Soft Computing (2018) 22:7027–7039
https://doi.org/10.1007/s00500-018-3364-x

FOCUS

# Human activity learning for assistive robotics using a classifier ensemble

David Ada Adama[1] · Ahmad Lotfi[1] · Caroline Langensiepen[1] · Kevin Lee[1] · Pedro Trindade[1]

Published online: 19 July 2018


**Abstract**
Assistive robots in ambient assisted living environments can be equipped with learning capabilities to effectively learn and execute human activities. This paper proposes a human activity learning (HAL) system for application in assistive robotics. An RGB-depth sensor is used to acquire information of human activities, and a set of statistical, spatial and temporal features for encoding key aspects of human activities are extracted from the acquired information of human activities. Redundant features are removed and the relevant features used in the HAL model. An ensemble of three individual classifiers—support vector machines (SVMs), $K$-nearest neighbour and random forest—is employed to learn the activities. The performance of the proposed system is improved when compared with the performance of other methods using a single classifier. This approach is evaluated on experimental dataset created for this work and also on a benchmark dataset—the Cornell Activity Dataset (CAD-60). Experimental results show the overall performance achieved by the proposed system is comparable to the state of the art and has the potential to benefit applications in assistive robots for reducing the time spent in learning activities.

**Keywords** Human activity learning · Feature extraction · Classifier ensemble · Assistive robotics · Activity classification.

## 1 Introduction

Ambient assisted living (AAL) is an active research area that has attracted a lot of interest in recent years through the development of various solutions to enable independent living and promote quality of life and well-being for an ageing human populace (Blackman et al. 2016). AAL solutions utilise assistive robots and other technologies to aid in daily routine activities. These robots are incorporated in various applications which involve human–computer interaction that traverse humans of all ages. Such applications include care for older adults (Xiao et al. 2014; Jayawardena et al. 2016).

However, due to the dynamic nature of the environment in real-world applications, it is quite challenging to have assistive robots execute functions easily. A specific case is assistive robots that can interact with older adults as carers. These robots learn tasks by observing a human carer execute the tasks. Such robots learn human activities by extracting descriptive information of the activities in order to classify them as they are executed. This process involves a transfer of knowledge/information of the activity performed which is referred to as *transfer learning* (Weiss et al. 2016).

Regardless of the method applied to learning an activity by a robot, there is a knowledge gap contained in the varied information acquired of a person executing an activity and a robot carrying out a similar activity. Transfer learning helps to bridge this gap by providing faster learning of activities and better collaboration of assistive robots in AAL environments (Helwa and Schoellig 2017). A conceptual overview of the processes involved in learning of human activities for assistive robotics is given in Fig. 1. It is evident in this context that the ability to correctly recognise a human activity and correctly learn (as highlighted in steps 1–3 of Fig. 1) such activity plays a significant role in the amount of knowledge which can be transferred to an assistive robot to be used in learning.

To obtain information of human activities as they are executed, recent research has made use of visual sensors (e.g. RGB-D sensors) (Sung et al. 2011, 2012; Han et al. 2017) and non-visual sensors (e.g. wearable sensors) (Capela et al. 2015) which make it a lot easier to obtain information of activities. Although non-visual sensors have certain advan-

Communicated by F. Chao, Q. Zhang.

✉ Ahmad Lotfi
ahmad.lotfi@ntu.ac.uk

[1] School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

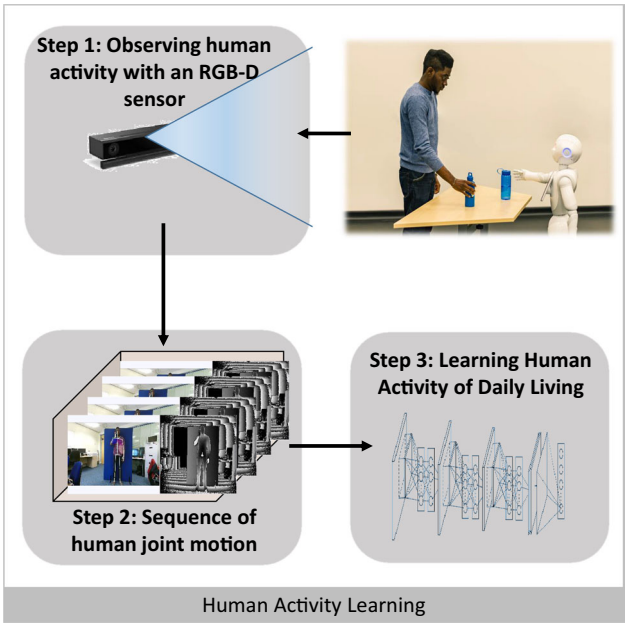



**Fig. 1** A conceptual overview of learning of human activity by an assistive robot using information from an RGB-depth sensor

tages, they are sometimes invasive and burdensome. The development of visual sensors like RGB-D sensors provides a better means to detect human pose used to build human activity recognition systems (Han et al. 2017). These sensors provide platforms for identifying body shape, depth maps and detecting skeleton of human joints in 3D space which can be exploited in learning activities.

The aim of this paper is to propose a human activity learning (HAL) system for assistive robotics. This will act as part of the process of transfer learning for assistive robots. The research presented in this paper is an extension of the system proposed earlier by Adama et al. (2018). The focus is on the three steps shown in Fig. 1. An RGB-D sensor is used to obtain 3D skeleton information of body joints during activities as they are executed by a human. Descriptive features are then extracted from the skeleton information obtained, and the most informative features are selected to be used in training a classifier model. These features are extremely valuable in evaluating the performance of the system because redundant and noisy features can have negative effect on the system performance. An ensemble of classifiers model is used in building the learning model for activities. The approach presented here employs three classifiers—multiclass support vector machines (MSVMs), $K$-nearest neighbour ($K$-NN) and random forest (RF)—in creating the ensemble model. These classifiers are classical algorithms used in machine learning problems. The proposed approach is not only focused on using the selected algorithms but a combination of them in an ensemble. The reason for using an ensemble of classifiers is to improve performance compared with a single classifier model (Tahir et al. 2012). The results discussed in subsequent sections show the improved performance.

The remaining sections of the paper are structured as follows. Section 2 presents a review of relevant related work in this area with emphasis on the main contributions. In Sect. 3, details of the methods applied in 3D data processing and feature representation are explained. Section 4 explains the classifier ensemble model approach for human activity learning. Section 5 presents experimental results and their evaluation, and Sect. 6 summarises the main results and provides discussion of the future work.

## 2 Related work

Learning and classification of human activities using computational intelligence/soft computing techniques is often referred to as human activity recognition (HAR) (Iglesias et al. 2010; Jalal and Kamal 2014). One of the main objectives is to extract descriptive information (i.e. features) from human activities to be able to distinctly characterise and classify one activity from another. An integral component of learning an activity is how information of the activity is obtained or observed. For human activities, information obtained using visual and non-visual sensors makes it a lot easier to understand and learn activities as they are performed. Visual sensors such as RGB cameras can be used to obtain descriptive information of an activity in 2D. However, this information is limited in effectively characterising an activity (Han et al. 2017). Additional depth information using RGB-D sensors provides several advantages as they are better suited to observing human activities and detecting human poses used to build activity recognition systems.

To effectively characterise activities from information obtained using RGB-D sensors, soft computing techniques such as machine learning and reasoning methods have been applied by many researchers (Koppula et al. 2013; Li et al. 2015; Han et al. 2017). These methods provide an understanding of how activities are learned and relationships between activities. However, there is some uncertainty regarding how one actor performing an activity would differ from another actor performing similar activity. This hinders HAR systems from going mainstream.

Data obtained from RGB-D sensors give information relevant for a robot to understand an activity. By exploring human pose detection using RGB-D sensors, activity recognition has advanced recently (Sung et al. 2011; Faria et al. 2014). Using RGB-D sensors extracts 3D skeleton data from depth images and body silhouette for feature generation. In Faria et al. (2014), the RGB-D sensor is used to generate a human 3D skeleton model with matching of body parts linked by its joints. They extract positions of individual joints from the

skeleton in a 3D form *x, y, z*. The authors in Jalal and Kamal (2014) use similar RGB-D sensor to obtain depth silhouettes of human activities from which body points information are extracted for the activity recognition system. Zhou et al. (2018) also used an RGB-D sensor to capture human skeleton information as part of a system for controlling a mobile robot using human gestures which is also a similar application proposed by Chao et al. (2017). Another approach is shown in the work in Gu et al. (2012) where the RGB-D sensor is used to obtain orientation-based human representation of each joint to the human centroid in 3D space. Raw data obtained from these sensors have to be pre-processed. This process is carried out to reduce redundancy in data for better representation of features of an activity.

Classification of human activities is carried out by extracting relevant features from data obtained using RGB-D sensors. In our previous work, a method for activity recognition using RGB-D data was proposed (Adama et al. 2018). The 3D joint position information extracted from the sensor is transformed into feature vectors by applying selected soft computing techniques to group key postures of an activity. The posture features are used as input to a learning algorithm for classification of human activities. SVM and KNN algorithms were used separately in classifying activities and the results compared. In the work by Faria et al. (2014), the authors proposed a combination of multiple classifiers to form a Dynamic Bayesian Mixture Model (DBMM) to characterise activities using features obtained from distances between different parts of the body. Hussein et al. (2013) applied statistical covariance of 3D joints (Cov3DJ) as features to encode the skeleton data of joint positions which are then used as input to an SVM model for activity recognition. Another approach applied by Wei et al. (2013) used a sequence of joint trajectories and applied wavelets to encode each temporal sequence of joints into features used in activity classification. Deep learning neural networks (Ijjina and Chalavadi 2017) have also more recently been applied in activity recognition problems with results showing robustness of the method in activity recognition. However, deep learning neural network systems require large amount of data to achieve for concise predictions of activities and in most cases more resources such as time and reliable processing architectures.

## 3 Methodology for human activity data processing and feature representation

The proposed approach to HAL described in this paper works by extracting features from 3D skeletal data and applying feature selection techniques for selecting the most informative features used in building a learning model for human activities. The overview of the system architecture shown in Fig. 2 illustrates the main stages within the process. This is divided into two stages as follows:

*Stage 1* Model learning

- Data input into the system from a dataset containing 3D skeleton information of human joints. These data are captured using an RGB-D sensor and pre-processed before it is used in training activity classifier ensemble model.
- Features representing activities are computed from the data. This step also includes the selection of optimal features relevant for learning activities.
- Training selected classifier models through supervised learning of activities. The output of this step is the learned classifier ensemble model ready to be utilised in activity classification.

*Stage 2* Activity classification

- Data input in this stage is similar to that described in the model learning stage. However, this has to be unseen data in order to validate the performance of the learned models. The data can be obtained from a dataset or on-the-fly from an RGB-D sensor.
- Similar features are extracted from the data to be classified. This stage differs from the model learning stage in that unlabelled activity data is used, while the model learning stage is based on labelled activity data. The features extracted from unlabelled activity data are passed into the learned classifier ensemble model for identification of activity classes.

### 3.1 3D activity data pre-processing

Human activity is composed of a continuous transformation of a series of human poses. Pre-processing the information is necessary to reduce irregularities in the data obtained from the sensor. RGB-D sensors provide information in three modes, namely RGB image, depth image and skeleton joint coordinates. However, this work uses only the skeleton joint coordinates information. A Microsoft Kinect V2 (Microsoft 2017) RGB-D sensor which has a skeleton model consisting of 25 joints as shown in Fig. 3 is used in this work. From the information obtained from the kinect sensor, 15 key joints as outlined in Fig. 3 are selected for use. Data are acquired from the sensor as frames containing different poses that make up an activity. 3D skeleton joint coordinates $J$ are obtained from pose approximation in each frame (Yang and Tian 2014) with coordinates relative to the sensor position where,

$$J = [j_1, j_2, j_3, \ldots, j_i], \quad \text{for } J \in \mathbb{R}^{3\times d} \tag{1}$$

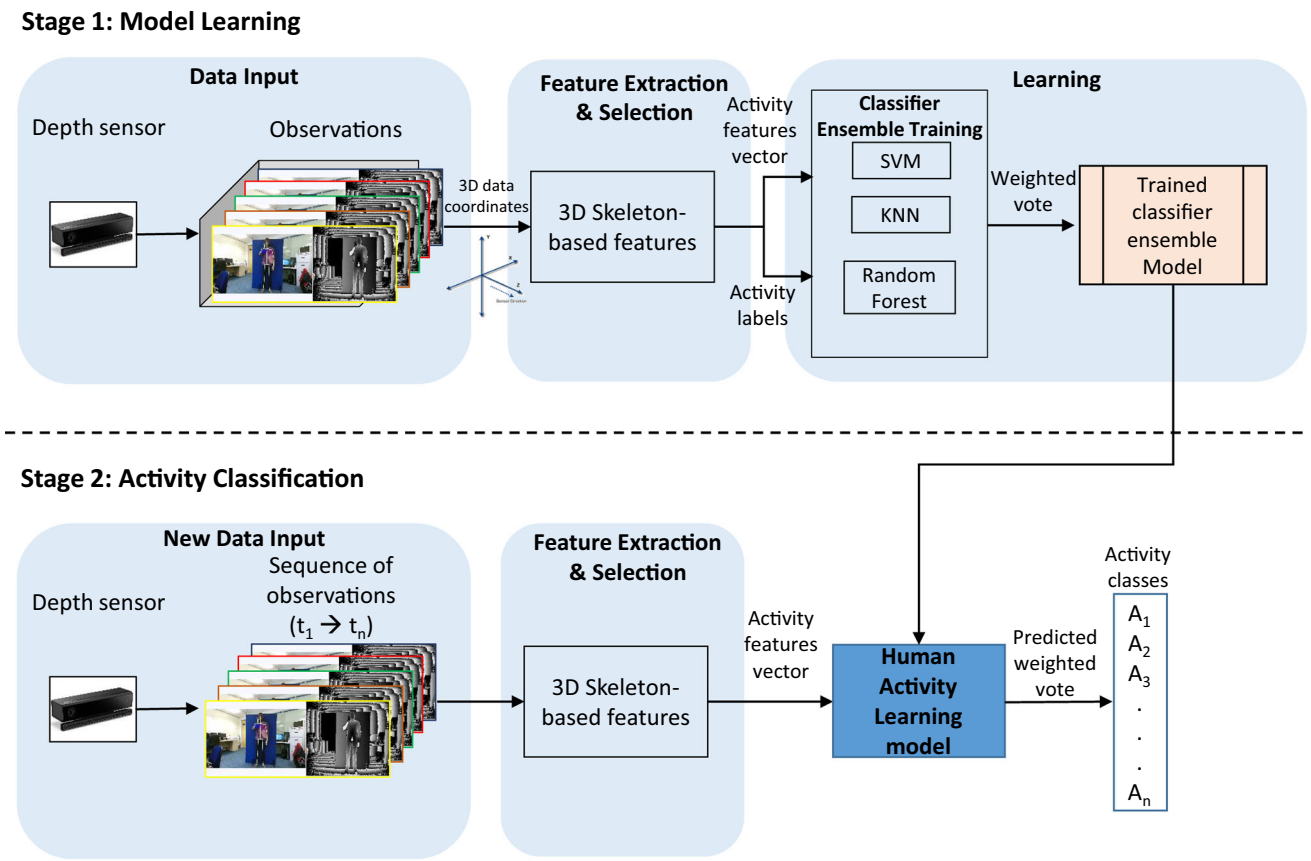



**Fig. 2** Architecture of proposed human activity learning model. Stage 1: Model Learning (top): learning human activities by training a set of classifiers (SVM, KNN and RF) from 3D skeleton features obtained from activity frames captured using an RGB-D sensor. Stage 2: Activity Classification (bottom): observations from human activity are used to extract/ select relevant features which are fed into the trained classifier models, and activities performed are detected

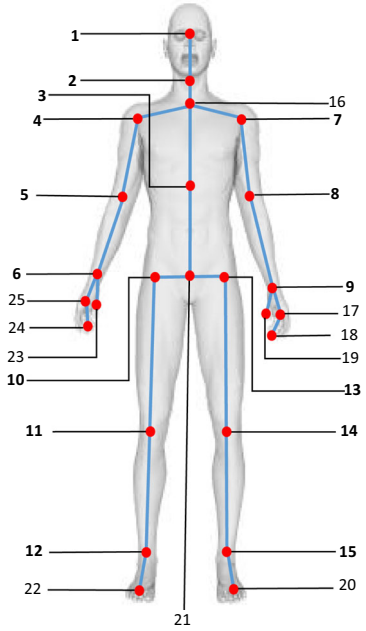
$j_i$ represents the $i$th joint with coordinates $x, y, z$ corresponding to horizontal, vertical and depth positions, respectively, and $d$ is the total number of skeleton joints used.

To make the joint coordinates invariant of the sensor position, the origin of the skeleton is translated along the vector $\overrightarrow{s_o j_t}$, where $s_o$ is the sensor coordinates origin and $j_t$ represents the torso centroid joint of the skeleton. Each joint coordinate position $\overrightarrow{j_i}$ ($j_i$ is a vector representing the $i$th joint coordinates of the skeleton) is computed with reference to the new origin of torso centroid $\overrightarrow{j_i}$ - $\overrightarrow{j_t}$. Thus, the skeleton is independent of the sensor position as shown in Fig. 4. Each sample posture of activity is then reformulated to the torso centroid origin.

Another stage of pre-processing is done to symmetrise the data in order to eliminate ambiguity in gestures performed by left- and right-handed people. This ensures each activity is represented in a variation of its original form as shown in Fig. 5. The symmetry is computed along the $y\text{-}axis$ of the origin (torso centroid).

## 3.2 Extraction and representation of 3D features

Extraction of descriptive information from acquired raw sensor information is crucial to any learning system as raw data do not provide adequate information for learning. This is carried out after the data are pre-processed. In this work, the features used are divided into two distinct categories: joint displacement-based features and statistical features in the time domain. Joint displacement-based features encode information relative to position and motion of body joints (Yang and Tian 2014; Han et al. 2017). This information considers displacement between joints of an activity pose and 3D position differences of skeleton joints across different time periods of an activity. Similarly, statistical time domain features encode information of variations across a collection of activity poses within a specified time domain. The following sections provide details of the features used in this work.




**Joints selected in this work**

1 - Head
2 - Neck
3 - Torso
4 – Left shoulder
5 – Left elbow
6 – Left hand
7 – Right shoulder
8 – Right elbow
9 – Right hand
10 – Left hip
11 – Left knee
12 – Left foot
13 – Right hip
14 – Right knee
15 – Right foot

**Fig. 3** Skeleton representation of Microsoft Kinect V2 with 25 joints. 15 key joints are used in this work as shown in the label definition in the figure

### 3.2.1 Displacement-based features

1. Spatial displacement between selected joint skeletal joint coordinates is computed as the Euclidean distance $\delta$ between any two joints described in Eq. 2. The joints are selected based on relevance to activities.

$$\delta_{(j_m,j_n)} = \sqrt{\sum_{x,y,z}(j_m - j_n)^2}, \tag{2}$$

for $1 \leq (\mathrm{m,n}) \leq i$ and $\mathrm{m} \neq \mathrm{n}$. $j_m$ and $j_n$ are any pair of selected joints with coordinates $x, y, z$.

2. Temporal joint displacement features consider 3D consecutive motion of joints $t_{cp}$ and overall motion dynamic of joints $t_{ci}$. $t_{cp}$ is computed as the joint coordinates position difference between the current pose $c$ and its preceding pose $p$ in Eq. 3 and $t_{ci}$ as the temporal difference between the each joint current pose from the initial pose $i$ in Eq. 4.

$$t_{cp} = [j_m^c - j_n^p]; \quad \text{for } j_m^c \in J^c \text{ and } j_n^p \in J^p \tag{3}$$

$$t_{ci} = [j_m^c - j_n^i]; \quad \text{for } j_m^c \in J^c \text{ and } j_n^i \in J^i \tag{4}$$

### 3.2.2 Statistical features in time domain

This is computed as the projected difference of joint coordinates $j_i$ of the current pose $c$ (also referred to as the current activity frame) from the mean, variance, standard deviation, skewness and kurtosis of joints coordinates for an activity sequence. These are computed as follows:

1. Joint coordinate-mean difference;

$$j_{(i,\mathrm{mean})} = j_i - j_{\mathrm{mean}} \tag{5}$$

where the mean of all positions for a joint coordinate is $j_{\mathrm{mean}} = \frac{1}{N}\sum_{c=1}^{N} j_i$ and $N$ is the sum of poses in an activity.

2. Joint coordinate-variance difference;

$$j_{(i,\mathrm{var})} = j_i - \frac{\sum_{c=1}^{N}(j_i - j_{\mathrm{mean}})^2}{N} \tag{6}$$

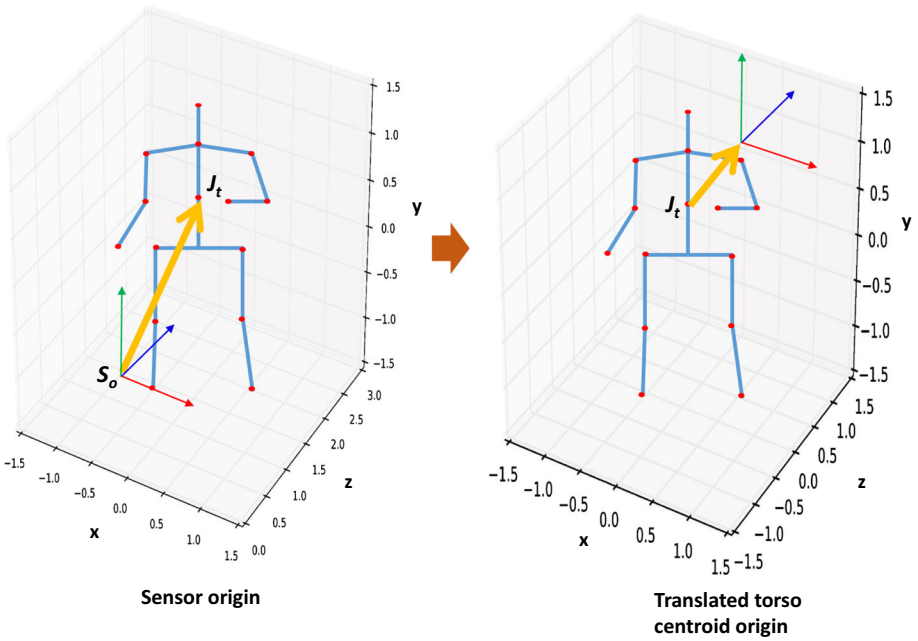



**Fig. 4** Translation of skeleton coordinate system from the sensor origin to the torso centroid origin

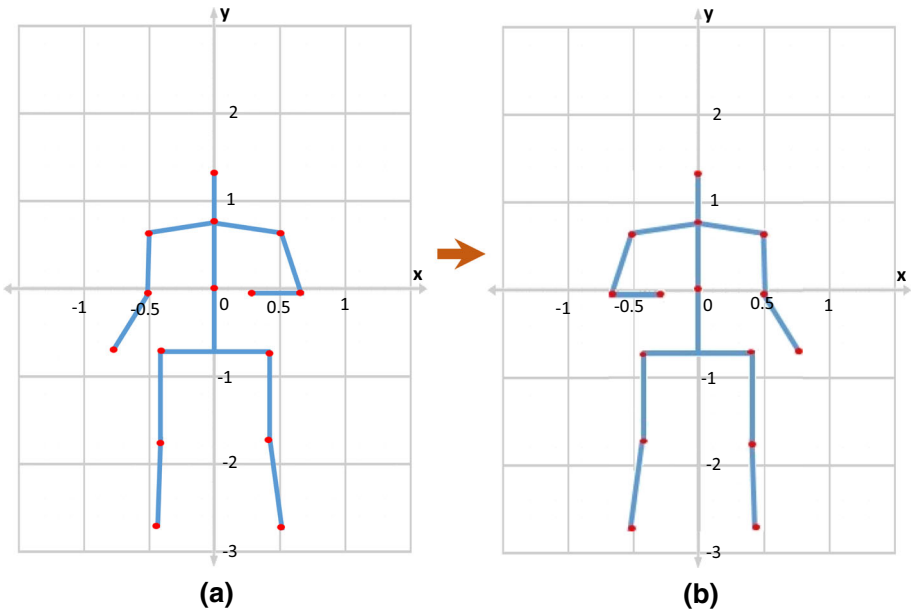



**Fig. 5** Skeleton symmetrisation of an activity posture about the *y-axis*. **a** represents the original activity posture, and **b** is the symmetry obtained of same posture

3. Joint coordinate-standard deviation difference;

$$j_{(i,\text{std})} = j_i - \sqrt{\frac{\sum_{c=1}^{N}(j_i - j_{\text{mean}})^2}{N}} \tag{7}$$

4. Joint coordinate-skewness difference;

$$j_{(i,\text{skw})} = j_i - \frac{\sum_{c=1}^{N}(j_i - j_{\text{mean}})^3}{(N-1)\sigma^3} \tag{8}$$

where $\sigma$ refers to the standard deviation of each joint coordinate for all poses in an activity.

5. Joint coordinate-kurtosis difference;

$$j_{(i,\text{kur})} = j_i - \frac{\sum_{c=1}^{N}(j_i - j_{\text{mean}})^4}{(N-1)\sigma^4} \tag{9}$$

All activity feature vectors computed are concatenated to form a matrix $A$ of extracted activity features in which the columns correspond to feature vectors and the rows correspond to features extracted from different frames of activities. $A$ is represented by the following;

$$A = [\delta, \quad t_{cp}, \quad t_{ci}, \quad \ldots, \quad j_{(i,\text{kur})}] \tag{10}$$

### 3.3 Feature normalisation

HAL systems can be problematic if the extracted features are not well processed. This is due to heterogeneity in features. A further pre-processing of extracted features is needed to deal with the issue of features heterogeneity before classification. This is done through feature normalisation which is often applied in many machine learning applications (Sung et al. 2012; Capela et al. 2015). Normalisation of each feature in the activity features matrix obtained in Eq. 10 is done according to:

$$a_{\text{norm}} = \frac{a_{cf} - \min(A_f)}{\max(A_f) - \min(A_f)} \tag{11}$$

where $a_{cf}$ is a feature on the current pose $c$ of the $f$th column feature vector. The obtained feature matrix after normalisation becomes $A_{\text{norm}}$.

### 3.4 Feature selection

Feature selection is performed on the normalised activity features matrix. This is important to any learning model as it enables faster training, reduces overfitting, improves accuracy and reduces model complexity (making it easier to interpret Gupta and Dallas 2014; Capela et al. 2015). In this paper, a filter method—Relief-F (Kononenko 1994) of feature selection—is applied. Filter methods are preferred to other methods such as wrapper methods since they do not require a fixed learning mechanism and therefore have more generalisation across different learning models (Gupta and Dallas 2014).

The Relief-F method uses a statistical approach rather than heuristic to provide relevance weights to rank potential features. The features ranked above a set threshold are selected for the model. In this paper, the threshold is determined from the number of features that provide the best substitution accuracy with the learning model. The performance achieved using the selected features is presented in the experimental results in Sect. 5.

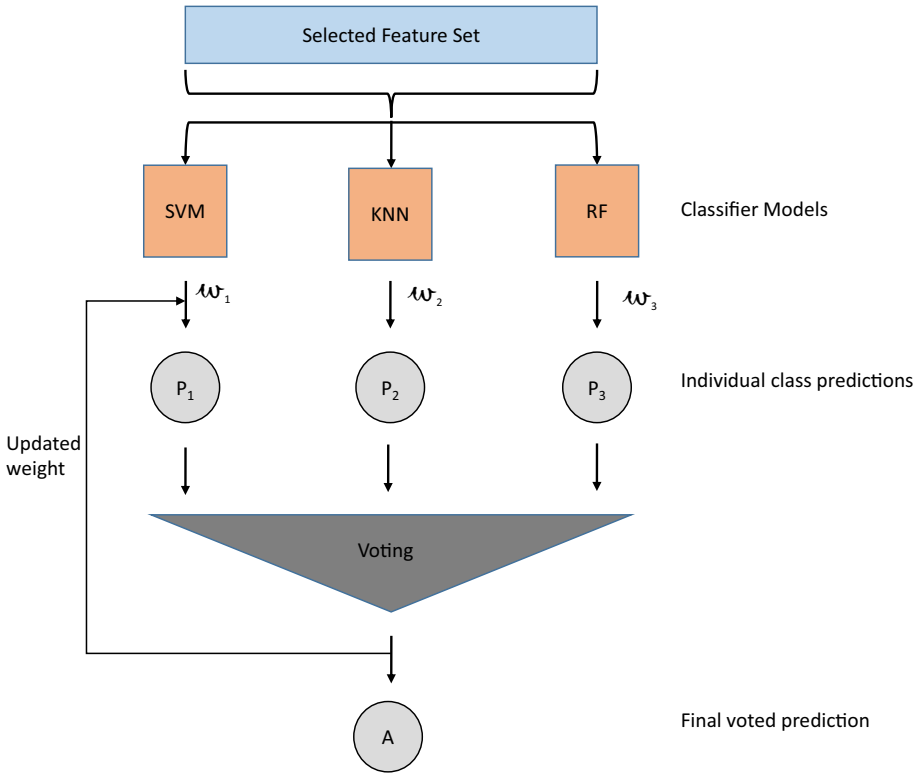



**Fig. 6** Overview of weighted voting architecture of classifier ensemble

## 4 Classifier ensemble model

The final stage in developing an activity learning system is training a classification model with the selected features to achieve a good learning performance score. Building on previous work by Adama et al. (2018) in which a selection of learning models was used separately to identify activities, this work employs a combination of different learning models in a framework referred to as a bagging ensemble of classifiers in order to achieve an improved performance of the system. The use of an ensemble of classifiers model generally allows for better predictive performance than the performance achievable with a single model (Diao et al. 2014; Yao et al. 2016). According to Tahir et al. (2012), ensemble models are learning models that construct a set of classifiers used in classifying new information based on a weighted vote of individual classifier predictions. Three base classifiers are used in this work to construct a bagging ensemble of classifiers: multiclass support vector machines (MSVMs), $K$-nearest neighbour and random forest classifiers. The pictorial overview of the bagging ensemble method applied is shown in Fig. 6.

The weighted votes work by computing the weighted majority vote $\hat{q}$ given in Eq. 12 through allocation of weights $\omega_r$ to each classifier $C_r$.

$$\hat{q} = \arg\max_i \sum_{r=1}^{3} \omega_r \times (C_r(s) = i), \tag{12}$$

where $C_r(s)$ is a classifier characteristic function in a set of unique classifier labels.

The weights assigned to individual classifiers in the ensemble are computed during the learning phase by weighted votes. At the initial stage, uniform weights are set and updated at each iteration of cross-validation. The updated classifier weights used in succeeding iterations are computed as ratios of the average precision obtained in the preceding iteration of each classifier in the ensemble.

The multiclass SVM model follows the configuration reported in Cippitelli et al. (2016) and Adama et al. (2018) which is an extension of a binary classifier. A *one-against-one* approach based on the construction of several binary SVM classifiers suitable for $M$ classes contained in a dataset—where $M > 2$—is implemented as one of the base classifiers. The $K$-NN classifier algorithm is one of the simplest machine learning algorithms used in classifying observations based on the closest training points in the feature space. An instance of observation is assigned to a class most common among its $k$ nearest neighbours by a majority of votes of its neighbours, where $k > 0$. Euclidean distance is used in most cases as a metric in finding nearest neighbours. In the proposed HAL model, a value of $k = 5$ nearest neighbour is used in the configuration. Random for-

est classifier consists of an ensemble of decision trees where each decision tree is trained from randomly selected samples of an original training set. In this work, RF is used with 10 decision trees. The configuration used is similar to Nunes et al. (2017) implementation of RF.

## 5 Experiments and evaluation

To evaluate the performance of the proposed HAL system, data collected from our experimental setup are used. This is used in order to verify the proposed system via a limited test we performed before it is tested on public datasets. Afterwards, the system is also evaluated using publicly available benchmark human activity dataset, Cornell Activity Dataset (CAD-60) (Sung et al. 2011). The following sections describe the experiments conducted in this work and discussion of the results obtained.

### 5.1 Experimental setup

Skeletal data are collected from three actors using a Microsoft Kinect V2 RGB-D sensor as mentioned previously in Sect. 3.1. The data are obtained at a frame rate of 30 frames per second (fps). Four activities are carried out, namely brushing teeth, pick up object (from the ground), sit on sofa and stand up. Each actor performs a single activity for a duration of 45–90 s. Sitting on sofa activity is performed by an actor going through a sequence of sitting, and getting up poses with more time spent in sitting, and standing activity is performed in a similar way with more time spent staying standing. The summary of the data collected is presented in Table 1.

The data acquired are pre-processed following the process earlier mentioned in Sect. 3.1. Key features representing activities are extracted from the processed data. Table 2 shows the number of activity features computed from the RGB-D sensor skeleton with 15 joints. The number of joints used in computing spatial displacement features is selected based on the importance of the joints while carrying out the selected activities. Nine features are computed which represent the Euclidean distance between both left and right hands, each hand and head, each hand and its corresponding foot, each shoulder and corresponding foot, each hip and corresponding foot. The other features are obtained for each joint coordinate—given that 15 joints are used, each feature description comprises $15 \times 3 = 45$ features extracted.

Features selected from the experimental dataset are fed into the learning model to test the performance of the system. A $K$-fold cross-validation test strategy is applied with $K = 4$. This involves splitting the data into fourfold in which threefold is used as training data for the model and the remaining fold is left out for validation. This process is repeated using each fold for validation, and the final result is the average performance of all test validation folds.

**Table 1** Summary of experimental human activity data collected from 3 actors using Microsoft kinect V2 RGB-D sensor

| Activity | Number of frames | | |
|---|---|---|---|
| | Actor 1 | Actor 2 | Actor 3 |
| Brushing teeth | 2202 | 1876 | 1781 |
| Pick up object | 1804 | 1663 | 1355 |
| Sit on sofa | 1489 | 1672 | 2736 |
| Stand up | 2126 | 2059 | 2100 |
| Total | 7621 | 7270 | 7972 |

Activities performed comprise: brushing teeth, pick up object, sit on sofa, stand up

**Table 2** Activity features computed from raw RGB-D sensor information of skeleton with 15 joints used in this work

| Feature description | Feature label |
|---|---|
| Spatial displacement $\delta$ between both hands, hands and head, hands and feet, shoulders and feet, hip and feet | 1–9 |
| Temporal joint coordinate displacement $t_{cp}$ | 10–54 |
| Temporal joint coordinate displacement $t_{ci}$ | 55–99 |
| Joint coordinate-mean difference $j_{(i,\text{mean})}$ | 100–144 |
| Joint coordinate-variance difference $j_{(i,\text{var})}$ | 145–189 |
| Joint coordinate-standard deviation difference $j_{(i,\text{std})}$ | 190–234 |
| Joint coordinate-skewness difference $j_{(i,\text{skw})}$ | 235–279 |
| Joint coordinate-kurtosis difference $j_{(i,\text{kur})}$ | 278–324 |
| Total number of computed features | 324 |

### 5.2 CAD-60 dataset and experiment

The CAD-60 dataset comprises RGB-D sequence of human activities acquired using an RGB-D sensor at a frame rate of 15 fps. The dataset contains RGB image, depth image and skeleton joint coordinates information of 15 skeletal joints of activities carried out. However, the proposed HAL system utilises only the skeleton joint coordinates information. Four different actors perform 12 activities in five different locations, namely bathroom, bedroom, kitchen, living room and office. The activities performed are: rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard, working on computer and a random + still activity. The random + still contains random movements sequence and a still pose performed by each actor. The stages described in the proposed HAL system are applied, with the CAD-60

**Table 3** Performance of the proposed HAL system on experimental dataset comprising four activities: brushing teeth, pick up object, sit on sofa, stand up

| Activity | Performance result | |
|---|---|---|
| | Precision (%) | Recall (%) |
| Brushing teeth | 40.38 | 62.19 |
| Pick up object | 100 | 94.69 |
| Sit on sofa | 100 | 100 |
| Stand up | 54.10 | 35.13 |
| Average | 70.65 | 68.43 |

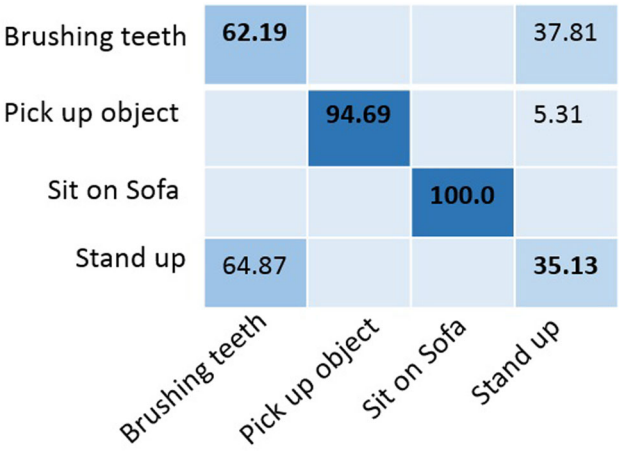



**Fig. 7** Confusion matrix of the proposed HAL system on experimental data

dataset as raw input to the system. The same number of features as shown in Table 2 is computed from the dataset.

Learning the activities is done as a grouping of activities in the various locations. This grouping shown in Table 4 follows the format used by all approaches reported in the state of the art in Table 6. For testing the trained model, a method of *leave-one-out* cross-validation is carried out in which the model is trained on three actors and tested on the *unseen* actor. This is also called a *new person* test strategy.

## 5.3 Evaluation and discussion

The proposed HAL system is evaluated on both datasets mentioned in Sects. 5.1 and 5.2 following the test methods described. The CAD-60 dataset tests are performed following similar test methods described by Sung et al. (2011) and other approaches by the authors in Table 6. Test results and discussions are presented in the following sections.

### 5.3.1 Experimental dataset results and evaluation

Table 3 shows the results obtained from the performance of the proposed HAL system on the experimental dataset. These are presented in terms of *precision* and *recall*. The system achieves an overall average precision of 70.65% and recall of 68.43% with the dataset. In Fig. 7, the confusion matrix shows the percentage of correctly classified activities along with the percentage of false classified activities. It can be noticed that the performance in activities of pick up object with recall of 94.69% and sit on sofa with recall of 100% are quite impressive. However, the model did not perform as impressively in correctly classifying brushing teeth and stand-up activities. This is due to the fact that both activities have closely related poses as brushing teeth is performed while in a stand-up pose. This gives rise to more stand-up data—i.e. 64.87%—characterised as brushing teeth which affects the overall performance achieved. In order to adequately test the robustness of a supervised learning system, the availability of more data samples is required for proper training and validation of learning models. However, the experimental dataset collected contains fewer data samples when compared with other human activity datasets such as the CAD-60 dataset. This can also be a reason for the performance achieved on the experimental dataset. Therefore, we also tested the HAL system with the CAD-60 dataset which contains more samples of human activity.

### 5.3.2 CAD-60 dataset results and evaluation

The results obtained from the performance of the proposed HAL system on the dataset are shown in Table 4. This is presented in terms of *precision* and *recall* of the HAL system. The proposed system achieved an overall average performance of 92.32% precision and 89.66% recall with features selected using the Relief-F feature selection method described in Sect. 3.4 and a performance 90.96% precision and 88.52% recall when all the features extracted are used. In Table 5, the result from different locations is shown. When compared with Table 4, the system achieved a better performance with selected features than with all the features as reported in Table 5. Table 6 shows the proposed system performance compared to the state of the art performances on the same dataset (Cornell University 2009). The table also shows information of the state-of-the-art works which employ extended modality of RGB-D sensor information which is a combination of skeletal joint coordinates information with either of RGB image and depth image sensor information modes. The proposed HAL system's performance indicates the features extracted in our system sufficiently discriminate the selected human activities from skeletal joints information.

Comparison of the proposed HAL system's performance with the state of the art on the CAD-60 dataset presented in Fig. 8 shows the proposed system is able to attain an

**Table 4** Performance of the proposed HAL system **with selected features** on the CAD-60 dataset using a "new person" test in different locations: bathroom, bedroom, kitchen, living room and office

| Location | Activity | Proposed HAL system | |
|---|---|---|---|
| | | Prec. (%) | Rec. (%) |
| Bathroom | Rinsing mouth | 100 | 99.97 |
| | Brushing teeth | 96.97 | 75.16 |
| | Wearing contact lens | 54.48 | 92.68 |
| | Random + still | 99.98 | 100 |
| | Average | 95.72 | 93.41 |
| Bedroom | Talking on phone | 98.58 | 74.55 |
| | Drinking water | 91.47 | 60.99 |
| | Opening pill container | 15.39 | 66.55 |
| | Random + still | 100 | 100 |
| | Average | 94.37 | 84.01 |
| Kitchen | Drinking water | 92.96 | 74.81 |
| | Cooking (chopping) | 31.04 | 66.67 |
| | Cooking (stirring) | 78.43 | 77.52 |
| | Opening pill container | 74.49 | 75.49 |
| | Random + still | 100 | 100 |
| | Average | 86.85 | 84.76 |
| Living room | Talking on phone | 82.36 | 88.29 |
| | Drinking water | 86.93 | 74.14 |
| | Talking on couch | 94.27 | 100 |
| | Relaxing on couch | 100 | 100 |
| | Random + still | 100 | 100 |
| | Average | 94.37 | 94.41 |
| Office | Talking on phone | 67.06 | 93.42 |
| | Writing on board | 87.36 | 73.19 |
| | Drinking water | 100 | 83.84 |
| | Working on computer | 100 | 100 |
| | Random + still | 100 | 100 |
| | Average | 93.28 | 91.71 |
| Overall average | | 92.32 | 89.66 |

**Table 5** Performance of the proposed HAL system **with all features** extracted from the CAD-60 dataset using a "new person" test. This shows the average performance from different locations

| Location | Performance result | |
|---|---|---|
| | Precision (%) | Recall (%) |
| Bathroom | 91.36 | 90.37 |
| Bedroom | 86.72 | 83.43 |
| Kitchen | 86.38 | 83.54 |
| Living room | 95.95 | 94.36 |
| Office | 94.41 | 90.92 |
| Overall average | 90.96 | 88.52 |

**Table 6** Overall average precision and recall of the proposed HAL system with the state of the art on the CAD-60 dataset in a "new person" setting in order of increasing precision reported by Cornell University (2009)

| Method | Prec. (%) | Rec. (%) | Extended modality |
|---|---|---|---|
| Sung et al. (2011, 2012) | 67.9 | 55.5 | ✓ |
| Piyathilaka and Kodagoda (2013) | 70.0 | 78.0 | – |
| Yang and Tian (2014) | 71.9 | 66.6 | ✓ |
| Ni et al. (2013) | 75.9 | 69.5 | ✓ |
| Gaglio et al. (2015) | 77.3 | 76.7 | – |
| Gupta et al. (2013) | 78.1 | 75.4 | ✓ |
| Koppula et al. (2013) | 80.8 | 71.4 | ✓ |
| Nunes et al. (2017) | 81.83 | 80.02 | – |
| Zhang and Tian (2012) | 86.0 | 84.0 | ✓ |
| Proposed HAL system (with all features) | 90.96 | 88.52 | – |
| Faria et al. (2014) | 91.1 | 91.9 | – |
| Parisi et al. (2015) | 91.9 | 90.2 | – |
| Proposed HAL system (with selected features) | 92.32 | 89.66 | – |
| Zhu et al. (2014) | 93.2 | 84.6 | ✓ |
| Shan and Akella (2014) | 93.8 | 94.5 | – |
| Cippitelli et al. (2016) | 93.9 | 93.5 | – |

The extended modality column indicates the mode of RGB-D sensor information used by different works, i.e. skeletal joint coordinates only (–) or skeletal joint coordinates information with a combination of either RGB image and depth image information modes (✓)

impressive performance. While some other proposed systems performance outperforms the HAL systems performance, the proposed HAL system differs from the other better performances in the following ways. The system proposed by Zhu et al. (2014) reported a performance of 93.2% precision and 84.6% recall. Although their precision exceeds that of the proposed HAL system, our system performs better in terms of recall. Also, the system by Zhu et al. (2014) uses a fusion of spatiotemporal interest point features obtained from combination of RGB-D sensor information modalities, i.e. depth image, RGB image and skeleton information as indicated in Table 6. This process can increase computational cost. The proposed HAL system utilises only the skeleton information offered by the RGB-D sensor to achieve such high performance. This shows that by adding more information for computer vision processing our system has the potential to achieve a higher performance.

However, the performance attained by Shan and Akella (2014) slightly out performs our proposed HAL system which is observed from the comparison of state-of-the-art results in Table 6. This approach performed tests excluding

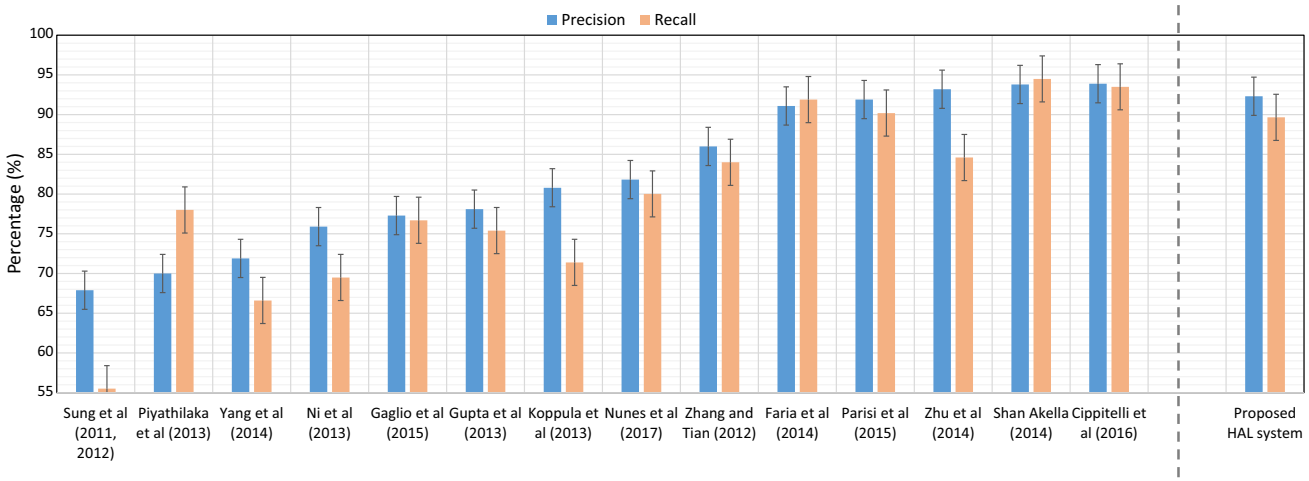



**Fig. 8** Precision and recall performance comparison of proposed HAL system with the state-of-the-art results on the CAD-60 dataset

**Table 7** Proposed classifier ensemble method performance comparison with single classifier performance on CAD-60 dataset

| Proposed by | Method | Prec. (%) | Rec. (%) |
|---|---|---|---|
| Yang and Tian (2014) | Naive Bayes Nearest Neighbour | 71.9 | 66.6 |
| Ni et al. (2013) | Latent SVM | 75.9 | 69.5 |
| Gaglio et al. (2015) | SVM | 77.3 | 76.7 |
| Koppula et al. (2013) | Structural SVM | 80.8 | 71.4 |
| Nunes et al. (2017) | RF | 81.83 | 80.02 |
| Zhang and Tian (2012) | SVM | 86.0 | 84.0 |
| Parisi et al. (2015) | Neural network | 91.9 | 90.2 |
| Proposed HAL system | Classifier ensemble | 92.32 | 89.66 |

the random + still activity performed by all actors in the dataset which is included in the tests performed using the proposed HAL system. This information is relevant in generalising the robustness of the system across varying human activities.

The system proposed by Cippitelli et al. (2016) on the CAD-60 dataset attained a higher performance of both precision and recall of 93.9 and 93.5%, respectively. Their system is tested with the dataset in a similar way observed in the system by Shan and Akella (2014) which excludes test on the random + still activity. Another reason could also be due to the fact that the proposed HAL uses all 15 skeleton joints of the CAD-60 dataset, whereas Cippitelli et al. (2016) used 11 selected skeleton joints to achieve the high performance. The selected joints do not include relevant joints such as the shoulders which are needed for our proposed application in assistive robots effectively executing human activities via transfer learning. However, the proposed HAL system with 15 skeletal joints achieves higher performance when compared with Cippitelli et al. (2016)'s performance with 15 skeletal joints of 87.9% precision and 86.7% recall using all 15 skeleton joints.

With the performance achieved using the proposed HAL system with both experimental and publicly tested CAD-60 datasets, this shows the systems potential in applications of assistive robots learning of human activities.

### 5.3.3 Comparison of classifier ensemble with single classifier performance

The method of using a classifier ensemble as proposed in this work shows the increase in activity learning accuracy when compared with other proposed methods which use single classifiers. Table 7 shows the performance of the proposed classifier ensemble method with other methods which apply single classifiers in learning human activities. Also, it can be noticed that majority of the other approaches apply SVM in recognising human activities which is also used in the proposed classifier ensemble method and results show the classifier ensemble outperforms the other single classifier methods. In addition, the proposed classifier ensemble approach proposed also has the benefit of attaining high activity learning performance with a small amount of training samples when compared to other widely used methods such as deep learning neural networks (Ijjina and Chalavadi

2017)—which require a lot of data and more time in training such networks for concise predictions.

## 6 Conclusion and future work

The work presented here proposes a system for human activity learning with the use of skeletal data obtained using an RGB-D sensor. We have shown explicitly the process of refining the raw sensor data obtained, computing relevant features and training the learning model. The main objective of this work is to have an activity learning system which is able to distinctly recognise activities as they are performed. The system can then be incorporated in an assistive robot to aid learning to perform such human activities. The performance attained by the proposed system on the CAD-60 benchmark dataset shows its reliability if used with an assistive robot.

Although we used a selection of three base classifiers in building the ensemble model, this could be extended to include more classifiers which may improve performance and also deep learning neural networks which are increasingly used in human activity recognition systems. The system could also be extended to learning activities on-the-fly as they are carried out by an actor. We plan to implement this in future. The direction of research following this work is to segment different aspects of each learned activity into representations that any assistive robot platform can adopt in reliably executing human activity.

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with animals performed by any of the authors.

**Informed consent** Informed consent was obtained from all individual participants included in the study.



## References

Adama DA, Lotfi A, Langensiepen C, Lee K (2018) Human activities transfer learning for assistive robotics. Springer, Cardiff, pp 253–264

Blackman S, Matlo C, Bobrovitskiy C, Waldoch A, Fang ML, Jackson P, Mihailidis A, Nygård L, Astell A, Sixsmith A (2016) Ambient assisted living technologies for aging well: a scoping review. J Intell Syst 25(1):55–69

Capela NA, Lemaire ED, Baddour N (2015) Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients. PLoS ONE 10(4):1–18

Chao F, Huang Y, Zhang X, Shang C, Yang L, Zhou C, Hu H, Lin CM (2017) A robot calligraphy system: from simple to complex writing by human gestures. Eng Appl Artif Intell 59:1–14

Cippitelli E, Gasparrini S, Gambi E, Spinsante S (2016) A human activity recognition system using skeleton data from RGBD sensors. Comput Intell Neurosci. https://doi.org/10.1155/2016/4351435

Cornell University (2009) Cornell Activity Dataset: state of the art results. http://pr.cs.cornell.edu/humanactivities/results.php. Accessed 15 Feb 2018

Diao R, Chao F, Peng T, Snooke N, Shen Q (2014) Feature selection inspired classifier ensemble reduction. IEEE Trans Cybern 44(8):1259–1268

Faria DR, Premebida C, Nunes U (2014) A probabilistic approach for human everyday activities recognition using body motion from RGB-D images. In: The 23rd IEEE international symposium on robot and human interactive communication. RO-MAN, IEEE, pp 732–737

Gaglio S, Re GL, Morana M (2015) Human activity recognition process using 3-D posture data. IEEE Trans Hum Mach Syst 45(5):586–597

Gu Y, Do H, Ou Y, Sheng W (2012) Human gesture recognition through a kinect sensor. In: IEEE international conference on robotics and biomimetics (ROBIO). IEEE, pp 1379–1384

Gupta P, Dallas T (2014) Feature selection and activity recognition system using a single triaxial accelerometer. IEEE Trans Biomed Eng 61(6):1780–1786. https://doi.org/10.1109/TBME.2014.2307069

Gupta R, Chia AYS, Rajan D (2013) Human activities recognition using depth images. In: Proceedings of the 21st ACM international conference on multimedia, pp 283–292

Han F, Reily B, Hoff W, Zhang H (2017) Space-time representation of people based on 3D skeletal data: a review. Comput Vis Image Underst 158(Supplement C):85–105

Helwa MK, Schoellig AP (2017) Multi-robot transfer learning: a dynamical system perspective. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 4702–4708

Hussein ME, Torki M, Gowayyed MA, El-Saban M (2013) Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In: Proceedings of the twenty-third international joint conference on artificial intelligence. AAAI Press, Beijing, China, pp 2466–2472

Iglesias JA, Angelov P, Ledezma A, Sanchis A (2010) Human activity recognition based on evolving fuzzy systems. Int J Neural Syst 20(5):355–364

Ijjina EP, Chalavadi KM (2017) Human action recognition in RGB-D videos using motion sequence information and deep learning. Pattern Recogn 72:504–516

Jalal A, Kamal S (2014) Real-time life logging via a depth silhouette-based human activity recognition system for smart home services. In: 11th IEEE international conference on advanced video and signal-based surveillance, AVSS 2014. IEEE, pp 74–80

Jayawardena C, Kuo IH, Broadbent E, MacDonald BA (2016) Socially assistive robot healthbot: design, implementation, and field trials. IEEE Syst J 10(3):1056–1067

Kononenko I (1994) Estimating attributes: analysis and extensions of relief. In: Bergadano F, De Raedt L (eds) Machine learning: ECML-94. Springer, Berlin, pp 171–182

Koppula HS, Gupta R, Saxena A (2013) Learning human activities and object affordances from RGB-D videos. Int J Robot Res 32(8):951–970

Li SZ, Yu B, Wu W, Su SZ, Ji RR (2015) Feature learning based on SAE–PCA network for human gesture recognition in RGBD images. Neurocomputing 151(Part 2):565–573

Microsoft (2017) Developing with kinect for windows. https://developer.microsoft.com/en-us/windows/kinect/develop. Accessed 28 Feb 2017

Ni B, Pei Y, Moulin P, Yan S (2013) Multilevel depth and image fusion for human activity detection. IEEE Trans Cybern 43(5):1383–1394

Nunes UM, Faria DR, Peixoto P (2017) A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier. Pattern Recogn Lett 99:21–31

Parisi G, Weber C, Wermter S (2015) Self-organizing neural integration of pose-motion features for human action recognition. Front Neurobot 9:3

Piyathilaka L, Kodagoda S (2013) Gaussian mixture based hmm for human daily activity recognition using 3D skeleton features. In: 2013 IEEE 8th conference on industrial electronics and applications (ICIEA), pp 567–572

Shan J, Akella S (2014) 3D human action segmentation and recognition using pose kinetic energy. In: 2014 IEEE international workshop on advanced robotics and its social impacts, pp 69–75

Sung J, Ponce C, Selman B, Saxena A (2011) Human activity detection from RGBD images. In: Proceedings of the 16th AAAI conference on plan, activity, and intent recognition, AAAIWS'11-16. AAAI Press, pp 47–55

Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from RGBD images. In: 2012 IEEE international conference on robotics and automation. IEEE, pp 842–849

Tahir MA, Kittler J, Bouridane A (2012) Multilabel classification using heterogeneous ensemble of multi-label classifiers. Pattern Recogn Lett 33(5):513–523

Wei P, Zheng N, Zhao Y, Zhu SC (2013) Concurrent action detection with structural prediction. In: Proceedings of the IEEE international conference on computer vision. IEEE, pp 3136–3143

Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data 3(1):9

Xiao Y, Zhang Z, Beck A, Yuan J, Thalmann D (2014) Human–robot interaction by understanding upper body gestures. Presence 23(2):133–154

Yang X, Tian Y (2014) Effective 3D action recognition using eigenjoints. J Vis Commun Image Represent 25(1):2–11

Yao G, Zeng H, Chao F, Su C, Lin CM, Zhou C (2016) Integration of classifier diversity measures for feature selection-based classifier ensemble reduction. Soft Comput 20(8):2995–3005

Zhang C, Tian Y (2012) RGB-D camera-based daily living activity recognition. J Comput Vis Image Process 2(4):12

Zhou D, Shi M, Chao F, Lin CM, Yang L, Shang C, Zhou C (2018) Use of human gestures for controlling a mobile robot via adaptive cmac network and fuzzy logic controller. Neurocomputing 282:218–231

Zhu Y, Chen W, Guo G (2014) Evaluating spatiotemporal interest point features for depth-based action recognition. Image Vis Comput 32(8):453–464