Activity Recognition for Natural Human Robot Interaction

Addwiteey Chrungoo[1], SS Manimaran and Balaraman Ravindran[2]

[1] School of Engineering and Applied Science, University of Pennsylvania, Philadelphia PA 19104, USA
[2] Department of Computer Science, Indian Institute of Technology Madras, Chennai, India

Abstract. The ability to recognize human activities is necessary to facilitate natural interaction between humans and robots. While humans can distinguish between communicative actions and activities of daily living, robots cannot draw such inferences effectively. To allow intuitive human robot interaction, we propose the use of human-like *stylized gestures* as communicative actions and contrast them from conventional activities of daily living. We present a simple yet effective approach of modelling pose trajectories using directions traversed by human joints over the duration of an activity and represent the action as a histogram of direction vectors. The descriptor benefits from being computationally efficient as well as scale and speed invariant. In our evaluation, the descriptor returned state of the art classification accuracies using off the shelf classification algorithms on multiple datasets.

1 Introduction

As robots are employed to perform wide range of tasks, especially in human environments, the need to facilitate natural interaction between humans and robots is becoming more pertinent. In many roles, such as, indoor personal-assistants, robots must be able to infer human activities and decipher whether or not a human needs assistance. For e.g., if a robot could recognize whether a person is drinking water, it could offer to pour more and react appropriately based on the person's response. In such scenarios, in addition to recognizing the drinking activity, the robot needs to be capable of recognizing communicative actions, so as to infer whether it should pour more or stop. This is similar in principle to *how humans assist others*, i.e., either they assist if assistance is sought or they foresee the need for assistance based on perception and acquired knowledge. Though past works [10] have focussed on estimating human intent to take such decisions, this work is motivated by the need for interaction between the robot and human as a factor in deciding on an appropriate behaviour. Incorporating such natural interactions is not easy when robots work in highly cluttered environments where people carry out activities in different ways leading to high variability [14, 7]. However, to best support humans, assistive robots need to

behave interactively like humans, making it imperative to correctly understand the human actions involved.

As a result, we are particularly interested in developing a concise representation for a wide variety of actions; both communicative and conventional activities of daily living. We propose the use of human-like *stylized gestures* as communicative actions and contrast them from conventional activities of daily living. *Stylized gestures* are symbolic representations of activities and are widely used by humans across cultures to communicate with each other when verbal communication is not possible.We hypothesize that such actions have distinct motion intrinsics as compared to conventional activities of daily living and can hence be used effectively to communicate with robots in the absence of verbal means. Before we can begin to develop a system for activity recognition, we need an efficient representation mechanism for human motion.

In this work we introduce a novel activity descriptor: Histogram of Direction vectors (HODV) that transforms 3D spatio-temporal joint movements into unique directions; an approach that proves to be highly discriminative for activity recognition. As shown in Figure 1, we represent skeletal joint movements over time in a compact and efficient way that models pose trajectories in terms of directions traversed by human joints over the duration of an activity. The issue we address in this paper is as follows: Learn to recognise various human actions given a direction-vector histogram representation using three dimensional joint locations as raw data. Further, learn to distinguish communicative actions to instruct a robot from conventional activities of daily living and obtain a descriptive labelling of the same. We show that our proposed approach is efficient in distinguishing Communicative and Non Communicative activities in our novel RGBD dataset and also performs equally well on two public datasets: Cornell Activity Dataset (CAD -60) and UT-Kinect Dataset using off the shelf classification algorithms.

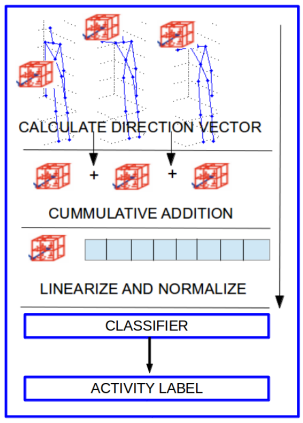

Fig. 1: The general framework of the proposed approach

1.1 Contributions and Outline

The contributions of this work are are as follows: Firstly, we introduce the problem of communicative vs non-communicative actions. Secondly, we propose a novel and computationally efficient activity descriptor based on pose trajectories. We provide analysis of our algorithm on two public datasets and demonstrate how the algorithm could be used for both Communicative/Interactive and Non-Communicative/Non-Interactive activity recognition. We will also release an annotated RGBD Human Robot Interaction dataset consisting of 18

unique activities including 10 *stylized gestures* as well as 8 conventional activities of daily living (within the same dataset) along with full source code of our algorithm.

The rest of the paper is organized as follows. Section 2 presents a brief literature review. Section 3 explains our dataset, while section 4 and 5 describe our algorithm and experimental results in detail respectively. We conclude the paper in section 6 and also present directions for future work.

2 Related Work

Human activity recognition has been widely studied by computer vision researchers for over two decades. The field, owing to its ability to augment human robot interaction, has recently started receiving a lot of attention in the robotics community. In this section, we restrict ourselves largely to research relevant to robotics, and for an in-depth review of the field, one can refer to recent survey papers [2].

Earlier works focussed on using IMU data and hidden Markov models(HMMs) for activity recognition. Authors in [18] proposed a model based on multi sensor fusion from wearable IMUs. They first classified activities into three groups, namely: Zero, Transitional and Strong displacement activities, followed by a finer classification using HMMs. Their approach was however restricted to very few activity classes and was computationally expensive. Mansur et al.[8] also used HMMs as their classification framework and developed a novel physics based model using joint torques as features; claimed to be more discriminative compared to kinematic features [12]. Zhang et al.[17] followed a vision based approach and proposed a 4D spatio-temporal feature that combined both intensity and depth information by concatenating depth and intensity gradients within a 4D hyper-cuboid. Their method was however dependant on the size of the hyper-cuboid and could not deal with scale variations. Sung et al.[12] combined human pose and motion, as well as image and point-cloud information in their model. They designed a hierarchical maximum entropy Markov model, which considered activities as a superset of sub-activities.

While most of these works focussed on generating different features, work on improving robot perception, including recognizing objects and tracking objects [4] led to the incorporation of domain knowledge [13] within recognition frameworks. Authors in [5] proposed a joint framework for activity recognition combining intention, activity and motion within a single framework. Further, [7, 10] incorporated affordances to anticipate activities and plan ahead for reactive responses. Pieropan et al.[9] on the other hand introduced the idea of learning from human demonstration and stressed the importance of modelling interaction of objects with hands such that robots observing humans could learn the role of an object in an activity and classify it accordingly.

While past works excluded the possibility of interaction with the agent, this work aims to understand activities when interaction between robots and humans is possible and realistic, especially, in terms of the human providing possible

instructions to a robot while also performing conventional activities of daily living. The focus of our work is to utilize distinctions in motion to differentiate between communicative/instructive actions and conventional activities of daily living. Having said this, we do not see motion information alone as a replacement, but as a complement to existing sensory modalities, to be fused for particularly robust activity recognition over wide ranges of conditions.

3 OUR DATASET

Recent advances in pose estimation [11] and cheap availability of RGBD cameras, has lead to many RGBD activity datasets [12, 14]. However, since none of the datasets involved communicative/interactive activities alongside conventional activities of daily living, we collected a new RGBD dataset involving interactive as well as non interactive actions. Specifically, our interactive actions were between a robot and a human; where the human interacts with the robot using *stylized gestures*; an approach commonly used by humans for human-human interaction.

The activities were captured using a kinect camera mounted on a customized pioneer P3Dx mobile robot platform. The robot was placed in an environment wherein appearance changed from time to time, i.e., the background and objects in the scene varied. In addition, the activities were captured at various times of the day leading to varied lighting conditions. A total of 5 participants were asked to perform 18 different activities, including 10 Communicative/Interactive activities and 8 Non-Interactive activities, each performed a total of three times with slight changes in viewpoint from the other instances. *'Catching the Robots attention', 'Pointing in a direction', 'Asking to stop', 'Expressing dissent', 'Chopping', 'Cleaning', 'Repeating', 'Beckoning', 'Asking to get phone' and 'facepalm'* were the 10 Robot-Interactive activities. In Robot-Interactive activities like *'Facepalm'*, the human brings his/her hand up to his head, similarly, the activity *'chopping'* involved a human repeatedly hitting one of his hands with the other hand, creating a stylized chopping action and so on. The non interactive activities were more conventional activities of daily living like *'Drinking something', 'Wearing a backpack', 'Relaxing', 'Cutting', 'Feeling hot', 'Washing face' 'Looking at time'* and *'Talking on cellphone'*.

We stress that our dataset is different from publicly available datasets as we represent a new mix of activities, more aligned with how humans would perform these in real life. In addition, the dataset involves wide variability in how the activities were performed by different people as subjects used both left and right hands along with variable time durations. For e.g., in the *'Drinking something'* activity, some subjects took longer to drink water and brought the glass to their mouth couple of times, while others took the glass to their mouth just once. The wide variety and variability makes recognition challenging. We have made the data available at: `http://rise.cse.iitm.ac.in/activity-recognition/`

4 Action Representation

Activities usually consist of sequences of sub-activities and can be fundamentally described using two aspects: a) Motor Trajectory and b) Activity context. For eg., in a drinking activity, a subject picks a glass or a cup, brings it closer to his/her mouth and returns it. While there are numerous possibilities behind the context of the activity, as a glass could contain juice while a cup could contain coffee, thereby giving more meaning to the activity 'drinking' and answering a question: *What is probably being drunk?* The motor trajectory followed by most people for a generic drinking activity would predominantly be similar. We aim to exploit this similarity and introduce a *local motion* based action representation called *Histogram of Direction Vectors*, defined as the distribution of directions taken by each skeleton joint during all skeleton pose transitions during an activity.

The intuition behind the descriptor is that directions have a clear physical significance and capturing motion intrinsics as a function of direction should be discriminative across classes. We describe the 3D trajectory of each joint separately and construct the final descriptor by concatenating the direction vector histogram of each joint.

4.1 Direction vectors from skeletons

The algorithm takes RGBD images as input and uses the primesense skeleton tracker [1] to extract skeleton joints at each frame. For each joint i, P_f^i represents the 3D cartesian position of joint i at time frame f. The joint locations are then normalized by transforming the origin to the human torso, thereby making them invariant to human translation. Direction vectors are then calculated for each joint i by computing the difference between joint coordinates of frame f and frame $f + \tau$, where τ is a fixed time duration (e.g., 0.1 seconds) in terms of frame counts. Mathematically, direction vectors are estimated for each joint at every frame as:

$$\boldsymbol{d}_f^i = \left[P_f^i - P_{f+\tau}^i\right], \forall f \in [1, 2, \ldots, f_{max} - \tau] \tag{1}$$

The next section explains the construction of our action descriptor, Histogram of direction vectors, and the final descriptor used to classify activities.

4.2 Histogram of direction vectors

At each frame f, the local region around a joint i is partitioned into a 3D spatial grid. We chose 27 primary directions in the 3D space and represented the direction taken by a joint by the nearest primary direction in that grid. The grid entries represent real world directions such as, up, down, up-left, down-right and so on; resulting in a total of 27 directions. The direction vector corresponding to a joint i is mapped onto the index of one of 27 directions, by estimating the 3D euclidean distance between grid coordinates σ_q and the direction vector

$\boldsymbol{d}_f^i$; with a vector being allotted a particular direction index q corresponding to the minimum distance. The goal is to find the specific direction index q^* that represents the direction which is at minimum euclidean distance from the direction vector.

$$q^* = \text{argmin}\|\mathrm{d_f^i} - \sigma_\mathrm{q}\| \quad \forall \mathrm{q} \in [1, 2, \ldots, 27] \tag{2}$$

where σ_q is the coordinate of grid index q.

Let Q^f denote the vector of directions, with Q_q^f denoting the entries of vector Q^f at index q. The grid index q^* is then used to update vector Q^f. To attain the total number of times a particular direction was taken during an activity, we perform cumulative addition of vector Q^f at each frame as shown in equation 4 where h^* is a vector revealing the number of times each direction was taken by a joint during the course of an activity.

$$Q_q^f = \begin{cases} 1 & \text{if } q = q^* \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$h^* = \sum_f Q^f \tag{4}$$

$$h_i = \frac{h^*}{\|h^*\|_1} \tag{5}$$

The vector h^* is then normalized to compute the feature vector h_i for joint i. Normalizing the vector h^* gives us a histogram h_i, representing the probability of occurrence of each direction for a particular joint i, during the course of an activity. Further, each histogram h_i is concatenated to generate the final feature vector $H = [h_1, h_2, \ldots, h_i]$; namely the Histogram of direction Vectors.

5 Experimental Results

In this section we present detailed analysis of our experiments. In addition to our dataset, we test our algorithm on two public datasets: The Cornell activity dataset (CAD-60) [12] and the UTKinect-Action Dataset [14]. Our results reveal that the proposed approach performs comparable to the state of the art approaches, which in general, are computationally expensive and involve complicated modelling. We show how our algorithm, despite being very simple, returns better results; while being computationally inexpensive as well as lower in dimensionality. We use an SVM (LIBSVM) as our classification algorithm along with histogram intersection as the kernel choice. We optimize the cost parameter using cross validation.

5.1 Our Dataset

On our dataset, we ran experiments using three different settings. In the first, we classified actions into their respective categories using the entire dataset. In the second setting, we manually separated the activities into Communicative/Interactive activities and Non-Interactive activities and ran our classification algorithm on the two groups independently. In the third setting, we trained

a two class classifier and labelled the activities as belonging to either of the two groups. All experiments were performed using 5 fold cross subject cross validation, such that, at a time, all instances of one subject were used for testing and the instances from the other subjects were used for training. None of the instances used for training were ever present in the test set at the same time.

It was our observation that not all joints contributed towards an activity. This lead to many joints being binned into the grid representing *no movement*, leading to reduced accuracy. To counter this phenomenon, we masked the feature vector i.e., made the contribution of the corresponding *no movement* bin zero and renormalized. Feature masking resulted in increased accuracy in not only our dataset (Figure 2) but also the CAD 60 and UTKinect Action Datasets.

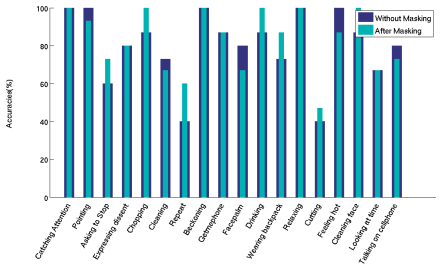

Fig. 2: Comparison on accuracies with and without feature masking

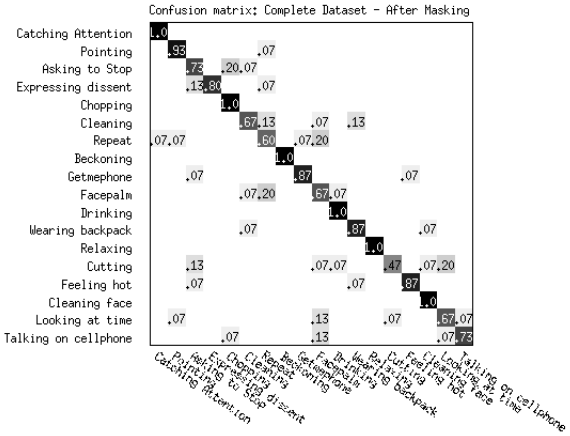

Fig. 3: Confusion matrix of entire dataset using Feature Masking

Figure 3 shows the confusion matrix of our first experimental setting. Most activities are classified with good accuracy apart from *Repeat* and *Facepalm*, mostly because of the similar motion trajectories. Also, as visible in Figure 2 activities such as *Asking to stop, Repeat, Drinking, Wearing backpack* and *Cleaning face* were better classified after feature masking. The average accuracy attained without feature masking was 80%, while with feature masking the average accuracy improved to 82.59%.

Figure 4 shows the confusion matrix of our second experimental setting. The average classification accuracy for Interactive actions was 84%, while for Non Interactive actions, the average accuracy was 86.67%. Like in the previous setup, the algorithm was able to accurately classify actions which had distinct motion trajectories but gets confused with actions with very similar motion like *Repeat* and *Facepalm*.

In the third experimental setup, we classified an activity into either of the two groups. The algorithm achieved an average classification accuracy of 89.26%. Interactive actions were classified with an accuracy of 92.67% while Non Interactive activities were classified correctly with an accuracy of 85%. This classification

paradigm could be essential for the development of hierarchical models where the first level could be an Interactive Vs Non-Interactive classification, followed by a finer categorization into an exact activity.

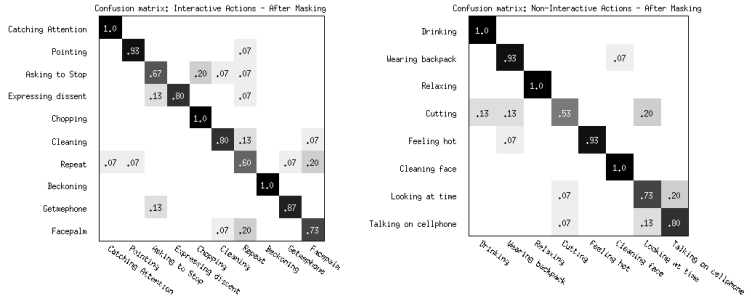

Fig. 4: Left: Confusion matrix of Interactive/Communicative actions after Feature Masking. Right: Confusion matrix of Non-Interactive actions after Feature Masking

Our algorithm is able to distinguish between Interactive and Non Interactive activities with good accuracy. It works well even when subjects take different time duration to complete an activity. Further, since we follow a histogram based representation, classification is invariant to the number of times an action is performed within an activity. For. e.g., a circle could be made once or five times. As long as the feature vector is normalized and if an action is symmetric (activities involving mirror directions eg: waving), the number of times the action is performed or the starting point of the activity would not hamper classification. The descriptor also benefits from being computationally efficient as the only calculations involved for each joints are:

- Calculation of direction vectors, which can be performed in constant time.
- Updating appropriate Histogram bins which is linear in the number of frames and can be performed real-time as and when new frames are captured.

This makes HODV an efficient, yet effective feature vector for classifying human activities.

5.2 Cornell Activity Dataset (CAD 60)

The dataset comprises of 60 RGBD video sequences of humans performing 12 unique activities of daily living. The activities have been recorded in five different environments: Office, Kitchen, Bedroom, Bathroom, and Living room; generating a total of 12 unique activities performed by four different people: two males and two females. We used the same experimental setup (4 fold cross-subject cross validation) and compare precision-recall values for the 'New Person' setting as described in [12]. Table 1 shows a comparison of our algorithm with other state of the art approaches. All of the algorithms mentioned in table 1 use visual features in addition to skeleton data. This work is largely restricted to the use of skeleton data for classification. Hence it would be fair to compare with an approach that uses just skeleton data. The precision recall scores in [12] without visual features is 67.20 and 50.20 respectively. Considering that we use only skeleton data, our approach still outperforms other algorithms.

5.3 UTKinect Action Dataset

Table 1: Comparison of our algorithm with other approaches on the CAD 60 dataset

Method	Precision	Recall
Sung et. al[12]	67.90	55.50
Yang, Tian[15]	71.90	66.60
Ni. et al[3]	75.90	69.50
Gupta et. al[6]	78.10	75.40
Koppula et. al[7]	80.80	71.40
Zhang, Tian[16]	86.00	84.00
Our Descriptor	71.76	70.23
Our Descriptor + Masking	83.77	82.06

The UTKinect Action Dataset [14] presents RGBD video sequences and skeleton information of humans performing various activities from different views. 10 subjects perform 10 different activities namely: *walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands* and *clap hands*. Each subject performs an activity twice.

There are a total of 200 instances of different activities in this dataset. Since each skeleton is described by 20 joints, our feature vector is of dimensions 20×27, i.e., a total of 540 features were used for classification in this dataset. For this dataset, we compare our approach with the state of the art methodology called histogram of 3D skeleton joint positions (HOJ3D)[14] using Leave one Sequence out Cross validation (LOOCV) and cross subject validation as defined previously in this paper. This dataset has activities which look very similar e.g., Sit down and Stand Up. Our high accuracies reveal the superiority of our algorithm in distinguishing such actions, which despite looking similar, have distinct trajectory directions, aptly captured by our approach. The overall accuracies attained on the dataset are shown in Table 2. Clearly, our approach generates better accuracy as compared to the Histogram of 3D joints algorithm under the LOOCV setting. The performs drops a bit under the cross subject crossvalidation scheme. Authors in [14] do not report cross subject results.

6 Conclusion

Table 2: Comparison of our algorithm with HOJ3D on the UT-Kinect dataset

Method	Accuracy
HOJ3D [14] (LOOCV)	90.92
Ours (Cross Subject)	84.42%
Ours (LOOCV)	87.44%
Ours + Masking (Cross Subject)	89.45%
Ours + Masking (LOOCV)	91.96%

This paper presented the problem of Communicative vs Non-Communicative actions and human activity recognition in general. We proposed a novel and computationally efficient activity descriptor, Histogram of Direction Vectors, which aptly captured motion intrinsics and returned good accuracies on our new RGBD dataset. The descriptor proved beneficial in distinguishing between Interactive/Communicative and Non-Interactive activities. Further, results on two public datasets depict its potential in conventional activity recognition frameworks. As part of future work, we would like to combine the descriptor with visual features to cater to cases where the motion trajectories are very similar.

References

1. Nite Skeleton Tracking. `http://http://wiki.ros.org/nite`, accessed: 2014-07-30

2. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. 43(3), 16:1–16:43 (Apr 2011), http://doi.acm.org/10.1145/1922649.1922653
3. BingBing Ni, Pierre Moulin, S.Y.: Order-preserving sparse coding for sequence classification. In: Proceedings of European Conference on Computer Vision. ECCV '12 (2012)
4. Collet, A., Martinez, M., Srinivasa, S.S.: The moped framework: Object recognition and pose estimation for manipulation. The International Journal of Robotics Research (2011)
5. Gehrig, D., Krauthausen, P., Rybok, L., Kuehne, H., Hanebeck, U., Schultz, T., Stiefelhagen, R.: Combined intention, activity, and motion recognition for a humanoid household robot. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. pp. 4819–4825 (Sept 2011)
6. Gupta, R., Chia, A.Y.S., Rajan, D.: Human activities recognition using depth images. In: Proceedings of the 21st ACM International Conference on Multimedia. pp. 283–292. MM '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2502081.2502099
7. Koppula, H., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. IJRR 32(8), 951–970 (2013)
8. Mansur, A., Makihara, Y., Yagi, Y.: Action recognition using dynamics features. In: Robotics and Automation (ICRA), 2011 IEEE International Conference on. pp. 4020–4025 (May 2011)
9. Pieropan, A., Ek, C., Kjellstrom, H.: Functional object descriptors for human activity modeling. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on. pp. 1282–1289 (May 2013)
10. Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. In: In RSS (2013)
11. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: In In CVPR, 2011. 3
12. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on. pp. 842–849 (May 2012)
13. Teo, C., Yang, Y., Daume, H., Fermuller, C., Aloimonos, Y.: Towards a watson that sees: Language-guided action recognition for robots. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on. pp. 374–381 (May 2012)
14. Xia, L., Chen, C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. pp. 20–27. IEEE (2012)
15. Yang, X., Tian, Y.: Effective 3d action recognition using eigenjoints. J. Vis. Comun. Image Represent. 25(1), 2–11 (Jan 2014), http://dx.doi.org/10.1016/j.jvcir.2013.03.001
16. Zhang, C., Tian, Y.: Rgb-d camera-based daily living activity recognition. In: Journal of Computer Vision and Image Processing Vol. 2, No. 4 (December 2012)
17. Zhang, H., Parker, L.: 4-dimensional local spatio-temporal features for human activity recognition. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. pp. 2044–2049 (Sept 2011)
18. Zhu, C., Sheng, W.: Human daily activity recognition in robot-assisted living using multi-sensor fusion. In: Robotics and Automation, 2009. ICRA '09. IEEE International Conference on. pp. 2154–2159 (May 2009)