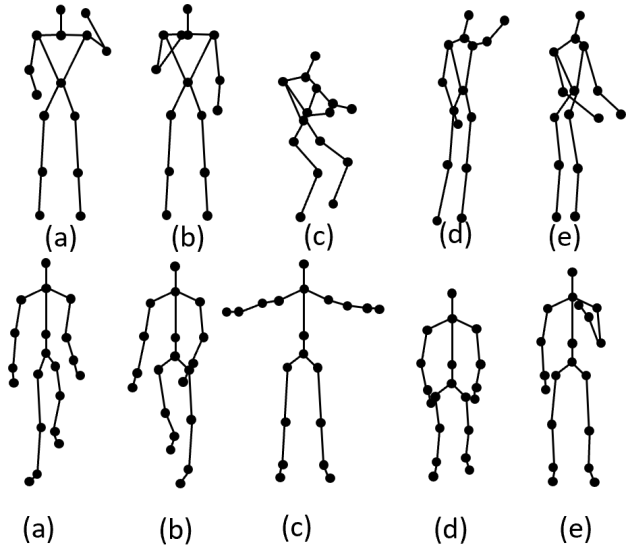


Figure 3. Top row: skeleton frames from the 15 joints skeleton dataset, a) talking on the phone, b) drinking, c) working on the computer, d) writing on the whiteboard, and e) cooking (chopping). Sample frames from the 20 joints skeleton dataset in row 2. (f) kick right leg (g) kick left leg (h) jumping jacks (i) seated (j) drinking.

been performed and recorded indoors employing a single stationary Microsoft Kinect sensor, and thus, are represented as RGB-D form. Each activity lasted between 3-4 minutes, with the data recorded at 30 frames per second (FPS). The dataset comprises the following activities: standing (STN), raising the right hand (RRH), raising the left hand (RLH), kicking with the right leg (KRL), kicking with the left leg (KLL), waving with the right hand (WRH), waving with left hand (WLH), performing jumping jacks (JJK), walking (WLK), sitting down (SDN), being seated (SIT), standing up (STU), making a phone call (TPN), drinking (DRK), pick object from the floor (PKU), sitting and reading book (RBS), and sweeping the floor (SWP). A few skeleton representations of the selected activities from the CAD60 and experimental datasets are shown on the top and bottom rows, respectively, in Figure 3.

C. Features

Selected features, including skeleton joint locations [26][6], joint orientation [57], and Euclidean distance between joints [58]. Figure 4 depicts the joint skeleton system at a particular frame instance comprising a set of m joints $i = \{1, 2, \ldots, m\}$. Each dataset contains a different value for m due to different skeleton tracking systems being used while recording the datasets. Each skeleton joint J_i may be described by its 3-dimensional (3D) positional information in the 3-dimensional space, $\mathbf{p}_i = (x_i, y_i, z_i) \in \mathbb{R}^3 : J_i \in \{J_1, \ldots, J_m\}$, as well as by its orientation information in Quaternion form, $\mathbf{o}_i = (q1_i, q2_i, q3_i, q4_i) \in \mathbb{R}^4 : J_i \in \{J_1, \ldots, J_m\}$.

The skeleton posture of a subject is formed by joining the 3D positional information of the skeleton joints at a

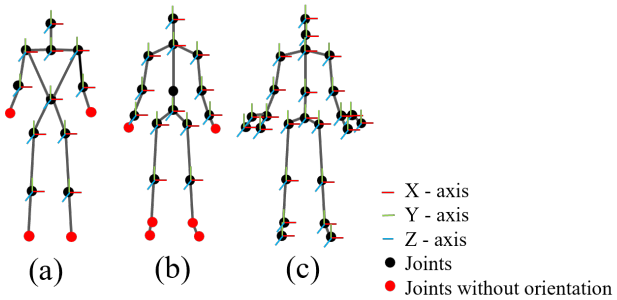


Figure 4. Skeleton joints used for this study (a) 15 skeleton joints (b) 20 skeleton joints (c) 25 skeleton joints.

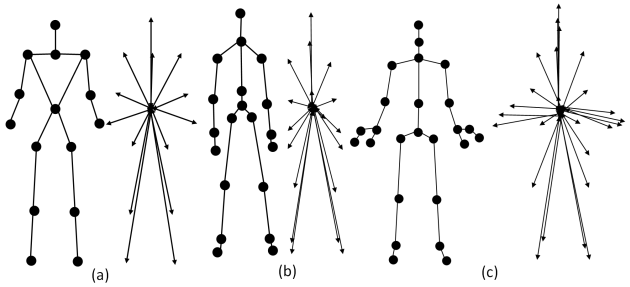


Figure 5. Pairwise Euclidean distance between the hip centre joint and the remaining joints in (a) fifteen skeleton (b) twenty skeleton and (c) twenty-five skeleton joints dataset.

specific frame. To ensure view invariance, this information is typically transformed to a selected centre of coordinates.

$$p_{ref\,i} = \left(x_{ref\,i},\ y_{ref\,i},\ z_{ref\,i}\right) = p_{hc} - p_i, \forall\ J_i \in J_1, \ldots,\ J_m \qquad (1)$$

Where $\mathbf{p}_{\text{ref}_i}$ is the translated 3D positional information of joint i with respect to the reference hip joint $\mathbf{p}_{\text{hc}}$. Euclidean distance between a reference skeleton joint to other joints may be extracted from $\mathbf{p}_{\text{ref}_i}$. Again, the hip-center joint $\mathbf{p}_{\text{hc}}$ is chosen as the reference joint, giving a total of m Euclidean distances, as depicted in Figure 5, representing the relative distance from the hip-center joint, denoted as d_i.

$$d_i = \sqrt{{x_{ref\,i}}^2 + {y_{ref_i}}^2 + {z_{ref_i}}^2}, \forall\ J_i \in J_1, \ldots,\ J_m \qquad (2)$$

From the orientation information $\mathbf{o}_i$ of joint i, the more familiar Euler form orientation representation $\mathbf{o}_{\mathbf{ref}_i} = \{\phi_i, \theta_i, \psi_i\}$ can be obtained, where ϕ_i, θ_i, and ψ_i are roll, pitch, and yaw angles of joint i, respectively.

$$\mathbf{o}_{\text{ref}_i} = \begin{bmatrix} \phi_i \\ \theta_i \\ \psi_i \end{bmatrix} = \begin{bmatrix} \tan^{-1}\left(\frac{2(q1_i.q2_i+q3_i.q4_i)}{1-2(q1_i^2+q2_i^2)}\right) \\ \sin^{-1}\left(2(q1_i.q3_i - q4_i.q2_i)\right) \\ \tan^{-1}\left(\frac{2(q1_i.q4_i+q2_i.q3_i)}{1-2(q3_i^2+q4_i^2)}\right) \end{bmatrix}, \forall\ J_i \in \{J_1, \ldots, J_m\} \quad (3)$$

$\mathbf{p}_{\text{ref}\,i}$, d_i, and $\mathbf{o}_{\text{ref}\,i}$ of the joint J_i may be used as features to represent each joint. For a specific frame instance, denoted as I_j, the instances in that particular frame j may be represented as a set of features $f_i \in \{\mathbf{p}_{\text{ref}\,i}, d_i, \mathbf{o}_{\text{ref}\,i}\}$ of the different skeleton joints $\forall J_i \in \{J_1, \ldots, J_m\}$.

$$I_j = f_i, \forall\ J_i \in J_1, \ldots,\ J_m \quad (4)$$

The kth activity, A_k can then be represented as a sequence of n of these frame instances, I_j, which are composed of a collection of selected features from the joint skeletons:

$$A_k = I_j, \forall\ J_i \in J_1, \ldots,\ J_m \quad (5)$$

D. Classification of Activities

The representation of the activity A_k may then be used as input into classification models, for activity recognition and activity pattern discovery. Activity recognition is essentially a supervised process, which identifies an activity A_k as belonging to either one of a set of K-recognised activities. This requires the presence of labelled training data to train the classification models before they can be used for activity recognition tasks. On the other hand, activity pattern discovery only attempts to group activities with almost similar characteristics. However, it can be used in an unsupervised manner and does not require pre-training.

1) Human activity recognition

For a dataset $D \in \{A_1, A_2, \ldots, A_k\}$ where $A_k = \{I_j\}$ contains k number of activity representations, with corresponding class label $C \in \{c_1, c_2, \ldots, c_k\}$ where $c_k \in \{1, 2, \ldots, K\}$. Each activity A_k has a corresponding class label c_k, belonging to one of the K activity classes. The aim is to train a classification model using the training samples, such that for a new activity representation A_*, it can be classified as one of the K activity classes, i.e., to have a corresponding class label $c_* \in \{1, 2, \ldots, K\}$. Five supervised classification models from the Scikit-learn machine learning library [59] were used for evaluation, including multiclass Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) with RBF kernel, and Multi-Layer Perceptron (MLP) with two hidden layers classification models. Hyperparameter tuning was avoided, with only the default parameter values used for each algorithm. The classification process has been divided into the training and testing phases, with each dataset divided into 80% training and 20% testing. A 10-fold cross-subject validation was performed to validate the training dataset. During the training phase, activity representations A_k with their corresponding class label c_k have been used to train the classifiers. The trained classification models are then used to classify unknown activity representations A_* into one of a set of K recognized activities and determine the performance of the different classification models.

2) Unsupervised human activity pattern discovery

While the supervised models try to learn a function from labeled activity data to accurately predict a new activity sample, the goal of the unsupervised method is to discover the patterns of the data. Typically, clustering involves the method of organizing similar objects from a given collection of objects based on specific patterns, aiming to maximize similarity within each class while minimizing similarity between different groups. The activity representations A_k may also be used for an unsupervised pattern discovery task. Five clustering algorithms, including K-means, spectral, agglomerative, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), and Gaussian Mixture Models (GMM) clustering algorithms, have been considered for this study. As the task can be performed in an unsupervised manner, no training data is required, and the dataset can be directly used for testing the models. The clustering of the activities within the dataset is based on a known k number of activities.

E. Performance measure

Accuracy score was used as a performance measure for supervised human activity recognition, while F1-score and Silhouette coefficient score were used as performance measures for unsupervised classification tasks. Accuracy score considers the correctly predicted observations out of the total number of observations and is defined as follows:

$$Accuracy\ Score = \frac{TP + TN}{TP + FP + FN + TN} \quad (6)$$

where, TP = true positive, TN = true negative, FP = false positive, FN = false negative.

F1-score considers the precision and recall for each class and can generate a better assessment of performance for some cases. In situations where the costs of false positives and false negatives significantly differ, relying solely on accuracy may not provide an accurate assessment. Instead, it is advisable to consider both precision and recall. Precision measures the proportion of correctly identified positive cases among all cases predicted as positive, while recall measures the proportion of correctly identified positive cases among all actual positive cases. For this study f1-score was used as an evaluation metric for evaluating clustering performance utilizing the available true labels.

$$F1 - score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (7)$$

where Recall and Precision are defined as:

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (8)$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negetive} \quad (9)$$

Cluster validation index is used to approximate the number of activities which clustering algorithms are able to distinguish assuming true labels are unavailable. The silhouette coefficient can be described as follows:

$$s = \frac{b - a}{\max(a, b)} \quad (10)$$

where a is the average distance between all data points in the same cluster and b is the average distance between all activity clusters. The silhouette coefficient ranges between 0-1 and the highest score is considered the best number of clusters.

4. Results

A. Features used

3D positional $\mathbf{p}_i$ and orientation $\mathbf{o}_i$ information of all joints were obtained and used to derive three different features: $\mathbf{p}_{\text{ref}\,i}$, d_i, and $\mathbf{o}_{\text{ref}\,i}$, which may be used to encode a particular frame instance I_j. The kth activity, A_k, can then be represented as a sequence of n of these frame instances, I_j. These representations of activities were used as input to the classification and clustering models for supervised activity recognition and unsupervised pattern discovery tasks. In this paper, the effectiveness of simple features, including the transformed joint position, Euclidean distance between the hip-center to other body joints, as well as joint orientation features, has been evaluated for different cases:

1. Features using the 3D translated positional information, $f_i = \{\mathbf{p}_{\text{ref}\,i}\}$, for all i, while considering activities to be represented with a fixed n number of frame instances from Table I. Henceforth, this set of features shall be referred to as transformed joint positions.

2. Features using the relative distance from the hip-center joint, $f_i = \{d_i\}$, for all i, while considering activities to be represented with a fixed n number of frame instances from Table I. Henceforth, this set of features shall be referred to as the pairwise Euclidean distance between the hip-center (PED-HC) joint to other body joints feature sets or PED-HC.

3. Features using the orientation information of the joint, $f_i = \{\mathbf{o}_{\text{ref}\,i}\}$, for all i, while considering activities to be represented with a fixed n number of frame instances from Table I. Henceforth, this set of features shall be referred to as joint orientations.

B. Supervised Classification Task

Classification results of the supervised classification models, using transformed joint positions, joint orientations, and pairwise Euclidean distance between the hip-center (PED-HC) joint to other body joints feature sets for activity representations on the different datasets are shown in Figure 6.

Random forest classifier with the transformed joint positions achieved the highest accuracy of 68% on the NTU RGB+D 120 dataset among all the features and algorithms used, as shown in Figure 6(a). The NTU RBG+D 120 dataset is the most challenging dataset with a large number of activities and human subject variations, containing 120 activities performed by 106 subjects. Most classifiers struggled to model the large number of activities in the dataset. Accuracies were lower with the joint orientation and PED-HC features.

With the transformed joint features, KNN, SVM, and RF achieved over 95% accuracy on the CAD60 dataset, with the lowest accuracy of 93% achieved by the DT classifier. Similarly, average accuracy on the Florence 3D dataset using the transformed joint features was above 95%. Other features (PED-HC and joint orientations) demonstrated lower overall accuracies. The highest and lowest accuracy were 98% and 92%, respectively, achieved with RF and DT classifiers. The SVM and KNN classifier with the transformed joint position achieved the highest accuracy of over 90% on the MSR Action dataset while the RF classifier achieved 89% accuracy. Accuracy scores on the experimental dataset with the same feature set varied between 87% - 98%. The RF classifiers achieved 98% accuracy while the lowest was 87% with the DT classifier. Transformed joint and joint orientation features achieved overall higher accuracy.

K-Nearest Neighbour (KNN), Random Forest (RF), and support vector machine (SVM), classifiers achieved relatively high accuracies for all the feature sets on most of the datasets that have a limited number of actions performed by a smaller number of actors. The performance of the DT classifier was lower on most of the datasets for all the features used. Among the three features used, the pairwise Euclidean distance between the hip centre joints and all other joints i.e. PEF-HC, demonstrated the lowest overall accuracy on all of the datasets. Results from the datasets used indicate that the transformed joint positions can be used to model most of the activities efficiently.

The random forest (RF) algorithm achieved the highest overall accuracy on most of the datasets, with the lowest RF accuracy obtained on the NTU RGB+D 120 dataset. One can possibly tune the hyperparameters and apply a feature reduction technique similar to the one used by authors in [60] to further improve classification accuracy on the NTU RGB+D 120 dataset. However, since the study focused on investigating simplistic features and learning

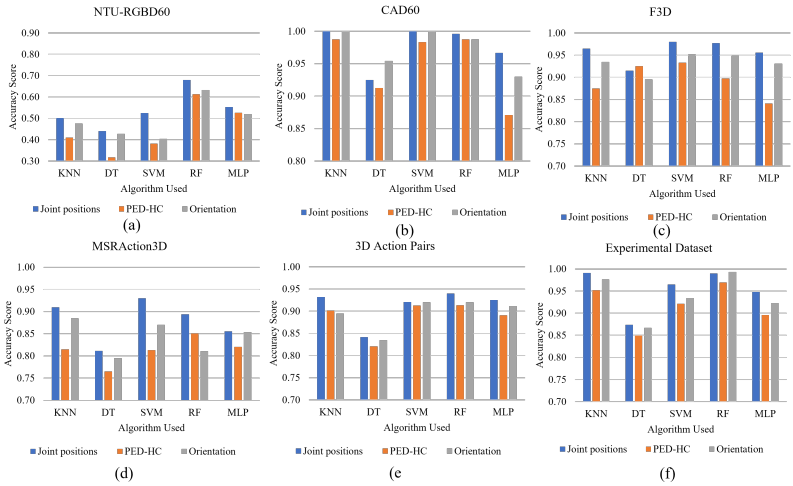


Figure 6. Average classification accuracy of the six datasets used (a) NTU RGB+D 120 dataset (b) CAD60 dataset (c) F3D dataset (d) MSR Action 3D dataset (e) 3D action pairs dataset, and (f) Experimental dataset.

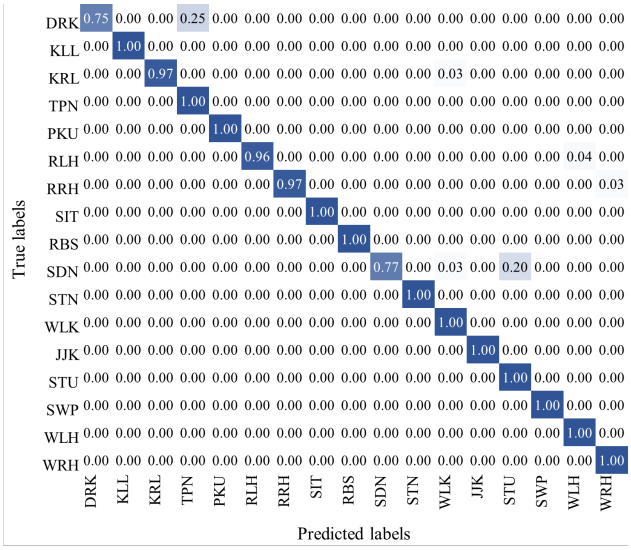


Figure 7. Confusion matrix of the SVM classifier on the experimental dataset, using features on joint position.

algorithms with the availability of data labels, we avoided hyperparameter tuning.

The performance of SVM is of particular interest since it has been widely used for various classification tasks as shown in the literature. For this study, the performance of SVM has shown varying results between the two considered datasets. Over 95% accuracy was obtained using the transformed joint position and joint orientation features, whilst the PED-HC feature gave an accuracy of 92% on the experimental dataset. The confusion matrix of the SVM classifier using the joint position feature on the experimental dataset is plotted in Figure 7. It can be seen that a few activity instances for drinking (DRK) and standing up (SDN) have been wrongly classified as talking on the phone (TPN) and sitting down (STU), respectively.

A comparison of our findings with state-of-the-art algorithms that used hand-crafted features listed in the review in reference [19] is shown in Table II. The authors have made the source codes available, with hyperparameter values available in reference [17]. HON4D, HOPC, LARP-SO-FTP, and HDG-jpd belong to the category of state-of-the-art skeleton features-based methods. These methods primarily rely on extracted skeletal joint information for action recognition. Among them, LARP-SO-FTP stands out with impressive performance across most datasets, demonstrating its effectiveness in capturing meaningful skeletal information. HON4D and HOPC also perform reasonably well, demonstrating the robustness of these skeleton-based techniques. However, HDG-jpd lags behind in terms of accuracy on several datasets, indicating that its approach may not be as effective in handling the complexities of different action recognition scenarios.

DT, RF, KNN, MLP, and SVM represent machine-learning methods that leverage transformed skeleton joints for recognition. In this category, Random Forest (RF) consistently performs exceptionally well across multiple datasets, showcasing its ability to harness skeletal information effectively. Support Vector Machine (SVM) also demonstrates strong performance, especially on the MSRAction3D dataset. KNN and MLP show competitive results but may have limitations in dealing with complex action recognition tasks. Decision Trees (DT), on the other hand, exhibit mixed performance, with some notable disparities across datasets. These findings underscore the significance of carefully selecting the machine learning algorithm depending on the specific dataset characteristics and requirements for skeleton-based action recognition tasks.

Graph Convolutional Network (GCN) [46], 3Mformer [48], and PoseConv3D [47] fall into the category of deep learning-based methods that utilize skeleton joints. These deep learning approaches are inherently more complex and capable of capturing intricate patterns, delivering competitive performance even when compared to well-established machine learning algorithms. Additionally, the dataset NTU RGBD, renowned as one of the largest 3D skeleton-based datasets available, played a crucial role in enabling these advancements. While all of the deep learning-based methods achieve remarkable accuracy on the NTU RGBD dataset, 3Mformer stands out among them with an accuracy of 92.3%. All of the deep learning-based methods had shown mixed results on smaller datasets. PoseConv3D leads the table with an accuracy of 97.1% on the 3D Action Pairs dataset. These results indicate that deep learning methods have the potential to excel in 3D action recognition tasks, particularly when dealing with complex and diverse datasets such as the NTU RGBD dataset.

It was observed that machine learning algorithms with the transformed skeleton data outperform some of the state-of-the-art algorithms that used extracted skeleton features. In many cases, the random forest algorithms performed better than other algorithms, followed by KNN. The DT classifier achieved the lowest accuracy score. In summary, our findings emphasize the substantial progress achieved in skeleton feature-based action recognition, with a shift towards deep learning methods, and highlight the importance of large and diverse datasets such as NTU RGBD in driving innovation in this domain.

C. Unsupervised Classification Task

As shown in the previous section, simplistic features and machine learning algorithms can achieve good accuracy scores when the dataset is labeled. In this section, we present the results of grouping activities using clustering methods assuming data labels are unavailable. In order to simplify the unsupervised activity pattern discovery task, the clustering process was first evaluated based on a known number of clusters for datasets, i.e., K=10 and K=17 clusters for the CAD60 and experimental datasets, respectively. For the CAD60 dataset, only 10 actions were used for clustering since two activities were only performed once by all actors. F1-scores of all the considered clustering methods, using the feature sets used for activity representations, are summarized in Figure 9 and Figure 10, on the CAD60 and experimental datasets, respectively. The highest F1-score achieved on the CAD60 dataset is 87% with the K-means clustering using the transformed joint position feature. The K-means clustering achieved an average F1-score of 79%, whilst the agglomerative clustering gave an average F1-score of 77% on the CAD60 dataset.

The agglomerative clustering using joint position scored the highest F1-scores of 71% on the experimental dataset, whilst the K-means and GMM clustering achieved the highest F1-score of 69% using the transformed joint position feature. Overall, the highest F1-scores were achieved using transformed joint position features on the experimental dataset.

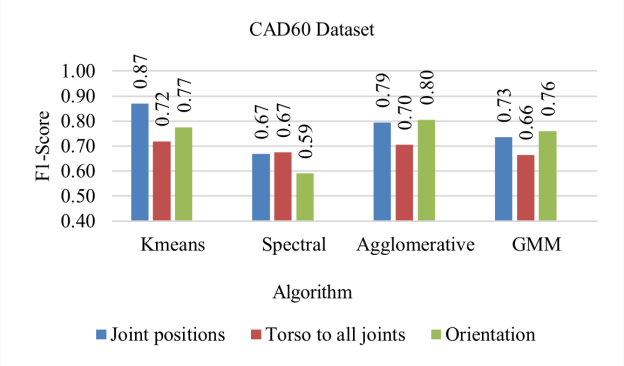


Figure 8. Comparison of average f1-score of all clustering methods on the CAD60 dataset

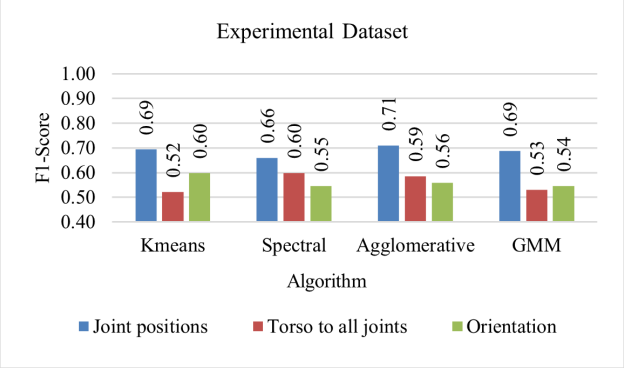


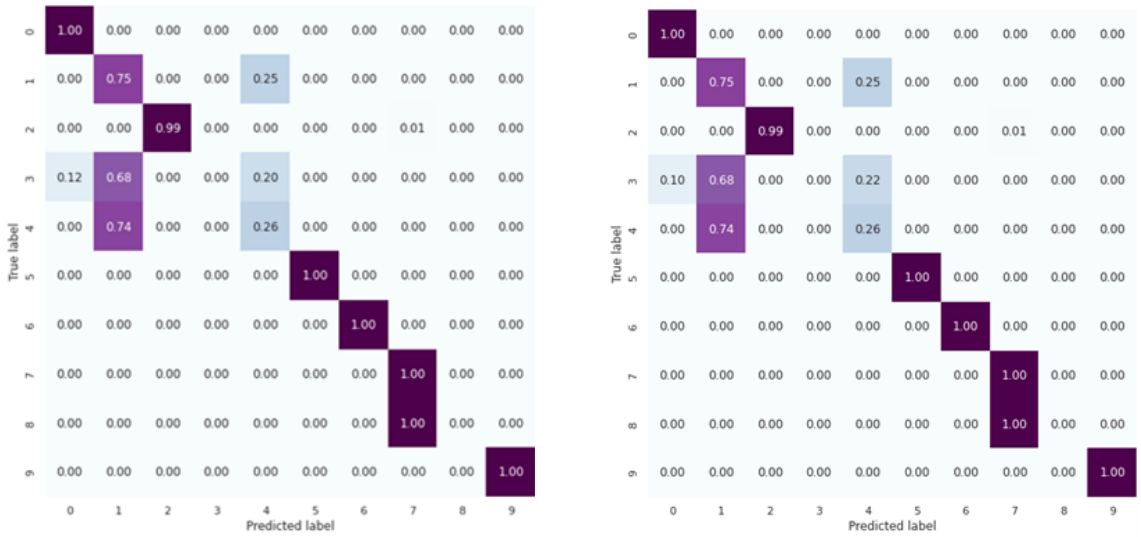
Figure 9. Comparison of average F1-score of all clustering methods on the experimental dataset

Comparing the F1-scores of the clustering algorithms in Figure 8 and Figure 9 shows that the unsupervised activity detection performs better on the CAD60 dataset than the experimental dataset; due to the fact that the experimental dataset contains activities with almost similar postures and complex activities involving longer time duration. Consequently, the clustering algorithms have failed to cluster these almost similar activities into respective groups. Since k-means and agglomerative clustering have demonstrated the highest F1-score on both datasets, only the performances of these two methods were further evaluated.

Since the number of expected clusters is known and the ground truth is available, this knowledge has been used to plot the confusion matrix by replacing the cluster membership with true labels. These are given in Figure 10(a) and Figure 10(b) for the K-means and agglomerative clustering, respectively. Though the aim of the clustering is to identify ten clusters from the CAD60 dataset, it can be observed that only seven clusters can be identified with the k-means algorithm; with activities with almost similar

TABLE II. Comparison of our findings with state-of-the-art algorithms

Method	MSRAction3D	3D Action Pairs	CAD-60	F3D	NTU RGBD	Experimental Dataset
HON4D [29] (Depth)	82.1	96.1	72.6	-	30.8	81.5
HOPC [31] (Depth)	85.51	92.42	47.65	-	40.0	73.12
LARP-SO-FTP [32] (Skel.)	89.45	94.65	78.66	64.2	52.2	71.02
HDG-jpd [30] (Skel.)	55.5	53.84	56.08	83.5	39.8	58.15
DT	81.15	84.15	92.5	71.42	54.5	87.36
RF	89.5	94.2	99.8	88.41	68.1	98.8
KNN	91.01	93.2	99.1	80.5	50.1	98.1
MLP	86.2	93.5	97.2	74.3	55.7	94.77
SVM	93.1	92.1	99.6	92.04	53.2	97.01
GCN [46]	92.03	96.03	-	-	89.5	98.3
3Mformer [48]	90.5	92.0	-	-	92.3	94.21
PoseConv3D [47]	96.5	97.1	-	-	91.7	96.15



(a) (b)

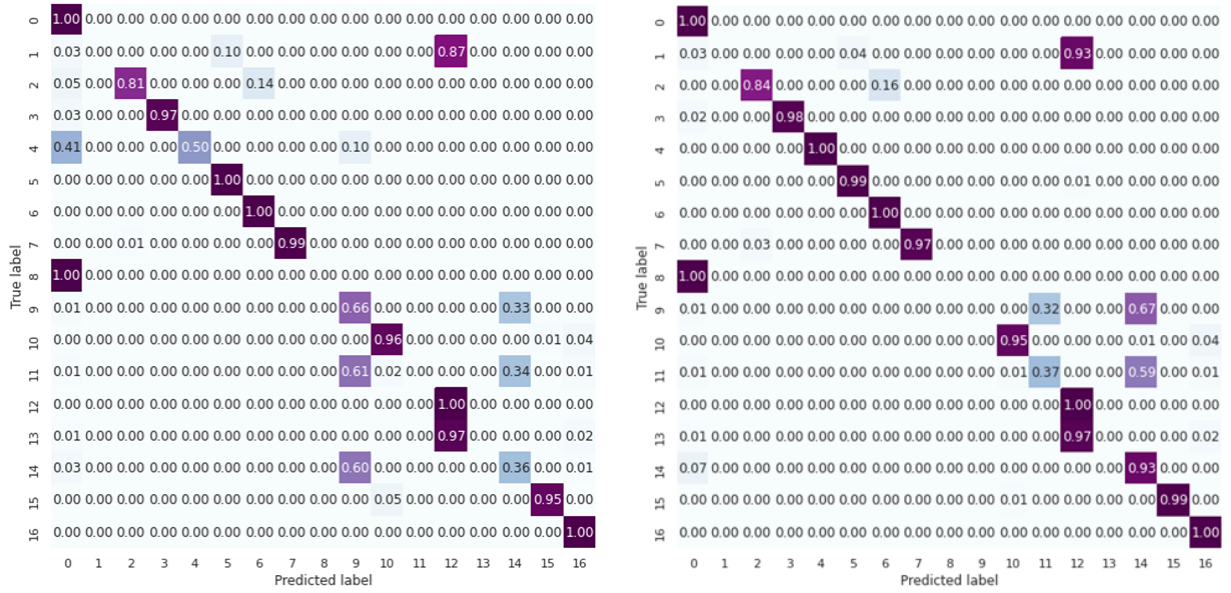
Figure 10. Confusion matrix of the CAD60 dataset with K=10, with (a) K-means clustering. (b) Agglomerative clustering

postures clustered into one larger cluster. For instance, on the CAD60 dataset, talking on the phone (labelled 1) activity has been confused with drinking (labelled 3) and brushing teeth (labelled 4) activities, whilst cooking (chopping) (labelled 7) and cooking (stirring) (labelled 8) in CAD60 dataset have been clustered into the same cluster as depicted in 10. The confusion matrices on the experimental dataset with the K-means and agglomerative clustering are given in Figure 11(a) and Figure 11(b), respectively.

The goal is to identify the 17 activities; however, only 10 clusters have been identified with the k-means algorithm, whilst only eleven clusters have been identified with the agglomerative clustering. Walking (labelled 8) activities have been confused with Standing (labelled 0) activities, and raising right hand (labelled 1), Phone call (labelled 12), and drinking (labelled 13) activities have been clustered into a single large cluster with both k-means and agglomerative clustering algorithms. Likewise, inaccurate clustering has been detected for sitting down (labelled 9), standing up (labelled 11) and picking up from the floor (14) activities.

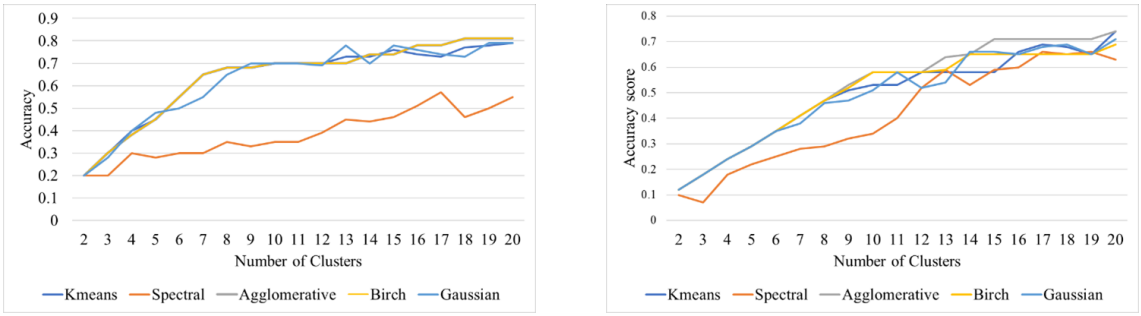
Figure 12 shows the number of clusters K, ranging from 2-20 and accuracy scores for both datasets. Investigations have shown that accuracy increases when the number of clusters is higher than the actual number of activities, due to the label shifting mechanism used. For instance, if true labels for a dataset with 3 activities are [1,1,1,2,2,2,3,3] and predicted labels are [1,1,2,2,3,3,4,4], label shifting may label them as [1,1,1,2,2,2,3,3] despite the presence of the fourth clusters; which may increase the accuracy of the clustering algorithms.

An individual may perform a variety of activities. It is therefore essential to validate the effectiveness of the unsupervised learning methods without specifying a known number of clusters. An internal cluster validation index, namely the silhouette coefficient score has been used to approximate the number of activities in both datasets to identify activities which clustering algorithms are able to distinguish. The silhouette coefficient scores for clusters ranging from 2-20 on both datasets are shown in Figure 13, for clustering performed using the k-means algorithm. The



(a) (b)

Figure 11. Confusion matrix of the CAD60 dataset with K=10, with (a) K-means clustering. (b) Agglomerative clustering



(a) (b)

Figure 12. Number of clusters and accuracy score on both datasets (a) CAD60 dataset (b) Experimental dataset

best score on the CAD60 dataset was 0.58% with 7 clusters, whilst, on the experimental dataset, the highest score of 0.39% occurred with 10 clusters. Seven and ten clusters for the CAD60 and experimental datasets, respectively, as given via the highest silhouette coefficients, are cross-validated with the confusion matrix shown in Figure 11(a) and Figure 11(b). As can be seen, the confusion matrix has indicated that the clustering algorithms clustered the activities into 7 and 10 clusters in Figure 11(a) and Figure 11(b), respectively, while the expected number of clusters were 10 and 17 clusters, for the CAD60 and experimental datasets, respectively. Some of the activities have been clustered well whilst similar activities have been grouped into larger clusters in both datasets. This knowledge of clustering can be used to group activities from a dataset which does not have labels and each cluster can be labelled as an activity which simplifies the labelling task.

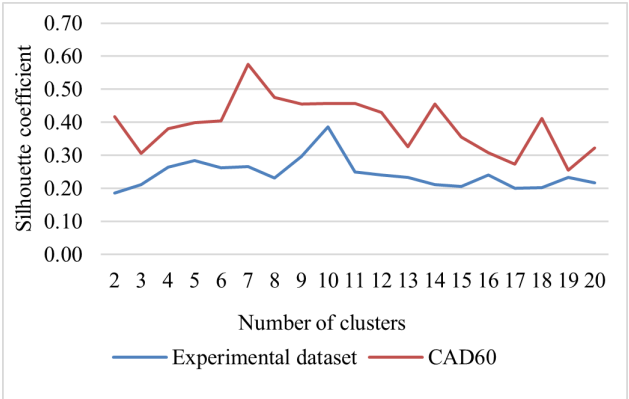


Figure 13. Comparison of average F1-score of all clustering methods on the experimental dataset

5. Conclusion

In this study, investigations have been presented using both supervised and unsupervised learning methods for human activity detection using skeleton data employing six different datasets. Analyses have shown that commonly used supervised learning algorithms such as random forest or KNN classifier can accurately model activities from the skeleton features derived from RGB-D sensors without extensive feature engineering/feature learning on smaller datasets. The average classification accuracies on some of the datasets were above 90% for most of the classifiers. While the state-of-the-art deep learning methods achieved over 90% accuracy on larger datasets. It is worth noting that our study is in line with a growing body of evidence suggesting that machine learning models excel with hand-crafted features on smaller datasets, while deep learning-based models perform better on larger datasets.

On the other hand, unsupervised learning algorithms for activity detection using the same set of features have been proven to be more challenging with the highest f1-score of 87% on the CAD60 dataset without using any labelled data. It remains challenging for unsupervised learning algorithms to distinct activities, which can only be differentiated by subtle differences when represented as skeleton posture. There are other issues to be addressed in addition to enhancing clustering performance, which indicates possible future research directions. Datasets used in the literature are manually segmented and labelled into respective activities, which remains an unresolved problem in human activity research. This investigation contributes to the broader understanding of the strengths and weaknesses of different methods in human action recognition, offering valuable insights for future research in this field.

Acknowledgment

This research was funded by Universiti Brunei Darussalam Bursary Award program.

References

[1] Y. Kwon, K. Kang, and C. Bae, "Unsupervised learning for human activity recognition using smartphone sensors," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6067–6074, 2014.

[2] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: A survey," *Procedia Computer Science*, vol. 155, pp. 698–703, 2019.

[3] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," no. 952215, 2022.

[4] S. Seneviratne, Y. Hu, T. Nguyen, G. Lan, and S. Khalifa, "A survey of wearable devices and challenges," no. July, 2017.

[5] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.

[6] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using skeleton data from rgbd sensors," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.

[7] M. A. Hossen and P. E. Abas, "Social distance monitoring using a low-cost 3d sensor," 2023. [Online]. Available: https://doi.org/10.36227/techrxiv.23516505

[8] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 1010–1019, 2016.

[9] S. K. Yadav *et al.*, "Csitime: Privacy-preserving human activity recognition using wifi channel state information," *Neural Networks*, vol. 146, pp. 11–21, 2022.

[10] J. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[11] F. Cruciani *et al.*, "Comparing cnn and human crafted features for human activity recognition," in *Proceedings - 2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Internet of People and Smart City Innovation, SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*, 2019, pp. 960–967.

[12] P. Hadikhani, D. T. C. Lai, and W.-H. Ong, "A novel skeleton-based human activity discovery technique using particle swarm optimization with gaussian mutation," 2022. [Online]. Available: http://arxiv.org/abs/2201.05314

[13] H. Gjoreski and D. Roggen, "Unsupervised online activity discovery using temporal behaviour assumption," *Proceedings - International Symposium on Wearable Computers, ISWC*, vol. Part F1305, no. September, pp. 42–49, 2017.

[14] S. Abudalfa and H. Qusa, "Evaluation of semi-supervised clustering and feature selection for human activity recognition," *International Journal of Computer and Digital Systems*, 2019.

[15] M. A. Hossen, O. W. Hong, and W. Caesarendra, "Investigation of the unsupervised machine learning techniques for human activity discovery," in *Proceedings of the 2nd International Conference on Electronics, Biomedical Engineering, and Health Informatics*, 2022, pp. 499–514.

[16] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[17] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[18] H.-B. Zhang *et al.*, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb 2019.

[19] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 15–28, 2020.

[20] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," pp. 1–8, 2020. [Online]. Available: http://arxiv.org/abs/2002.05907

[21] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2014, pp. 804–811.

[22] R. Slama, H. Wannous, and M. Daoudi, "Grassmannian representation of motion depth for 3d human gesture and action recognition," in *2014 22nd International Conference on Pattern Recognition*, Aug 2014, pp. 3499–3504.

[23] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. D. Bimbo, "Space-time pose representation for 3d human action recognition bt - new trends in image analysis and processing – iciap 2013," 2013, pp. 456–464.

[24] L. Gan and F. Chen, "Human action recognition using apj3d and random forests," *Journal of Software*, vol. 8, Sep 2013.

[25] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[26] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun 2012, pp. 20–27.

[27] A. Taha, H. H. Zayed, M. E. Khalifa, and E.-S. M. El-Horbaty, "Human activity recognition for surveillance applications," in *Proceedings of the 7th International Conference on Information Technology*, 2015, pp. 577–586.

[28] M. A. Hossen, A. G. Naim, and P. E. Abas, "Evaluation of 2d and 3d posture for human activity recognition," *AIP Conference Proceedings*, vol. 2643, no. 1, p. 40013, 2023.

[29] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 716–723.

[30] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014*, 2014, pp. 626–633.

[31] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2430–2443, 2016.

[32] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 588–595.

[33] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[34] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng, "Online human gesture recognition from motion data streams," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 23–32.

[35] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1809–1816.

[36] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[37] R. Murugaiya, P. E. Abas, K. Mohanchandra, and D. S. Liyanage, "Robust cepstral feature for bird sound classification," *International Journal of Electrical and Computer Engineering*, vol. 12, 2022.

[38] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations (ICLR 2015) - Conference Track Proceedings*, 2015, pp. 1–15.

[39] F. A. Azis, H. Suhaimi, and E. Abas, "Waste classification using convolutional neural network," in *Proceedings of the 2020 2nd International Conference on Information Technology and Computer Communications*, 2020, pp. 9–13.

[40] M. A. Humayun, H. Yassin, and P. E. Abas, "Native language identification for indian-speakers by an ensemble of phoneme-specific, and text-independent convolutions," *Speech Communication*, vol. 139, pp. 92–101, 2022.

[41] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *American Journal of Veterinary Research*, vol. 39, no. 9, pp. 1442–1446, Aug 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[42] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, vol. 2015-January, 2015, pp. 577–585.

[43] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "Rnn fisher vectors for action recognition and image annotation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9910 LNCS, 2016, pp. 833–850.

[44] G. Ch, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *ICCV (IEEE International Conference on Computer Vision)*, 2015, pp. 3218–3226.

[45] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2018, pp. 7444–7452.

[46] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2669–2676.

[47] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[48] L. Wang and P. Koniusz, "3mformer: Multi-order multi-mode transformer for skeletal action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5620–5631.

[49] A. Fataniya and H. Modi, "Comprehensive analysis of deep learning-based human activity recognition approaches based on accuracy," *International Journal of Computer and Digital Systems*, vol. 12, pp. 1097–1118, 2022.

[50] P. A. Colpas, E. Vicario, E. De-La-Hoz-Franco, M. Pineres-Melo, A. Oviedo-Carrascal, and F. Patara, "Unsupervised human activity recognition using the clustering approach: A review," *Sensors (Switzerland)*, vol. 20, no. 9, 2020.

[51] H. Mohammadzade and M. Tabejamaat, "Sparseness embedding in bending of space and time; a case study on unsupervised 3d action recognition," *Journal of Visual Communication and Image Representation*, vol. 66, p. 102691, 2020.

[52] W. H. Ong, T. Koseki, and L. Palafox, "An unsupervised approach for human activity detection and recognition," *International Journal of Simulation: Systems, Science and Technology*, vol. 14, no. 5, pp. 42–49, 2013.

[53] J. S. et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[54] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 842–849.

[55] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 479–485.

[56] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2020.

[57] A. Franco, A. Magnani, and D. Maio, "Joint orientations from skeleton data for human activity recognition," in *International Conference on Image Analysis and Processing*, vol. 2, 2017, pp. 152–162.

[58] M. Sharif, M. A. Khan, T. Akram, M. Y. Javed, T. Saba, and A. Rehman, "A framework of human detection and action recognition based on uniform segmentation and combination of euclidean distance and joint entropy-based features selection," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, 2017.

[59] F. P. et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, 2012.

[60] M. Ramashini, P. E. Abas, U. Grafe, and L. C. D. Silva, "Bird sounds classification using linear discriminant analysis," in *2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, 2019, pp. 1–6.

Md Amran Hossen is a PhD candidate at the Faculty of Integrated Technologies, Universiti Brunei Darussalam. His academic journey began at the same institution, where he pursued a bachelor's degree in Computer Science, followed by a master's degree by research. His primary research revolves around the innovative application of various machine learning algorithms to comprehend and analyse human activities. His expertise encompasses Unsupervised Machine Learning, Deep Learning and Computer Vision Algorithms.

Pg Dr. Emeroylariffion Abas is a Senior Assistant Professor at the Faculty of Integrated Technologies, Universiti Brunei Darussalam. He is also the Director of the Office of AVC (Innovation and Enterprise). Before joining the University, he completed his bachelors and Phd from Imperial College London majoring in Systems Engineering.