Human Activities Transfer Learning for Assistive Robotics

David Ada Adama, Ahmad Lotfi, Caroline Langensiepen, and Kevin Lee

School of Science and Technology
Nottingham Trent University, Nottingham, NG11 8NS, UK
david.adama2015@my.ntu.ac.uk,

Abstract. Assisted living homes aim to deploy tools to promote better living of elderly population. One of such tools is assistive robotics to perform tasks a human carer would normally be required to perform. For assistive robots to perform activities without explicit programming, a major requirement is learning and classifying activities while it observes a human carry out the activities. This work proposes a human activity learning and classification system from features obtained using 3D RGB-D data. Different classifiers are explored in this approach and the system is evaluated on a publicly available data set, showing promising results which is capable of improving assistive robots performance in living environments.

Keywords: Activity Recognition, Activity Classification, Assistive Robotics

1 Introduction

Assistive robots deployed in living environments for applications such as elderly care should learn tasks by observing human carers performing routine duties. To achieve this goal, the assistive robots must be equipped with abilities to learn activities. This requires extracting descriptive information of the activities and classify them while they are performed by a human.

Learning human activities by an assistive robot can be classified under two methods [1]; *Independent Learning* which learn an activity from scratch or learning by making use of transferred knowledge and information which is referred to as *Transfer Learning*. Independent learning is a method whereby an assistive robot learns to perform an activity independently without any prior knowledge of the activity. For example, an assistive robot learning an activity such as cooking (chopping vegetables) or opening a pill container without prior information of how a person would perform the activity. This requires more time in learning and more cost in-cured which are limitations of the method. On the other hand, transfer learning methodology allows information acquired from prior experience to assist in learning an activity [3].

In the context of this paper, an assistive robot can learn to perform an activity from knowledge acquired as it observes a person perform similar activity. This enables faster learning of activities and allows collaboration and adaptation of

robots within living environments. Regardless of the method applied to learning an activity, the availability of descriptive information affects the understanding of an activity. Variations in information and understanding about an activity performed by a person and a robot performing similar activity can be defined as contained within a *knowledge gap* and transfer learning helps to bridge this gap.

Human activities are diverse in nature with imprecision, vagueness, ambiguity and uncertainty in information about the way activities are performed. Thus, variabilities are encountered when an assistive robot tries to learn activities. This affect correct classification of human activities which is relevant in improving the amount of knowledge that can be used by a robot in learning. To capture imprecisions and uncertainties, fuzzy logic has proven to be a suitable method which allows incorporation of imprecisions and uncertainty expressiveness within information [3][4] can be applied to classify human activities. Combining this method with transfer learning would improve assistive robots learning human activities from observing while activities are performed. Other learning techniques applied to learning/classifying human activities are limited in their ability to handle vagueness, imprecision and uncertainties in activities when considering acquiring knowledge that can be transferred across different learners.

In this paper, a method for learning and classifying activities carried out by humans in the context of assistive robotics are presented. Set of features representing daily activities are extracted from human activities and these features are used as input to a classifier to find relevant structures within the features. Classification of activities is done by exploiting different classification techniques; a multiclass Support Vector Machine (SVM), K-Nearest Neighbour (K-NN) and also, Fuzzy C-means (FCM) clustering technique. A cross-validation test is performed on the trained classifier to measure their performance in predicting activities. The aim of the proposed work presented in this paper is to build a human activity learning and classification system that can be incorporated in an assistive robot to improve human-robot interaction in living environments.

The structure of this paper is 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 for feature extraction and classification of activities. Initial results are presented in Section 5. Section 6 presents conclusions and future work to be undertaken.

2 Related Work

Learning and classification of human activities is often referred to as Human Activity Recognition (HAR) [9][10]. One of the main objectives is to extract descriptive information (i.e. features) from human activities to be able to distinctly characterize and classify one activity from another. An integral component of learning an activity is how information of the activity is obtained (i.e. observation). For human activities, information obtained using visual and non-visual sensors makes it a lot easier to understand and learn activities as they are per-

formed. 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 characterizing an activity [11]. Additional depth information using RGB-D sensors provide several advantages as they are better suited for observing human activities to detect human pose used to build activity recognition systems.

To effectively characterize activities from information obtained using RGB-D sensors, machine learning and reasoning methods have been applied by many researchers [12][13][14]. These methods provide an understanding of how activities are learned and relationships between activities. However, there is some uncertainty that exist in how one actor performing an activity would differ from another actor performing similar activity. This hinders HAR systems from going mainstream.

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 [15][16]. Using RGB-D sensors extracts $3D$ skeleton data from depth images and body silhouette for feature generation. In [15], 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 [17] 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 [18] 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.

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 is proposed [19]. The $3D$ joint position information extracted from the sensor are transformed feature vectors by applying K-means clustering to group key postures of an activity. The posture features are used as input to a neural network for classification of the human activities. Authors in [15] proposed a combination of multiple classifiers to form a Dynamic Bayesian Mixture Model (DBMM) to characterize activities using features obtained from distances between different parts of the body. Also, [20] applied statistical covariance of 3D joints (Cov3DJ) as features to encode the skeleton data of joint positions. Another approach seen in [21] used a sequence of joint trajectories and applied wavelets to encode each temporal sequence of joints into features.

3 Activity Features

In the proposed system, the process starts by obtaining RGB-D sensor information from the performed activities. The architecture of the proposed system is shown in Figure 1. Incoming data is obtained using a Kinect RGB-D sensor [22]

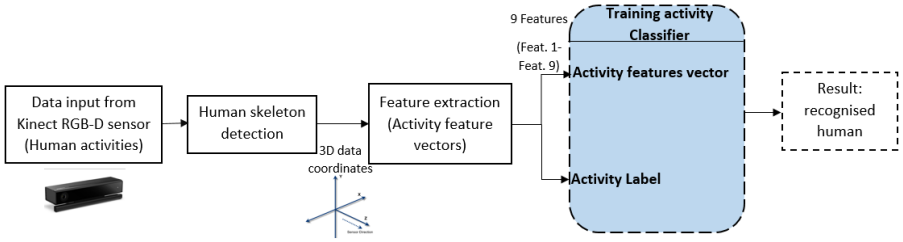

Fig. 1: Architecture of proposed system.

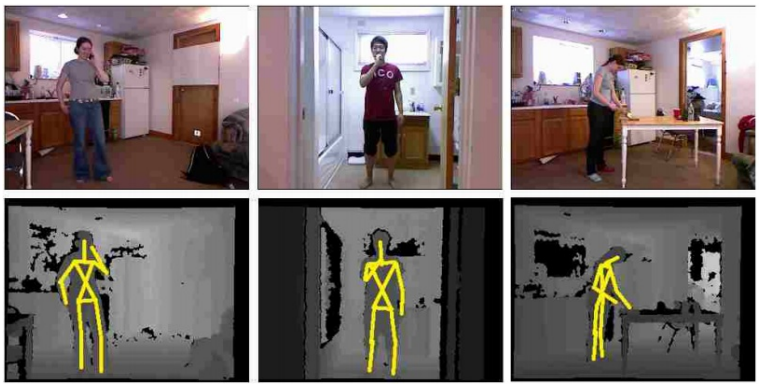

Fig. 2: Frames of human activities performed in a living environment extracted using an RGB-D sensor [23].

which tracks human joint movements and their transitions over time. Data pre-processing and $3D$ skeleton-based feature selection are performed before before they are applied to the classifier. More details are provided below.

3.1 Data Pre-processing

Data is obtained from $3D$ $\{x, y, z\}$ skeleton detection of an actor performing an activity. The skeleton of the actor is tracked using an RGB-D sensor for obtaining positions of joints of the human body. The data representing an activity consist of N number of frames (observations). An example frames of human activities obtained using the RGB-D sensor which shows the tracked skeleton of human joints is shown in Figure 2 [23]. The Kinect RGD-D sensor considers the skeleton frame of reference from the sensor. However, for better representation of the features of an activity, the frame of reference for all joints relative to the torso centroid coordinates is considered.

For a skeleton frame consisting of joints j, the torso centroid coordinate is represented as j_t. The distance between the i^{th} joint j_i and j_t is given as $d_i = j_i$

- j_t. This distance is computed for all joints in each frames of an activity. After computing distances, each frame n is represented by a vector containing joints distances relative to the torso $V_n = \{d_1, d_2, d_3, ..., d_i\}$.

3.2 3D Skeleton-Based Features

Feature extraction is an important aspect of any activity recognition system as raw data obtained from activities do not provide enough information to allow implementing an activity recognition system. The joint distance vectors obtained from the pre-processing stage is converted into a set of useful features that model human activities.

Features obtained in human activity recognition systems can be computed using human skeleton joint position coordinates obtained from an RGB-D sensor. The features are often based on raw joint positions and displacement-based representations when considering temporal and spatial data. In this work, displacement features from skeleton joint coordinates are used. We exclude temporal information to make the system independent of speed of joint movements.

The features used in this work are similar to the ones proposed by [15]. These features are obtained from joint displacement positions of a person performing an activity.

The features are based on distance between both left and right hands, as a lot of attention is drawn towards the pose of the hands when performing an activity. Distance between hands and head, between hip and feet, shoulder and feet, between the initial hand (for both hands and elbows) position of the first frame and the next frames. These are computed using the Euclidean distance equation given as $\delta_{(j_{b1}, j_{b2})}$.

$$\delta_{(j_{b1}, j_{b2})} = \sqrt{(j^x_{b1} - j^x_{b2})^2 + (j^y_{b1} - j^y_{b2})^2 + (j^z_{b1} - j^z_{b2})^2} \tag{1}$$

where the joints of a human skeleton are represented by j_b for b= {*face, hand, shoulder, hip, feet and torso*}. Each joint coordinate is represented in 3D {x, y, z}. The Euclidean distance computed represent features f of an activity

To classify different activities, each activity is represented by a set of feature vectors which characterize the activity as explained above and classification is done on this feature vector. Therefore, an activity A is characterized by features $A = \{f_1, f_2, f_3, ..., f_m\}$, where f_m is the mth feature vector for the activity.

3.3 Features Normalization

Features extracted from an activity can be heterogeneous and this could introduce problems during classification if one of the selected features varied more than another. To avoid this problem, data normalization is performed on the selected activity features and the normalized features are used as input to train and validate the classifiers. In order to normalize our features, the mean and standard deviation of each feature vector is determined and we create new feature set that has zero-mean and a unit standard deviation using equation 2.

This is done to remove distortion due to data heterogeneity before classification is done with the normalized features.

$$\text{Normalized feature} = \frac{f_m - \mu_m}{s_m}, \tag{2}$$

where, s_m is the standard deviation and μ_m is the mean of an activity feature f_m.

4 Activity Classification

The final stage in learning human activities is classification of activities using the extracted feature vectors. This step aims to associate feature vectors to the correct activity. As stated in Section 1 different classification techniques are used in order to classify activities. Support Vector Machine (SVM), K-Nearest Neighbour (K-NN) and also, Fuzzy C-means (FCM) are frequently used in many classification problems and they are also exploited here. However, the FCM algorithm is not commonly used but poses to be a good method for classifying activities. In this algorithm, several features which characterize an object are assigned to different classes with different membership grades. A benefit of using this method for classification is that an initial knowledge of the feature vectors is not required as membership functions are formed automatically by the method.

4.1 Support Vector Machine (SVM)

Considering the application of SVM in classifying activities we apply a method used in [24] where a multi-class SVM is applied to activity recognition. The multi-class SVM is an extension of the SVM from binary classifier. A *"one against-one"* approach which is based on the construction of several binary SVM classifiers is stated to be the most suitable for practical use. This method is necessary for M classes dataset, where $M > 2$. A training phase is carried out during which the activity features are given as input to the multi-class SVM together with activity labels. In the test phase, activity labels are obtained from the classifier.

4.2 K-Nearest Neighbour (K-NN)

The K-NN is among one of the simplest machine learning algorithms and is a method of classifying objects based on closest training points in the feature space. An object is assigned to a class most common among its k nearest neighbours (where k is a positive integer) by a majority of votes of its neighbours. In most cases, Euclidean distance is used as the metrics in finding the nearest neighbours to an object. Applying this method in the proposed approach, in the training phase, the activity feature vectors and activity labels of the training set are stored. During the classification phase, the user defined constant k and unlabelled activity feature vectors are classified by assigning a label most frequent among the k training samples.

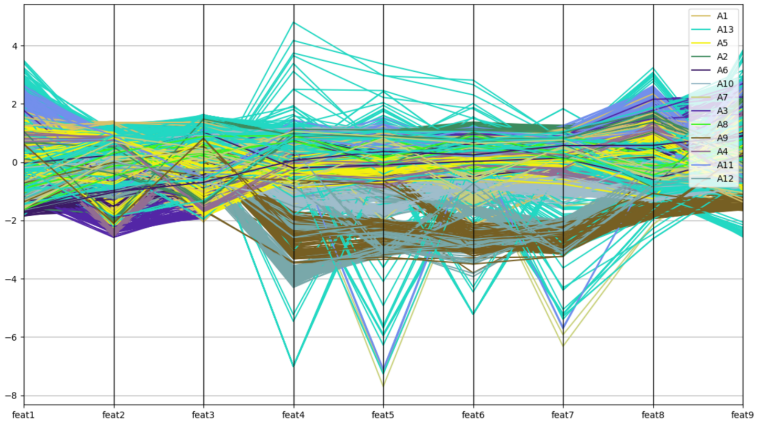

Fig. 3: Parallel coordinate plot showing selected 9 features ($f1 - f9$) of 13 activities ($A1$ - $A13$) obtained from the CAD-60 human activity dataset.

4.3 Fuzzy C-means Algorithm

Fuzzy c-means (FCM) algorithm is a method of clustering which allows one piece of data to belong to two or more clusters. It is frequently used in pattern recognition. Although, FCM is primarily used to cluster data, it could also be employed as a classifier to provide a measure of belonging to each cluster. This is an interesting approach for activity recognition as it will provide a measure of membership to each of the identified classes. Readers are referred to [2] for more details about FCM.

5 Experimental Results

The proposed approach described in this paper is evaluated using publicly available human activity dataset, CAD-60 data set [16]. This data comprises RGB-D sequence of human activities acquired using an RGB-D sensor. 12 activities and an addition of a random + still activity performed by four different participants in five different locations namely; bathroom, bedroom, kitchen, living room and office environments. The activities are listed as follows, with the labels corresponding to the labels shown in the results diagram.

A1 Rinsing mouth,
A2 Brushing teeth,
A3 Wearing Lens,
A4 Talking on the Phone,
A5 Drinking water,
A6 Opening pill container,
A7 Cooking (chopping),
A8 Cooking (stirring),

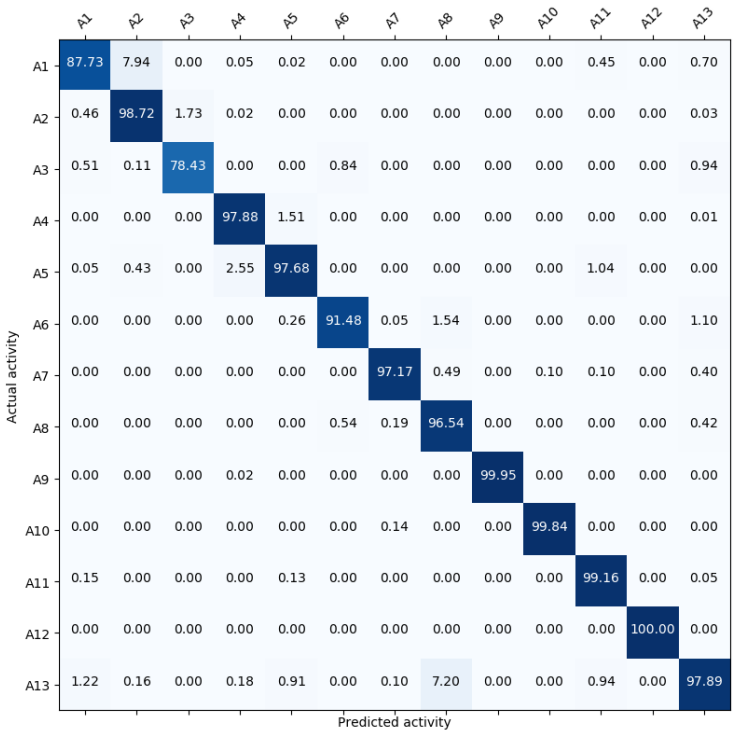

Fig. 4: Confusion matrix plot showing the performance of SVM classifier for classification of 13 activities ($A1$ - $A13$) obtained from the CAD-60 human activity dataset.

A9 Talking on the couch,
A10 Relaxing on couch,
A11 Writing on board,
A12 Working on computer and
A13 Random + still activity.

The first step is data pre-processing which is performed on the data set to obtain each joint coordinate relative to the torso coordinate. Features are then calculated from the pre-processed data using method described in Section 3.2 and 9 features are obtained for each activity. These features are used as input to the classifiers. In Figure 3, a parallel coordinate plot showing the 9 features selected across a sample of observations of the different activities is shown. This shows how the features corresponding to different activities are appear to be similar, thus making the process of classification complicated.

We present the classification results for SVM and K-NN classifiers in terms of *Precision*, *Recall* shown in Table 1 and confusion matrices presented in Figure 4 and 5 for the overall classification of the activities. For testing of the trained

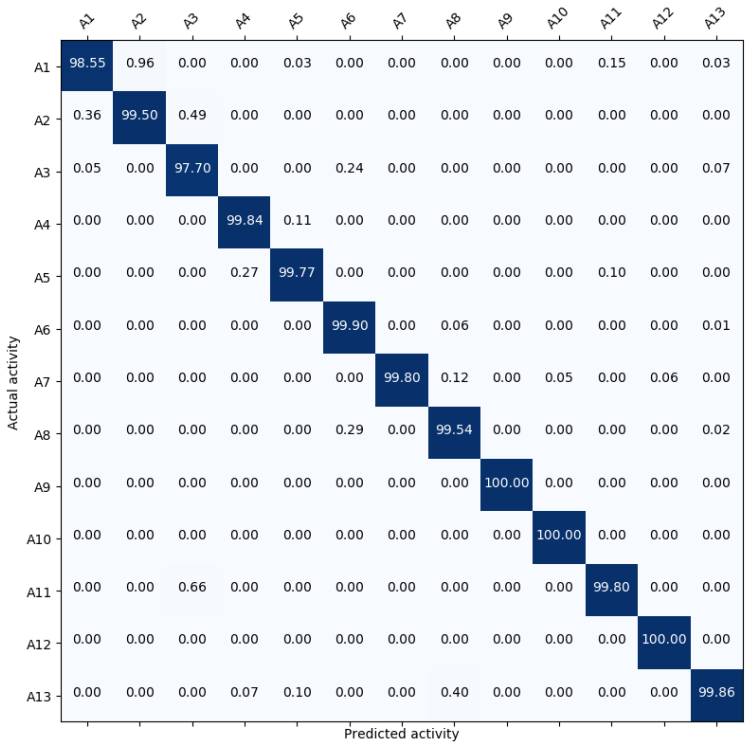

Fig. 5: Confusion matrix plot showing the performance of K-NN classifier for classification of 13 activities ($A1$ - $A13$) obtained from the CAD-60 human activity dataset.

classifier, we use a method of *leave-one-out* cross-validation strategy in which 70% of the data set is used in training the classifier and the rest 30% is used for testing and validation of the classifier.

It can be observed from the results presented in Table 1 that Using the SVM classifier, we obtain classification accuracy of 97.02% on the test activities data. In Figure 4, a confusion matrix for SVM classification results is shown, where the last column of the matrix (i.e. column 13) has the random activity which is a neutral activity performed by the participants (activities that were not classified with high confidence). This is included to show the confidence of our approach. In Figure 5, similar result is also shown when we use a K-NN classifier. However, with the K-NN classifier we attain an accuracy of 99.73% on the test activities data. Note that the results presented are for classification on *'have seen'* test activities data after the classifiers are trained.

For the classification using the FCM algorithm, 13 clusters are selected which represent the number of activities in the data set to be classified. The metrics usually applied to clustering results analysis are; *Purity*- which is an external

Table 1: Result of SVM and K-NN classifier used in overall classification of 13 human activities obtained from the CAD-60 data set. The table presents precision and recall scores for both classifiers when 9 features are used for classification on 'have seen' person.

Activity	SVM Classifier		K-NN Classifier	
	Prec	Rec	Prec	Rec
Rinsing mouth	97.03	88.39	99.58	98.55
Brushing Teeth	92.96	98.69	99.04	99.49
Wearing Lens	98.61	76.69	98.83	97.70
Talking on the phone	97.29	97.78	99.66	99.84
Drinking water	97.05	97.86	99.75	99.77
Opening pill container	97.60	92.47	99.46	99.90
Cooking (chopping)	99.75	96.40	100.0	99.80
Cooking (stirring)	91.03	96.97	99.42	99.53
Talking on the couch	100.0	99.95	100.0	100.0
Relaxing on couch	100.0	99.89	99.94	100.0
Writing on board	97.77	98.95	99.75	99.80
Working on computer	100.0	100.0	99.94	100.0
Random + still activity	96.50	97.89	99.86	99.86
Average	97.35	97.02	99.63	99.73

evaluation criterion for cluster quality, *Normalized Mutual Information* (NMI)- and *Rand Index* (RI). The best result for FCM classification is obtained when we apply a fuzziness coefficient $\phi = 1.4$. Higher values of ϕ result in more overlap between clusters and lower values result in less overlap between clusters which could result in hard clustering. After clustering we obtain the results shown in Table 2.

Table 2: Fuzzy C-means classification result of 13 human activities obtained from the CAD-60 data set. The table shows the metrics used in evaluating the results when 9 features are used for classification

Evaluation metric	Score
Purity	0.55
Normalized Mutual Information (NMI)	0.52
Rand Index (RI)	0.30

6 Conclusions

In this work, classification methods for human activities using 9 features extracted from human activities data collected using an RGB-D sensor is presented. This is part of an on-going research for transfer learning of human activities using

assistive robots. It can be observed from Figure 3, the complexity of human activities using the selected features. This requires proper selection of informative features which provide relevant information that could be used in distinctly characterizing activities. Thus, future work will focus on feature extraction methods for human activities.

The purpose of classifying human activities is to be able to build a system to distinctly characterize activities as they are performed in living environments in order to have assistive robots learn to perform the activities.

References

1. K. Weiss, T. M. Khoshgoftaar and D. Wang.: A survey of transfer learning. Journal of Big Data, vol. 3, no. 9, (2016).
2. Bezdek, J. C.,:Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981).
3. Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., Zhang, G.: Transfer learning using computational intelligence: a survey. Knowledge-Based Systems, vol. 80, (2015), pp. 14-23,.
4. Shell, J., Coupland, S.: Fuzzy transfer learning: methodology and application. Information Sciences, vol. 293, (2015), pp. 59-79.
5. Bellman, R. E., Zadeh, L. A.: Decision-making in a fuzzy environment. Manage. Sci., vol. 17(4), (1970), pp. 141-164.
6. Shell, J., Coupland, S.: Towards Fuzzy Transfer Learning for Intelligent Environments. Ambient Intelligence, vol. 7683, (2012), pp. 145-160.
7. Behbood, V., Lu, J., Zhang, G.: Fuzzy refinement domain adaptation for long term prediction in banking ecosystem. IEEE Trans. Industr. Inf., 10 (2) (2014), pp. 16371646.
8. Behbood, V., Lu, J., Zhang, G.: Fuzzy bridged refinement domain adaptation: long-term bank failure prediction. Int. J. Comput. Intell. Appl., 12 (01) (2013).
9. Iglesias, J. A., Angelov, P., Ledezma, A., Sanchis, A.: Human Activity Recognition Based on Evolving Fuzzy Systems. International Journal of Neural Systems, vol. 20(05), (2010), pp. 355-364.
10. Zhang, H., Yoshie, O.: Improving human activity recognition using subspace clustering. IEEE Machine Learning and Cybernetics (ICMLC), vol. 3, (2012), pp. 1058-1063.
11. Han, F., Reily, B., Hoff, W., Zhang, H.: Space-Time Representation of People Based on 3D Skeletal Data: A Review. Computer Vision and Image Understanding, vol. 158, (2017), pp. 85-105.
12. Koppula, H. S., Gupta, R., Saxena, A.: Learning Human Activities and Object affordances from RGB-D videos. The International Journal of Robotics Research, vol. 32, (2013), pp. 951-970.
13. Shao-Zi Li, Bin Yu, Wei Wu, Song-Zhi Su and Rong-Rong Ji. Feature learning based on SAEPCA network for human gesture recognition in RGBD images, Neurocomputing, vol. 151, (2015), pp. 565573.
14. Kviatkovsky, I., Rivlin, E. and Shimshoni, I.: Online Action Recognition Using Covariance of Shape and Motion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

15. Diego R. Faria, Cristiano Premebida, Urbano Nunes. A Probabilistic Approach for Human Everyday Activities Recognition using Body Motion from RGB-D Images. 23rd IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN. IEEE, (2014) pp. 732737.
16. Jaeyong Sung, Colin Ponce, Bart Selman, Ashutosh Saxena. Human Activity Detection from RGBD Images. In Proceedings of the 16th AAAI Conference on Plan, Activity, and Intent Recognition (AAAIWS11-16). AAAI Press, (2011), pp. 4755.
17. Ahmad Jalal and S. Kamal. Real-time life logging via a depth silhoueebased human activity recognition system for smart home services. In 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS, (2014), pp. 7480.
18. Ye Gu, Ha Do, Yongsheng Ou, and Weihua Sheng. Human gesture recognition through a kinect sensor. In IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, (2012), pp. 13791384.
19. David Ada Adama, Ahmad Lotfi, Caroline Langensiepen, Kevin Lee, and Pedro Trindade. Learning Human Activities for Assisted Living Robotics. In Proceedings of 10th Conference on Pervasive Technology Related to Assistive Environments (PETRA'17), Island of Rhodes, Greece, June 21-23, 2017.
20. Mohamed E Hussein, Marwan Torki, Mohammad Abdelaziz Gowayyed, and Motaz El-Saban. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence. AAAI Press, Beijing, China, (2013), pp. 24662472.
21. Ping Wei, Nanning Zheng, Yibiao Zhao, and Song-Chun Zhu. Concurrent action detection with structural prediction. In Proceedings of the IEEE International Conference on Computer Vision. IEEE, (2013), pp. 31363143.
22. Microsoft, Developing with Kinect for Windows. https://developer.microsoft.com/en-us/windows/kinect/develop
23. Cornell activity datasets CAD-60. http://pr.cs.cornell.edu/humanactivities/data.php
24. Enea Cippitelli, Samuele Gasparrini, Ennio Gambi, and Susanna Spinsante.:A Human Activity Recognition System Using Skeleton Data from RGBD Sensors. Computational Intelligence and Neuroscience, vol. 2016.