Learning Human Activities for Assisted Living Robotics

David Ada Adama*
School of Science and Technology
Nottingham Trent University
Nottingham, UK NG11 8NS

Ahmad Lotfi†
School of Science and Technology
Nottingham Trent University
Nottingham, UK NG11 8NS

Caroline Langensiepen
School of Science and Technology
Nottingham Trent University
Nottingham, UK NG11 8NS

Kevin Lee
School of Science and Technology
Nottingham Trent University
Nottingham, UK NG11 8NS

Pedro Trindade
School of Science and Technology
Nottingham Trent University
Nottingham, UK NG11 8NS

ABSTRACT

Assistive living has gained increased focus in recent years with the increase in elderly population. This has led to a desire for technical solutions to reduce cost. Learning to perform human activities of daily living through the use of assistive technology (especially assistive robots) becomes more important in areas like elderly care. This paper proposes an approach to learning to perform human activities using a method of activity recognition from information obtained from an RGB-D sensor. Key features obtained from clustering and classification of relevant aspects of an activity will be used for learning. Existing approaches to activity recognition still have limitations preventing them from going mainstream. This is part of a project directed towards transfer learning of human activities to enhance human-robot interaction. For test and validation of our method, the CAD-60 human activity data set is used.

CCS CONCEPTS

•**Computing methodologies** → **Activity recognition and understanding;** *Supervised learning by classification; Feature selection; Vision for robotics; Image representations;*

KEYWORDS

Activity recognition, Assistive robotics, Ambient assisted living, Feature extraction

1 INTRODUCTION

With the increase in the ageing population in developed societies, Ambient Assisted Living (AAL) technology helps to reduce cost of elderly care. One means of achieving this is by deploying assistive robotics in an AAL environment. This requires the robot to learn tasks that human carers will perform as part of routine duties.

Human Activity Recognition (HAR) has gained a lot of interest in recent years [9][10][11]. In applications involving human-computer interactions (for example, gaming and assisted living environments), HAR enables elaborate explanation/interpretation of activities performed by humans. This is key to understanding how these activities are learned and how one experience relates to another. This improves the collaboration and adaptation of assistive robots.

Prior to assistive robots performing a human activity, they need to learn these activities to effectively perform them. This requires grouping human movements with their descriptive semantics. Observing activities as they are performed through the use of visual or non-visual sensors makes it a lot easier to obtain information of human activities in an environment[4][19][20]. It would be extremely hard to understand and interpret activities using a normal visual sensor such as RGB cameras which provide 2D visual data [6]. These sensors provide limited information for an activity performed in a real world environment. However, recent development in RGB-D sensors show that they are better devices for observing human activities. These sensors provide a means of better observing the world to detect human pose used to build activity recognition systems [4][20]. They also provide a platform for exploiting depth maps, body shape and skeleton joint detection of humans in 3D space which are used in developing sophisticated recognition algorithms.

This paper proposes a supervised learning method to human activities recognition by exploiting 3D human skeleton data provided by an RGB-D sensor. The proposed method shows the observation of activities as they are performed and how features are extracted after a clustering method is applied. Thus, providing a means of classifying different activities. The features extracted simplify the process of learning the activities. The work presented here is part of a research project directed towards improving human-robot or robot-robot interaction through Transfer Learning [12]. This will enable more adaptable robots learn to perform human activities of daily living from transferred knowledge. For example, in Figure 1, two robots are shown performing an activity in which one acts as the *teacher* while the other acts as a *learner* using transferred knowledge gained from visual observation as the *teacher* robot performs the activity. The goal of our research is to develop a state

*David A. Adama is a PhD research student who has conducted the research as part of his thesis.

†Ahmad Lotfi is the corresponding author. The corresponding email address is: ahmad.lotfi@ntu.ac.uk

PETRA'17, Rhodes Island, Greece

DOI: 10.1145/nnnnnnn.nnnnnnn

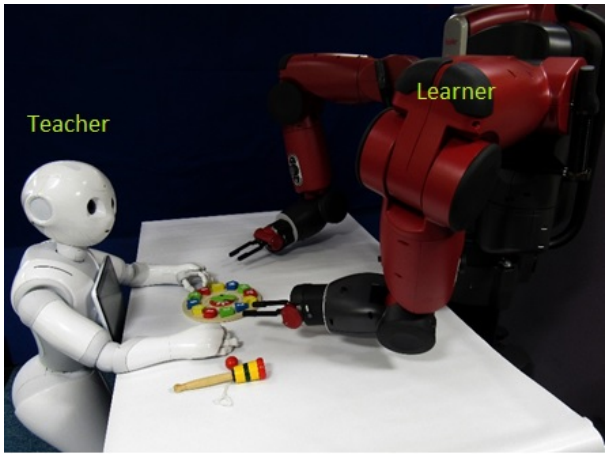

Figure 1: An example of robot-robot interaction through learning from activity recognition.

of the art system which will have HAR and learning performance comparable to that of humans.

This paper is organized as follows: In section 2, a review of related work in this area is presented. Section 3 gives details of the method applied to our approach and some initial results are presented in Section 4. Section 5 presents conclusions and future work to be undertaken.

2 RELATED WORK

For an assistive robot to perform an Activity of Daily Living (ADL), an understanding of the activity is required which is the process of learning. An important aspect in the process of learning is the recognition of activities as they are performed by human actors. This creates the need for proper observation of the environment by the robot to rightly interpret the activity.

HAR from information obtained from RGB-D sensors gives very important information relevant for a robot to understand an activity. By exploring human pose detection using RGB-D sensors, activity recognition has seen more advancement in recent times [4][19]. Using RGB-D sensors extracts 3D skeleton data from depth images and body silhouette for feature generation. In [4], the RGB-D sensor is used to generate 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.* Authors in [10] use similar RGB-D sensor to obtain depth silhouette of human activities from which body points information are extracted for the activity recognition system. Another approach is shown in the work in [5] 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 preprocessed. This process is carried out to reduce redundancy in data for better representation of features of an activity.

HAR from video camera sensors can be broken down into two aspects [18]; feature-based and model-based. Feature-based techniques such as Histogram of Oriented Gradients (HOG)[21], subspace clustering based approach (SCAR)[25] are used to extract features for recognizing human activity from data acquired using sensors. Model-based approach has to do with construction of a human model for recognition either as a 2D, 3D or skeletal model. Authors in [1][23] construct models using kinematic approach that extract features from frame sequences for human structure representations. A combination of both feature-based and model-based approaches is seen in [20] where a maximum entropy Markov Model (MEMM) for classification of activities using features from skeleton tracking combined HOG. Authors in [3] also used neural network technique to propose an end-to-end hierarchical recurrent neural network (RNN) for representing skeleton based construction. They make use of the raw positions of human joints as input to the RNN.

All methods reported above are using feature extraction techniques to obtain feature vectors which describe the activity performed. HAR approaches which utilize human 3D joint positions information extracted from RGB-D sensors transform joint positions of individual frames into column vectors. A matrix is then formed to encode the sequence of frames at specific time intervals. 14 features obtained from distances between different body parts are used to characterize 12 activities[4]. This was used with a combination of multiple classifiers to form a Dynamic Bayesian Mixture Model (DBMM). Similarly, [8] applied statistical covariance of 3D joints (Cov3DJ) as features to encode the skeleton data of raw joint positions. Another approach seen in [24] used a sequence of joint trajectories and applied wavelets to encode each temporal sequence of joints into features. Since not all joints of a person provide substantial information for interpreting an activity performed, different methods are proposed to select key joints which are more descriptive [2][7][14][16] and they will be investigated for an optimal method to be used in this research.

In the next section, the proposed methodology for feature extraction and activity recognition from raw joint positions of skeleton data obtained from an RGB-D sensor will be explained.

3 ACTIVITY RECOGNITION SYSTEM

To recognize activities from RGB-D sensors, the first stage is to recognize skeleton raw joint positions. This is done by obtaining recorded frames of annotated data for activities performed by a human actor. The data captured includes a hierarchical nature of activities, human skeletal features and their transitions over time. The system can be set up in a living environment. An example of RGB and depth image frame extracted from an RGB-D sensor is shown in Figure 2. Once the raw joints positions information is obtained, important features from this data must be extracted before they are classified into different activities. The proposed system architecture is shown in Figure 3. The following key steps are identified;

- *Posture feature extraction*: Posture feature vectors representing key human poses that describe an activity are extracted by obtaining centroids of clusters using *K-means* clustering technique.
- *Activity classification*: Classification of activities from the extracted posture features.

Details of each step is briefly explained in the following sections.

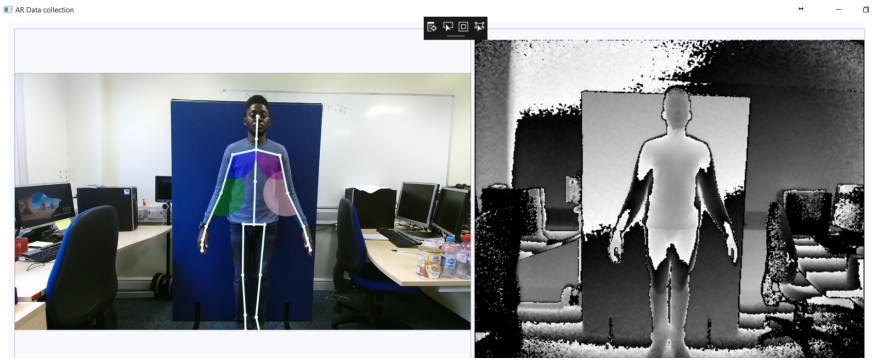

Figure 2: Frame of RGB and Depth image showing tracked human skeleton.

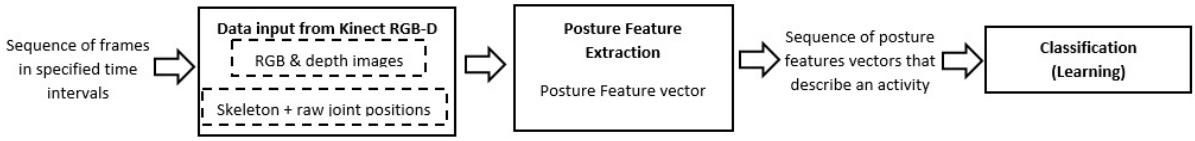

Figure 3: Overall system architecture of proposed system.

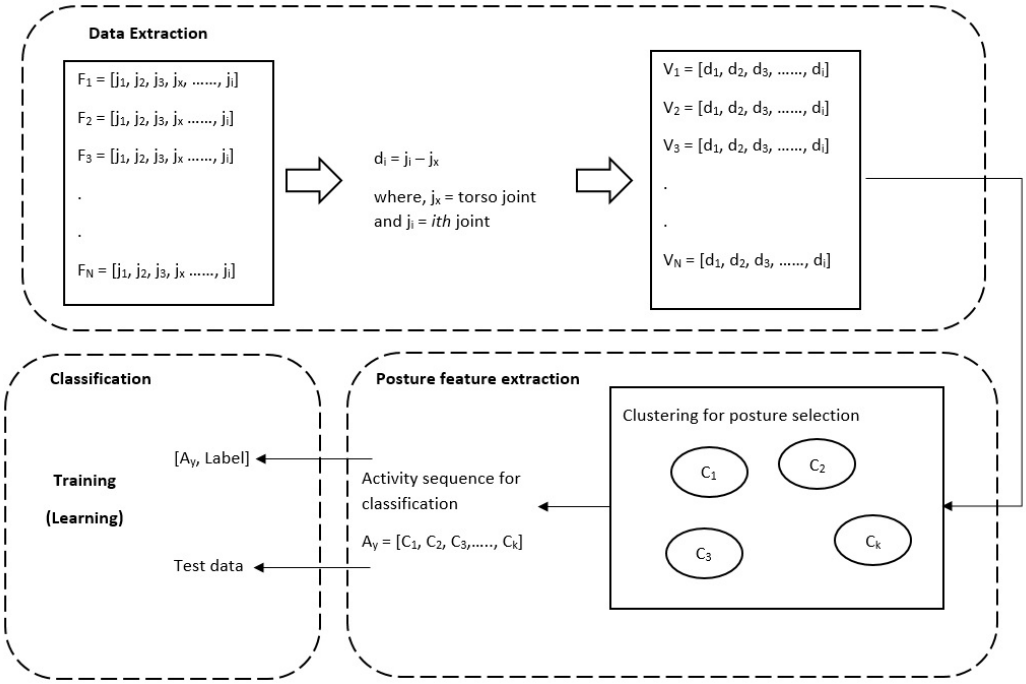

Figure 4: Steps to activity recognition in the proposed method.

3.1 Posture Feature Extraction

Posture feature extraction is a key step in our system for recognizing activities. Features obtained in activity recognition systems can be

computed using joint position coordinates extracted from RGB-D sensor skeleton data. This approach is one of the simplest employed by researchers. Two methods can be applied, these are based on raw joint positions and displacement-based representations- when considering temporal and spatial data.

The proposed system adopts a method which exploits displacement features from the skeleton joint coordinates. However, temporal information will be excluded to make the system independent of speed of movement. This method was introduced by [2]. For each skeleton frame, a vector F_N is used to represent the posture for joints j_i in 3D where N is the frame (or observation) number for an activity as shown in Figure 4. The Kinect RGB-D sensor considers the skeleton frame of reference to be the sensor. However, in order to compensate the position of the skeleton, the frame of reference will be considered for all joints relative to the torso joint coordinates.

Consider a skeleton which has i joints with j_x representing the torso coordinates. The distance vector d_i between the i^{th} joint and the torso (reference) is given by:

$$d_i = j_i - j_x \tag{1}$$

After computing the distance d_i for all the joints in a frame of an activity, a vector V representing the set of joints position distance relative to the torso is obtained. Each vector V represents a posture for a skeleton frame. A set of V_N vectors for all frames obtained represent an activity. This process is illustrated in Figure 4.

$$V_N = (d_1, d_2, d_3, ..., d_i) \tag{2}$$

The proposed approach applies *K-means* clustering technique for selection of key human pose that represent an activity thereby reducing the complexity in large activity data sets. This technique is used for representation of an activity by a subset of poses, without having to use all observations. V_N vectors representing frames of an activity are grouped in K clusters based on the squared Euclidean distance as a metric.

For an activity consisting of V_N vectors, K clusters are defined to represent key postures for the activity. A cluster C_K is a grouping of closely related joints positions component (joint coordinates) within frames of the activity. After computing, an output of K clusters $[C_1, C_2, C_3, ..., C_K]$ are obtained. Each cluster is a vector containing centroids for all joints that define a key posture within an activity. Once the clusters are formed, they are sorted out following the order the cluster elements occur during the activity. The clustering algorithm provides the ID of the cluster which an observation within the activity belongs. The sequence of the cluster vectors obtained represent the sequence of observations that constitute the performed activity. The centroids of each cluster represent key postures which are the most important features of an activity. These features form the activity features vector A.

$$A = [C_1, C_2, C_3, ..., C_K] \tag{3}$$

For example, considering an activity represented with 10 observations and 3 clusters, one of the outputs after applying $K - means$ clustering would be a sequence of cluster IDs representing the activity: $[2, 2, 1, 1, 1, 1, 3, 3, 3, 1]$, this means the first two observations are within cluster 2, the third - sixth observation in cluster 1, seventh - ninth observation in cluster 3 and the tenth observation in cluster 1. Therefore, the activity features vector will be $A = [C_2, C_1, C_3, C_1]$.

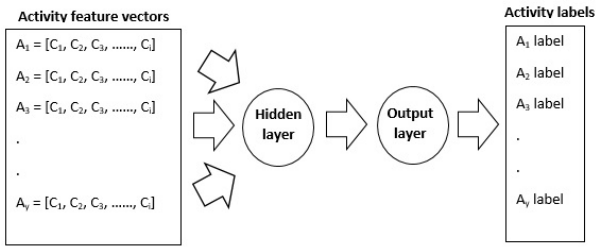

Figure 5: Artificial Neural Network classification structure.

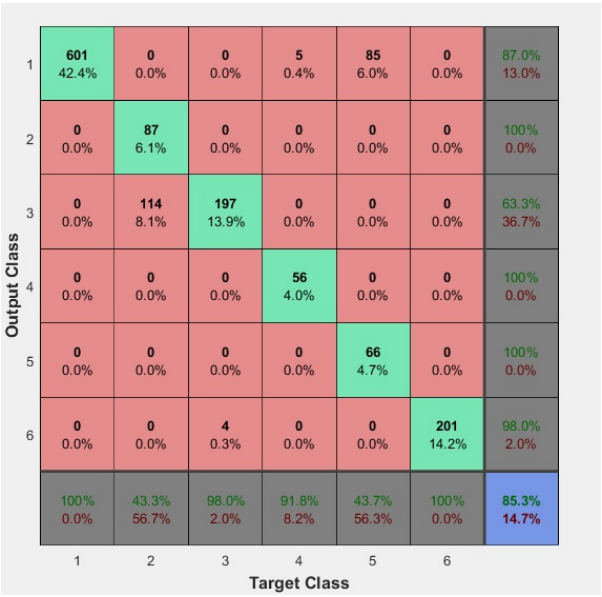

Figure 7: Confusion matrix showing the performance of our system for 6 activities.

One of the challenge of using *K-means* clustering algorithm in the proposed method is obtaining the optimal number of clusters that can effectively represent the posture features of an activity. Considering that, we use different number of clusters C_K for $K = (5, 10, 15, 20, 25, 30)$ and compare the results. The result is evaluated by minimizing the error Z which is the sum of the deviations between the centroids of the clusters and the actual joint positions through the sequence of frames for an activity.

$$Z = \sum_{i=1}^{n} (C_{k,n} - J_{i,n}) \tag{4}$$

where i is the i^{th} joint, $C_{k,n}$ is the centroid of the cluster k at the n^{th} observation/frame and $J_{i,n}$ is the ith joint at n.

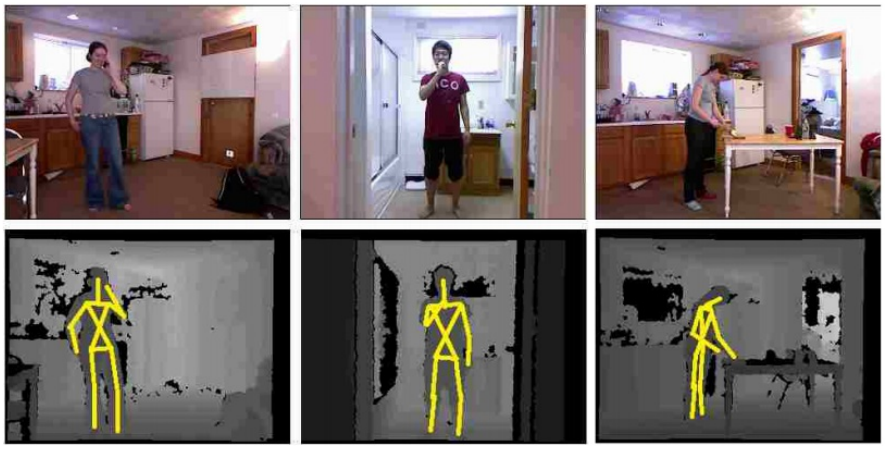

Figure 6: Selected frames of human actor performing activities in a living environment from CAD-60 dataset.

3.2 Activity Classification

Classification of the activity feature vector is computed to associate each extracted feature vector with the correct activity performed. To achieve the task of human activity classification, machine learning classification algorithms may be applied. One of such algorithm is an *Artificial Neural Network* (ANN) trained classifier.

The proposed system employs an ANN algorithm consisting of a hidden layer for classification of human activities. Details of the ANN algorithm for classification which is adopted in this work can be found in [15]. In training the classifier, feature vectors obtained from clustering which describe a performed activity are passed as inputs and the activity labels as targeted outputs of the ANN. An iterative learning process which is a key feature of ANN's is performed during which the weights of all neurons are adjusted to predict the correct activity label from input activity feature vectors. The structure of the ANN classification is given in Figure 5.

The input to the hidden layer neurons is the sum of the extracted activity feature vectors A and weights w_i passed through an activation function. For the initial iteration, weights are selected at random and subsequently, they are modified through successive iterations during training of the network based on the error propagated from each iteration. The error is the difference between the actual output label for the activity and the predicted value from each iteration and is calculated for each iteration.

4 EXPERIMENTAL RESULTS AND DISCUSSION

The proposed method of learning from activity recognition is tested on a well known human activity data set, CAD-60 data set [22]. The data had been extracted using a Microsoft Kinect RGB-D sensor [13]. It contains RGB frames as well as depth for sequence of human activities performed.

The data is collected at a frame rate of 15fps which is offered by the sensor and comprises of 12 activities performed by four people. However, for testing the proposed system, 6 activities out of 12 activities are selected which are;

(1) Brushing teeth
(2) Cooking-chopping
(3) Cooking-stirring
(4) Open pill container
(5) Talk on phone
(6) Write on board

The activities are selected based on the type of application of the proposed system for learning using an assistive robot capable of performing these activities. Therefore, only activities which require more participation of the arms when performed are chosen for tests. Selected frames for these activities are shown in Figure 6.

For posture feature extraction, the joint positions are extracted following the process as described in Section 3.1. However, in computing the optimal number of clusters which best represent activity feature vectors different number of clusters ranging from 5 to 30 are tested. The best results are obtained using 25 and 30 clusters. To illustrate this, a plot of cluster centroids and joint positions through a sequence of *brushing teeth* activity for one joint component (Right hand x-coordinate) is presented in Figure 8. The joint position is particularly selected because it is one of the most significant components in performing the activity.

After clustering, the activity features vectors are associated with their corresponding activity through a training process using an ANN classification algorithm with 10 hidden neurons. For evaluation of our classifier, we adopt a cross validation method of testing using a *new person* data which was not used in training the classifier. Test data from a *new person* consisting of 1416 observations of different activities is introduced to the trained network and results obtained achieve a recognition accuracy of 85.3% as shown in the confusion matrix plot of Figure 7. The confusion matrix shows the number/percentages of true positives, true negatives, false positives and false negatives of observations from the *new person* test data set. The axes represent the 6 activities chosen for this test.

After training and validation of the classifier using the extracted activity features vector, each frame of a *new person* test data is classified to the closest posture feature within the activity features vector. The features vector which represent centroids of closely

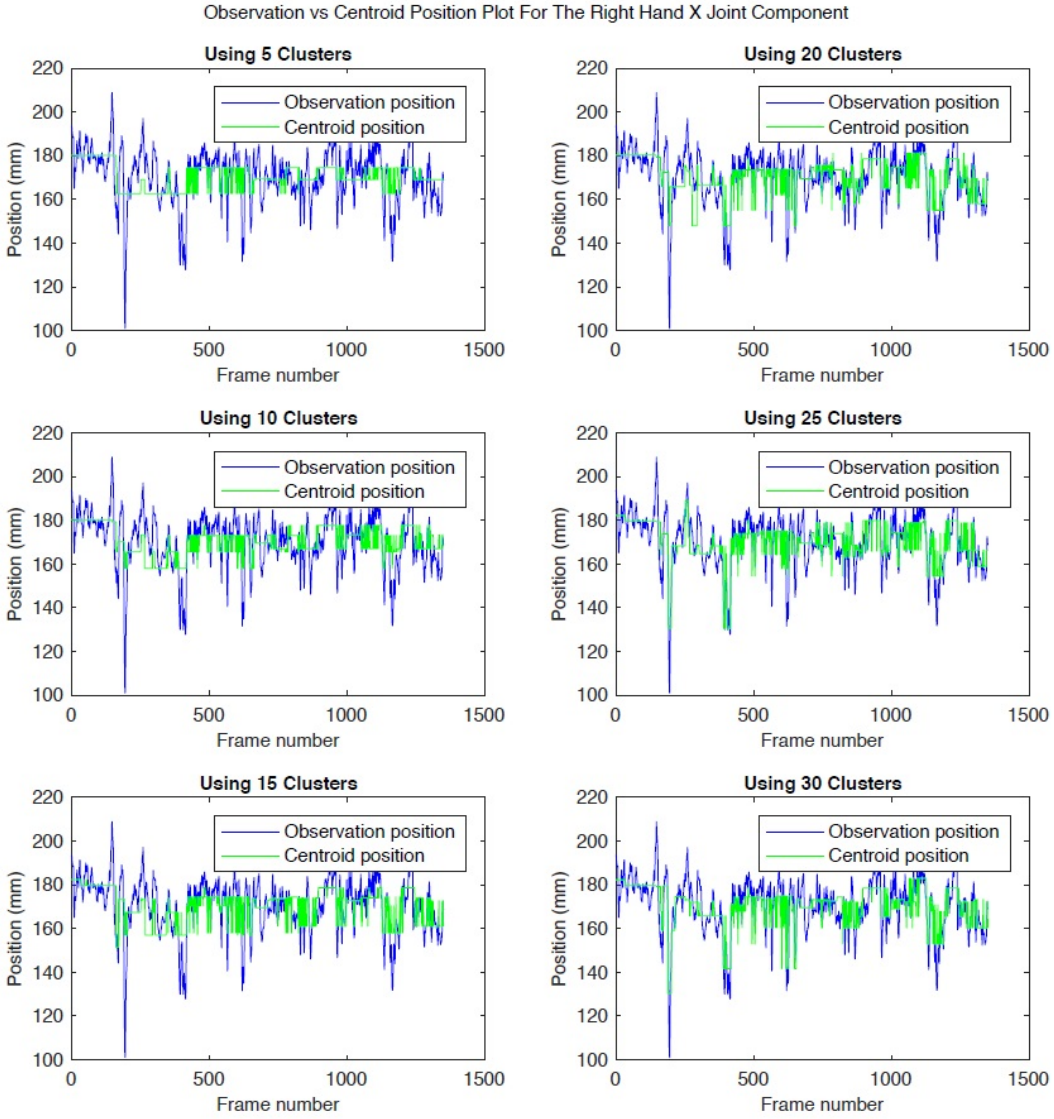

Figure 8: Result of different number of clusters in an activity for one of the key joints in performing an activity of brushing teeth from CAD-60 dataset.

related joint components for a pose and therefore an input frame vector of joints positions can be associated to a corresponding feature vector for the activity. The system does not have to wait for a complete sequence of a *new person* activity before classifying the activity performed. Considering that, the proposed system should be capable of fulfilling real-time performance which will be explored in subsequent research work.

5 CONCLUSION AND FUTURE WORK

The work in this paper presented a method of learning human activities via computing activity features vector by clustering raw joint positions extracted from human skeleton representation from an RGB-D sensor. The method can be applied to a human assistive robot which would be capable of performing these activities whenever the activity is detected. The robot can perform the activity via coordinate mapping of its joints to the human skeleton joints. However, this is beyond the scope of this paper.

The classification of activities result show that some observations from the test data are classified wrongly. This could be as a result of a feature vector representing a particular pose in the activity similar to that of another activity. Therefore, part of the future work is to investigate the degree of membership to which a feature vector belongs to different activities in order to distinctly classify the activity using more sophisticated computational intelligence method.

Another aspect not considered in this work is temporal information of performed human activities which could be very useful in differentiating similar joint coordinates position of differing activities. Also, in selection of the best amount of clusters K given a specific activity, other approaches could be explored for example Bayesian Information Criterion (BIC) [17] as a way to score and select an optimal value for K. This will be considered in further research.

This method can be used in assisted living/elderly care to have assistive robots used in performing most ADLs for elderly or people incapable of performing activities within a living environment effectively. The results show a very promising outcome in the area of assistive robots' coexistence in human living environments.

REFERENCES

[1] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. 2013. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters* 34, 15 (2013), 1799–1807.

[2] Alexandros Andre Chaaraoui, Jose Ramon Padilla-Lopez, Pau Climent-Perez, and Francisco Florez-Revuelta. 2014. Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert systems with applications* 41, 3 (2014), 786–794.

[3] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 1110–1118.

[4] Diego R. Faria, Cristiano Premebida, and Urbano Nunes. 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, 732–737.

[5] Ye Gu, Ha Do, Yongsheng Ou, and Weihua Sheng. 2012. Human gesture recognition through a kinect sensor. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 1379–1384.

[6] F. Han, B. Reily, W. Hoff, and H. Zhang. 2016. Space-Time Representation of People Based on 3D Skeletal Data: A Review. *Computer Vision and Image Understanding* (2016), 1–21.

[7] De-An Huang and Kris M Kitani. 2014. Action-reaction: Forecasting the dynamics of human interaction. In *European Conference on Computer Vision*. Springer, 489–504.

[8] Mohamed E Hussein, Marwan Torki, Mohammad Abdelaziz Gowayyed, and Motaz El-Saban. 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, 2466–2472.

[9] Jose Antonio Iglesias, Plamen Angelov, Agapito Ledezma, and Araceli Sanchis. 2010. Human activity recognition based on Evolving Fuzzy Systems. *International journal of neural systems* 20, 5 (2010), 355–364.

[10] Ahmad Jalal and S. Kamal. 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, 74–80.

[11] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. 2013. Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research* 32 (2013), 951–970.

[12] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. 2015. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems* 80, C (2015), 14–23.

[13] Microsoft. 2017. Developing with Kinect for Windows. https://developer.microsoft.com/en-us/windows/kinect/develop. (2017). Accessed: 2017-02-28.

[14] Orasa Patsadu, Chakarida Nukoolkit, and Bunthit Watanapa. 2012. Human gesture recognition using Kinect camera. In *Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on*. IEEE, 28–32.

[15] David Reby, Sovan Lek, Ioannis Dimopoulos, Jean Joachim, Jacques Lauga, and Stéphane Aulagnier. 1997. Artificial neural networks as a classification method in the behavioural sciences. *Behavioural processes* 40, 1 (1997), 35–43.

[16] Miguel Reyes, Gabriel Domínguez, and Sergio Escalera. 2011. Featureweighting in dynamic timewarping for gesture recognition in depth data. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 1182–1188.

[17] Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464.

[18] T Subetha and S Chitrakala. 2016. A Survey on human activity recognition from videos. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)*. IEEE, 1–7.

[19] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 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, 47–55.

[20] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2012. Unstructured human activity detection from RGBD images. In *2012 IEEE International Conference on Robotics and Automation*. IEEE, 842–849.

[21] Sabanadesan Umakanthan, Simon Denman, Clinton Fookes, and Sridha Sridharan. 2014. Multiple instance dictionary learning for activity representation. In *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 1377–1382.

[22] Cornell University. 2009. Cornell activity datasets CAD-60. http://pr.cs.cornell.edu/humanactivities/data.php. (2009). Accessed: 2017-02-28.

[23] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. 2014. Human Action Recognition by Representing 3D Skeletons As Points in a Lie Group. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. IEEE Computer Society, Washington, DC, USA, 588–595.

[24] Ping Wei, Nanning Zheng, Yibiao Zhao, and Song-Chun Zhu. 2013. Concurrent action detection with structural prediction. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3136–3143.

[25] Huiquan Zhang and Osamu Yoshie. 2012. Improving human activity recognition using subspace clustering. In *Machine Learning and Cybernetics (ICMLC), 2012 International Conference on*, Vol. 3. IEEE, 1058–1063.