Robot-centric activity recognition 'in the wild'

Ilaria Gori, Jivko Sinapov, Priyanka Khante, Peter Stone, and J. K. Aggarwal

University of Texas at Austin, Austin TX 78712, USA
{ilaria.gori,aggarwaljk}@utexas.edu,
{jsinapov,pkhante,pstone}@cs.utexas.edu

Abstract. This paper considers the problem of recognizing spontaneous human activities from a robot's perspective. We present a novel dataset, where data is collected by an autonomous mobile robot moving around in a building and recording the activities of people in the surroundings. Activities are not specified beforehand and humans are not prompted to perform them in any way. Instead, labels are determined on the basis of the recorded spontaneous activities. The classification of such activities presents a number of challenges, as the robot's movement affects its perceptions. To address it, we propose a combined descriptor that, along with visual features, integrates information related to the robot's actions. We show experimentally that such information is important for classifying natural activities with high accuracy. Along with initial results for future benchmarking, we also provide an analysis of the usefulness and importance of the various features for the activity recognition task.

1 Introduction

Robots are becoming increasingly sophisticated, and are bound to become pervasive in humans' every-day lives. To effectively collaborate with humans, it is useful for a robot to understand their activities and intentions automatically. This understanding is especially important in human-robot interaction scenarios: if the robot can properly interpret the behavior of humans, its communication with them will be facilitated. For example, in our scenario, a robot moves in a building monitoring the environment. If it could recognize when a person needs help, or whether someone wants to talk to it to ask for directions, or if it is being ignored, its social skills would improve dramatically.

Most existing datasets available to assess activity recognition methods are recorded from a still camera, and they comprise surveillance [10] or sports videos [13]. Others are composed of cinema movies or Youtube videos [9]. Yet others are recorded by asking the participants to perform specific activities [8]. None of these datasets perfectly reflect the types of human activities a robot is likely to perceive when interacting with people. In this work, we present a dataset taken from a robot's perspective, where the camera moves according to the robot's

This work has taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (CNS-1330072, CNS-1305287), ONR (21C184-01), AFRL (FA8750-14-1-0070), and AFOSR (FA9550-14-1-0087).

movements, and the activities are performed spontaneously by humans, often in relation to the presence or behavior of the robot. A few datasets do exist that have been recorded from a robot's perspective [17, 14, 3], even though the robots they used were not fully autonomous. Such datasets were collected by asking participants to perform specific actions. This lack of spontaneity may lead to a low recognition rate when the same activities are executed by people who are not asked to perform them. Besides, there is no guarantee that the chosen staged actions are those that people would naturally perform in front of a robot.

In contrast, our dataset was collected by a mobile robot, able to localize itself and navigate autonomously, moving in a populated environment, and recording people's actions. During the dataset collection, the robot moved autonomously, so that the behavior of the humans was not influenced by any external presence. The recorded data was then analyzed, and the action categories were determined by the activities spontaneously executed by the subjects. This dataset presents some unique characteristics: 1) There can be multiple people in the scene at the same time, each doing something different; 2) There may be occlusions; 3) Actions are performed at different scales and with different body orientations; 4) The robot moves continuously, therefore, there is ego-motion in the scene; 5) Some actions occur more often, while others are very rare, thus the data is highly unbalanced. As we show empirically in the experimental section, these characteristics make learning from our dataset very challenging. To the best of our knowledge, this is the first dataset recorded in such a spontaneous manner.

In this paper, the new dataset is exploited to tackle a human activity recognition task, even though it can be useful for different learning tasks as well. Unlike the previous action recognition methods, which limited their analysis to visual descriptors, we also explore features that are directly related to the robot's behaviors and movements, which, in this particular setting, influence its perceptions.

Our contribution is twofold. First, we present a new problem and make publicly available a novel challenging dataset recorded 'in the wild' from a robot's perspective. Second, we use this dataset to tackle an activity learning task. We provide results obtained using several state-of-the-art descriptors for the purpose of future benchmarking. We also present an analysis of the usefulness and importance of the various features for this specific task. In particular, we show that, in this setting, exploiting data associated with the robot's point of view consistently improves the results obtained when using only visual features.

2 Related Work

Since the early '90s, the computer vision research community has produced a plethora of methods for recognizing human activities (see [1] for a review). In most early approaches the video stream is captured by one or more stationary cameras, e.g., [2]. Methods have also been proposed for human activity recognition in movies [9], where the camera is not always stationary. In some studies, the camera is attached to a person and data is collected as the person performs

a variety of activities, e.g., playing a sport [7], or interacting with others [5]. The primary focus of this past research has been on developing efficient and informative features that enable activity recognition using off-the-shelf supervised machine learning algorithms.

Most relevant to this paper are studies in which video streams were captured by a robot. Such studies are relatively new and include the works of [15, 17, 14, 3]. For example, [17] describes an experiment in which 8 participants are asked to perform 9 activities in front of a teleoperated robot. The data is subsequently used for the development of an activity classification system. Similarly, in [14], 8 participants are asked to perform up to 7 activities in front of a teddy-bear equipped with a camera and mounted on a rolling chair. In [3], the researchers include a larger number of activities (18) performed by 5 participants.

Most robot-centric human activity recognition methods, including the ones described above, are subject to several limitations: 1) The activities are prespecified by the experimenters; 2) The activities are performed by a relatively small number of people (typically 5-8) who are recruited to participate to the dataset collection; 3) The robot is typically either stationary or teleoperated. The present study overcomes these limitations in several significant ways. Our robot uses its autonomous navigation capability in a large and dynamic human-inhabited environment, as opposed to a structured laboratory environment. This results in much more realistic, but also more challenging, video streams. Also, the activities in our study were spontaneously performed by a large number of people who interacted with the robot, as opposed to the standard methodology of asking participants to perform certain actions.

3 Dataset

The robot used to record the dataset is shown in Fig. 1. It was built on top of the Segway Robotic Mobility Platform with an added caster wheel to keep the robot level to the ground. The robot's sensors include a Hokuyo URG-04LX laser rangefinder, used for mapping and localization, and a Kinect RGB-D (version 1.0) camera, used for obstacle avoidance. For this specific experiment, the robot was also equipped with the newer Kinect 2.0 RGB-D camera, which was used for visual person detection and tracking. The robot uses a hierarchical task-planning software architecture [18] based on the Robot Operating System [12].

The robot collected data by autonomously patrolling through an undergraduate and a graduate student lab which were connected by a doorway. To collect the dataset, the robot traversed the environment for 1-2 hours per day, for 6 days. Over the course of the experiment, the robot travelled a total of 14.037 km. As soon as the Kinect 2.0 detected a person, our program started recording all the information described in the next paragraph and summarized in Table 1. Many people just ignored the robot or passed by it. Others engaged in various interactions such as blocking it, waving at it, or taking a picture of it. At the end, we labeled the actions into a number of categories that we observed at least 6 times in the recorded videos. The labeling was carried out by two authors of

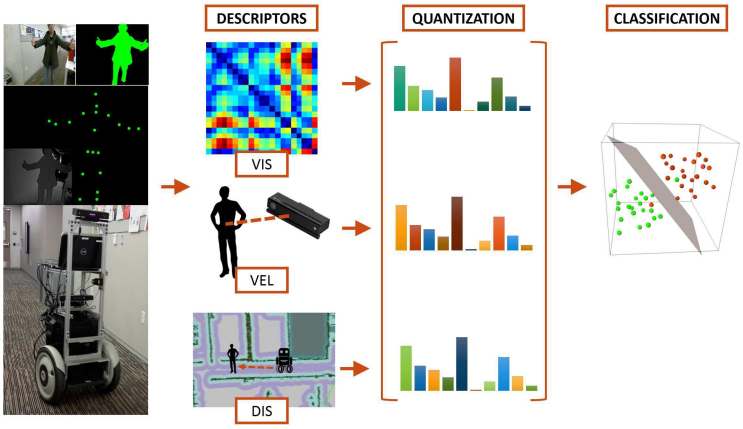

Fig. 1: An overview of our system. Descriptors on visual (VIS), velocity (VEL) and distance (DIS) information are extracted, quantized, concatenated and finally fed to the classifier.

this paper; therefore, it is subjective, prone to errors or different interpretations. The resulting activity categories are: *approach, block, pass by, take picture, side pass, sit, stand, walk away, wave*. There were several more, e.g., three persons approached the robot and pretended to punch it, or started to dance in front of it, but they were too rare and were not included in our subsequent analysis. Notably, the Kinect performance in tracking is not perfect, especially when the robot moves continuously. It may happen, for instance, that a wall, or a chair, or a column is recognized as a person. These samples are gathered in the class *false*, which is used in the classification procedure as well. In total, there are 10 class labels and 1204 selected samples. We plan, however, to record several more hours of activities in the future and expand the released dataset as more data becomes available.

For each video we provide RGB images (3-channel images of dimension 512×424), depth images (16-bits, 1-channel images of dimension 512×424), and the position of the skeletal joints for each person, in 3D and on the image. Some images extracted from our dataset can be observed in Fig. 2. Each video segment is also annotated with the robot's position and orientation in the map, estimated using Monte-Carlo Localization as implemented in ROS, and the robot's raw odometry information, computed by the Segway RMP ROS driver. Each video has been annotated with an activity label, assigned as described in the previous paragraph. This feature set constitutes the first main contribution of this paper, and we intend for it to be useful to the community. While our analysis focuses on human activity recognition, the dataset can be used for other tasks as well.

Table 1: Raw features provided in the newly collected dataset.

Features	Dimension	Range	Sampling Rate (Hz)
RGB images	512×424×3	{0,255}	50
Depth images	512×424	{0,65535}	50
3D Joints	21x3	$\mathbb{R}$	50
2D Joints	21x2	{0,512}×{0,424}	50
Robot's pose on the map	7	$\mathbb{R}$	1.5
Robot's odometry	7	$\mathbb{R}$	100
Activity label	1	{1,10}	–

4 Activity Recognition

We use the newly collected dataset to carry out an activity recognition experiment. In this setting, the task of the robot is to annotate a video with the correct activity label. The raw recorded data was too highly dimensional to be used as direct input to a classifier. Hence, we manipulated the raw sensory data to obtain higher-level feature descriptors. In particular, this section describes what descriptors have been extracted, and how they have been quantized, so that each video is represented by a single vector. Our main proposition is to concatenate robot-centric descriptors – i.e. descriptors related to the robot's perspective – with visual features, as we hypothesize that they will improve the performance of the classifier. Figure 1 shows an overview of the recognition system.

Visual Features : We extract five different visual descriptors and we compare them in the experimental evaluation section. The first one has been proposed in [16], and builds a histogram of the joints in 3D (HOJ3D). The second one has been presented in [6], and computes the covariance of the joint positions over time (COV). The third one has been described in [3], and generates Histograms of Direction Vectors (HODV). The fourth one is based on raw depth images and has been published in [11] (HON4D). Finally, we rely on a simple descriptor that builds a matrix of pairwise relations between joints. We will refer to it as the Pairwise Relation Matrix (PRM). The intuition behind this descriptor is that, while absolute joint positions are not translation invariant, their relations are independent from the absolute position of the person. At the same time, they provide a good representation of the skeleton configuration. Let $J = (\mathbf{j_1}(t), \mathbf{j_2}(t), \ldots, \mathbf{j_m}(t))$ be the set of 3D joints tracked by the Kinect at time t. We build a $m \times m$ matrix $\mathbf{R(t)}$ for each frame t, where m is the number of joints. Each element of $\mathbf{R(t)}$ is equal to

$$R(t)^{i,k} = ||\mathbf{j_i}(t) - \mathbf{j_k}(t)||. \tag{1}$$

Since the resulting matrix is symmetric, we use only the values under the diagonal, therefore the final descriptor belongs to $\mathbb{R}^{\frac{m(m-1)}{2}}$.

Human-Robot Velocity Features : The movements of a person as perceived by the robot are different from his or her movements with respect to an absolute point of view. For example, consider the motion of a person walking away from the robot. We hypothesize that the robot's perception of this movement will depend on how the robot itself is moving. If the robot is still, it will perceive the walk away movement as it is, but if it is moving towards the person at high speed, it may perceive the person as approaching it. To avoid this ambiguity, we need to know how the human is moving with respect to the robot using an absolute point of reference. Let $\mathbf{p_h^r(t)} \in \mathbb{R}^3$ be the position of the human with respect to the robot at time t, as perceived from the Kinect sensor. Then, let $\mathbf{p_r^m(t)} \in \mathbb{R^3}$ be the position of the robot with respect to its starting point at time t, and $\mathbf{R_r^m(t)} \in \mathbb{R}^{3\times3}$ be the rotation matrix describing the orientation of the robot with respect to the starting point. It is possible to compute the position $\mathbf{p_h^m(t)}$ of the human with respect to the starting point at time t as follows:

$$\left[\mathbf{p_h^m(t)}\ 1\right] = \begin{bmatrix} \mathbf{R_r^m(t)} & \mathbf{p_r^m(t)}^T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p_h^r(t)}^T \\ 1 \end{bmatrix}. \tag{2}$$

At this point, we compute the velocity vector between pairs of successive frames as follows:

$$\mathbf{x_d(t)} = \frac{\mathbf{p_h^m(t+1)} - \mathbf{p_h^m(t)}}{||\mathbf{p_h^m(t+1)} - \mathbf{p_h^m(t)}||}. \tag{3}$$

This quantity represents the real direction in which the human moves with respect to the robot. Since we do not need the last coordinate, which is always 0 (the robot and the person move on a plane), $\mathbf{x_d(t)} \in \mathbb{R}^2$.

Human-Robot Distance Features : Different activities present similar visual and motion properties, but may be distinguished on the basis of where the person is with respect to the robot. For instance, some humans tried to block the robot standing in front of it; their pose, however, is very similar to the pose of those persons that ignore the robot and stand at a certain distance from it. Therefore, we incorporate the distance between the human and the robot retrieved from the Kinect sensor for each frame, taking the hip joint as the point of reference. The human-robot distance descriptor belongs to $\mathbb{R}$.

Feature Quantization : The feature vectors that have been extracted for each frame (or from each pair of successive frames, in case of the Human-Robot Velocity feature vector) are quantized using k-means and represented using Bag Of Words (BOW), so that they generate a single feature vector for each video. A different dictionary is built for each descriptor. We then concatenate the feature vectors in a single vector, obtaining the final descriptor for each video, which belongs to $\mathbb{R}^{\sum_{s=1}^{n} k_s}$, where k_s is the size of the dictionary for the s-th descriptor.

Table 2: Comparison among different features and their combination

Method	Visual only	Visual + HR Velocity + HR Distance	Visual + HR Velocity + HR Distance + Robot Pose
COV [6]	0.3287	0.4397	0.4642
HOJ3D [16]	0.5135	0.6327	0.6507
HODV [3]	0.6242	0.6493	0.6605
PRM	0.5474	0.6597	0.6716
HON4D [11]	**0.7558**	**0.7629**	**0.7642**

5 Experimental Results

This section presents a comprehensive evaluation of feature descriptors and their combinations on our dataset for the activity recognition task. We present results using non-linear SVM with χ^2 kernel, since other kernels (e.g., linear, Gaussian, polynomial and intersection) and other classifiers (e.g., Naive Bayes, Random Forests) achieved comparable or worse results. We perform a stratified 6-fold cross validation, and we repeat the procedure 10 times, to take into account the randomness of the dictionary learning stage. As we anticipated, the dataset is very unbalanced with respect to the activity labels (i.e., some activities are much more frequent than others), thus the recognition accuracy is not a good measure to judge classification performance. Instead, we report the Cohen's kappa coefficient [4], which compares the classifier accuracy against chance accuracy:

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{4}$$

where $Pr(a)$ is the probability of correct classification by the classifier, and $Pr(e)$ is the probability of correct classification by chance.

Visual Features Comparison We assessed the 5 visual descriptors listed in Sec. 4, and we used default parameters to compute all of them. The descriptors reported in [11], [6] and [3] do not need a dictionary learning stage as they already represent the entire video; they belong to $\mathbb{R}^{22680}$, $\mathbb{R}^{1953}$ and $\mathbb{R}^{567}$ respectively. For the other two (PRM and HOJ3D [16]), we set the number of dictionary atoms to 300. The second column of Table 2 reports the results achieved by the different visual descriptors. Notably, the only depth-based descriptor that we have tested, HON4D [11], gets the highest kappa coefficient. In this specific case, where the Kinect is moving continuously, the joint estimation procedure is probably not as reliable as in situations where the Kinect is stable, while the depth images are probably not as affected by the robot's movements as the joint estimation algorithm. Hence, this result may be due to the fact that HON4D is computed on the raw depth images, and does not use joints at all.

Table 3: Precision, Recall and F-1 score of each activity class

Activity	Num Samples	HON4D Precision	HON4D Recall	HON4D F-1	PRM Precision	PRM Recall	PRM F-1
Picture	6	–	0	–	–	0	–
Wave	12	–	0	–	–	0	–
False	608	0.8845	0.9645	0.9227	0.8322	0.9378	0.8818
Block	23	0.7273	0.3130	0.4377	0.5167	0.1348	0.2138
Pass by	153	0.7993	0.8641	0.8304	0.7318	0.8510	0.7869
Walk away	68	0.9394	0.8662	0.9013	0.8652	0.8588	0.8620
Approach	33	0.5970	0.3636	0.4520	0.4817	0.2394	0.3198
Sit	150	0.8483	0.8273	0.8377	0.8196	0.7480	0.7822
Stand	106	0.6433	0.6840	0.6630	0.4875	0.4425	0.4639
Side pass	45	0.7817	0.3978	0.5272	0.6036	0.2267	0.3296

Results with Robot-Centric Descriptors We hypothesize that, in this setting, robot-centric descriptors are useful to improve the performance of visual descriptors. To evaluate this hypothesis, we concatenate the robot-centric descriptors described in Sec. 4 with visual features. Table 2 reports the results obtained by this concatenation in the third column: when robot-centric descriptors are concatenated with visual features, the kappa statistics improves consistently. The fourth column of Table 2 shows the classification rate as the robot's pose in the map is concatenated with the distance and velocity robot-centric features. Notably, making use of the robot's position increases the kappa rate even further. This may be because some activities are more likely to occur in certain regions of the map than at other locations.

Finally, Table 3 provides precision, recall and F1-score of each class using the two best combination of descriptors (HON4D + robot-centric descriptors, and PRM + robot-centric descriptors). Even though HON4D performs significantly better than PRM, it is unable to correctly classify the activities *picture* and *wave*, for which we get 0 true positives and 0 false positives, therefore the symbol "–" in the table. This is probably due to the fact that those are the classes with the smallest number of examples (6 and 12 respectively). For the same reason, the precision and recall on the actions with many samples are relatively high, while those on the actions with a few samples are low. This suggests that if the dataset was more balanced, the results would be more homogeneous. However, the fact that the dataset is unbalanced is one of the natural effects derived from recording activities in the wild. Therefore, learning activities when the number of samples per class differs a lot is one of the challenges of our dataset.

6 Conclusion

This paper considers a new, realistic problem in the field of robot-centric activity recognition: classifying spontaneous activities from a mobile robot's perspective. Unlike previous works, activities are not specified beforehand, and humans are

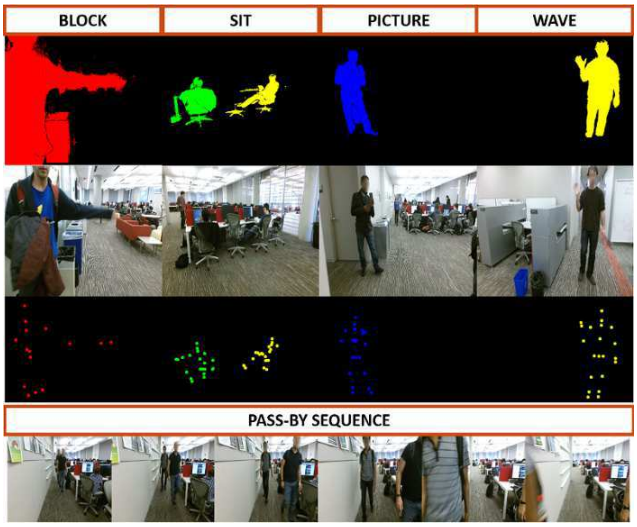

Fig. 2: Examples extracted from our newly recorded dataset. Top: shots of stationary activities. Bottom: action **pass by**. The robot moves forward and then turns.

not asked to perform them. Instead, an autonomous, mobile robot moved around in a building full of people, and recorded their spontaneous behaviors. The robot was left to act alone, therefore the persons who encountered it were not influenced by our presence. All the recorded data was successively analyzed, and the activity classes were determined from the observed videos. To the best of our knowledge, there is no dataset in the literature like the one we are proposing. We plan to release it upon publication, as we expect it to be useful to the community. To obtain satisfactory results on this data, visual features were concatenated with supplementary information directly related to the robot's movements. We showed experimentally that these descriptors consistently improve the results obtained using only visual features. We plan to use the new dataset as a platform to test various learning tasks, different learning algorithms, and multiple combinations of features.

Future work includes using the activity recognition system for 'activity-aware' navigation. For instance, when the robot recognizes that someone is taking a picture of it, it stops and waits until the activity is finished. We also plan to use multiple labels, since sometimes a certain action cannot be described by a single one. Finally, an important future direction is analyzing 'two-way' interactions, during which the robot reacts back to the human.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) 43(3), 16 (2011)

2. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence (2001)
3. Chrungoo, A., Manimaran, S., Ravindran, B.: Activity recognition for natural human robot interaction. In: Social Robotics, pp. 84–94. Springer (2014)
4. Cohen, J.: A coefficient of agreement for nominal scales. Educaional and Psychological Measurement 20(1), 37–46 (1960)
5. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1226–1233. IEEE (2012)
6. Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M.: Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. International Joint Conference on Artificial Intelligence (IJCAI) (2013)
7. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3241–3248. IEEE (2011)
8. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2010)
9. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
10. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J.K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
11. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
12. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: ROS: an open-source Robot Operating System. In: ICRA Workshop on Open Source Software. p. 5 (2009)
13. Rodriguez, M., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
14. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
15. Ryoo, M., Fuchs, T.J., Xia, L., Aggarwal, J., Matthies, L.: Robot-centric activity prediction from first-person videos: What will they do to me. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 295–302. ACM (2015)
16. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3d joints. IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2011)
17. Xia, L., Gori, I., Aggarwal, J.K., Ryoo, M.S.: Robot-centric activity recognition from first-person rgb-d videos. In: IEEE Winter Conference on Applications of Computer Vision (2015)
18. Zhang, S., Sridharan, M., Gelfond, M., Wyatt, J.: Towards an architecture for knowledge representation and reasoning in robotics. In: Social Robotics, pp. 400–410. Springer (2014)