Activity Recognition by Learning from Human and Object Attributes

Brian Reily, Qingzhao Zhu, and Hao Zhang

***Abstract*— Human-robot teaming is crucial to the success of many real-world applications, from search and rescue to home assistance. A robot must be capable of quickly and accurately recognizing activities of its human teammate. In this paper, we develop a new approach to human activity recognition, through simultaneously learning from observations of the teammate and of the attributes of objects involved in the activity. We propose to formulate activity recognition as a joint optimization problem that adopts structured sparsity to identify discriminative body parts and object attributes, and utilizes a regression-like loss function to integrate teammate and object cues to perform real-time activity recognition. To assess our approach, we perform preliminary experiments on a physical robot in a practical home assistance scenario.**

I. Introduction

Effective human-robot teaming is critical for the success of applications that require humans and robots to work together [1]. Human-robot teaming often require that robots be able to understand activities of human teammates with no explicit commands, with the objective to offer proactive assistance, without cognitively burdening human teammates [2]. Human activity recognition in the real world is a difficult problem, complicated by variations in human appearance and motions, and by technical challenges, such as changes in illumination or occlusions. Given the challenges, it is critical to obtain as much information from the scene as possible. This includes the human, as poses and movements are indicative of human activities, as well as further context from the objects in the scene or objects that the human is interacting with, which can provide distinctions between activities (e.g., Figure 1).

In this workshop paper, we introduce a new approach to human activity recognition based on learning from teammate features and object attributes. We formulate human activity recognition as a regression-like optimization problem, and design structured norms as regularization terms to promote sparsity and identify both discriminative skeletal joints and object attributes. This formulation is inspired by the fact that many human activities rely solely on a subset of joints (e.g., a waving activity uses only joints in the arm, not in the legs), or can be identified based upon context of objects in the scene (e.g., reading a newspaper or typing on a laptop at a table appear very similar if only the human pose is considered).

II. Approach

The proposed approach begins with a set of data instances $\mathbf{X} = \{\mathbf{T}, \mathbf{O}\}$, consisting of paired observations of a human teammate and observations of objects. $\mathbf{T} \in \mathbb{R}^{d_T \times N}$ denotes the matrix of observations of the teammate. Observations of the objects are encoded in the matrix $\mathbf{O} \in \mathbb{R}^{d_O \times N}$. Then, our objective is to assign each unknown data instance in $\mathbf{X}$ to one of C behavior categories, based on these observations of the teammate and objects in the scene. Behavior category labels for each data instance are denoted in the category indicator matrix $\mathbf{Y} = [\mathbf{y}^1; \ldots; \mathbf{y}^N] \in \mathbb{R}^{N \times C}$.

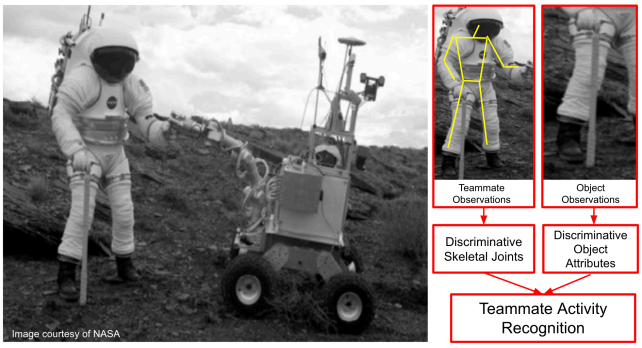


Fig. 1. A motivating example of integrating observations of the teammate and of the objects involved in the activity to understand human behaviors. By observing the human teammate and objects to recognize human activities, an autonomous robot has the potential to interact proactively without requiring direct commands from the human teammate.

We formulate human activity recognition based upon both skeletal observations and object observations as a regression-like optimization problem:

$$\min_{\mathbf{W},\mathbf{U}} \|\mathbf{T}^\top \mathbf{W} + \mathbf{O}^\top \mathbf{U} - \mathbf{Y}\|_F^2 + \lambda_1 \|\mathbf{W}\|_S + \lambda_2 \|\mathbf{U}\|_A \quad (1)$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C] \in \mathbb{R}^{d_T \times C}$ denotes a weight matrix indicating the importance of $\mathbf{T}$ to the behavior category labels, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_C] \in \mathbb{R}^{d_O \times C}$ is a weight matrix doing the same for $\mathbf{O}$, and $\lambda_{1,2}$ are trade-off hyperparameters. $\mathbf{w}_c \in \mathbb{R}^{d_T}$ is the weights of joints with respect to c-th category, with subsections $\mathbf{w}_c^j \in \mathbb{R}^{d_T^j}$ representing the weights of the j-th joint to the c-th category. Similarly, $\mathbf{u}_c \in \mathbb{R}^{d_O}$ represents weights of object attributes with respect to c-th category, with subsections $\mathbf{u}_c^{o_m} \in \mathbb{R}^{d_O^m}$ representing weights of the m-th attribute of the o-th object to the c-th category.

To identify discriminative body joints, we design a *skeletal norm* on the joint observation weight matrix $\mathbf{W}$:

$$\|\mathbf{W}\|_S = \sum_{c=1}^{C} \sum_{j=1}^{J} \|\mathbf{w}_c^j\|_2 \quad (2)$$

This norm enforces the ℓ_2-norm within a joint feature and the ℓ_1-norm between joints in order to force sparsity and weight only discriminative joints. In addition, we introduce a new

Brian Reily, Qingzhao Zhu and Hao Zhang are with the Human-Centered Robotics Laboratory at the Colorado School of Mines, Golden, CO, 80401, USA. email: {breily, zhuqingzhao, hzhang}@mines.edu

This work was partially supported by ARL DCIST CRA W911NF-17-2-0181, USAFA FA7000-18-2-0016, and ARO W911NF-17-1-0447.

(a) TurtleBot Platform (b) Scenario Setup

Fig. 2. The setup used to evaluate the home assistance scenario. Figure 2(a) shows the Turtlebot platform. Figure 2(b) illustrates the robot observing the human as an activity is performed.

attribute norm to learn the importance of object attributes of multiple objects in a scene, which is defined as:

$$\|\mathbf{U}\|_A = \sum_{c=1}^{C}\sum_{o=1}^{O}\sum_{m=1}^{M}\|\mathbf{u}_c^{om}\|_2 \tag{3}$$

The attribute norm selects discriminative attribute modalities for each object. Each object present has multiple attributes, and the ℓ_2-norm is applied to enforce similar weights within an attribute modality. The ℓ_1-norm is applied between these attribute modalities to enforce sparsity and force the identification of discriminative attributes.

III. Experimental Results

We implemented our approach on a physical robot in order to validate its performance in a real-world home assistance scenario. We deployed our approach on a Turtlebot robot participating in a home assistance scenario, as depicted in Figure 2. The Turtlebot has an ASUS Xtion Pro color-depth sensor onboard to extract 3D skeleton data and a lightweight netbook for processing.

In this scenario, five activities were defined, including *drinking wine*, *storing food*, *storing dishes*, *pouring wine*, and *eating*. Each activity was performed 20 times. In order to test the effectiveness of our approach in learning simultaneously from observations of the teammate and observations of the objects, these activities were defined to involve similar objects and human poses. For example, both drinking wine and pouring wine involve a glass and a bottle, but drinking wine is performed while sitting down and pouring wine is performed while standing up. Similarly, both eating and drinking wine are behaviors performed by a sitting human teammate, but involve different objects (respectively, a bowl and a spoon versus a wine glass and a bottle).

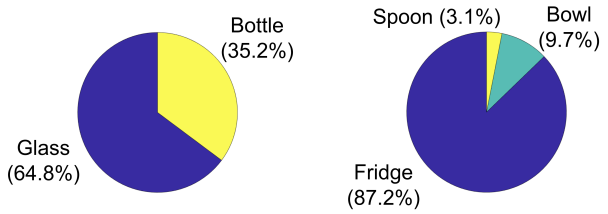
The quantitative experimental results are presented in Table I. We can observe that the proposed approach achieves an overall accuracy of 98.33% in this home assistance scenario. Comparison with baseline real-time approaches is also listed in Table I, which shows that our approach is superior to two standard real-time machine learning methods as baselines. With only the *skeletal norm* or only the *attribute norm*, the proposed approach achieves good accuracy but less than with the full formulation that uses both norms.

TABLE I

Accuracy obtained by our approach in the home assistance scenario and comparison to baseline real-time approaches.

Approach	Accuracy
Support Vector Machine	51.67%
Decision Forest	91.67%
Our Approach (only *skeletal norm*)	95.00%
Our Approach (only *attribute norm*)	96.67%
Our Approach	**98.33%**

In this set of experiments, we also tested our approach with a different set of attributes in order to evaluate its ability to identify discriminative objects. In this setup, each scene has five attribute modalities, where each modality is the probability that an object category appeared in that scene. The 5 object categories used are the *wine bottle*, *glass*, *fridge*, *bowl*, and *spoon*. The probabilities that an object appeared in a scene were obtained from the YOLO object detection, which uses a pre-trained neural network to identify common household objects. For example, for the activity of drinking wine, the probability of a bottle or glass appearing would be close to 1, and close to 0 for the remaining objects.



(a) Drinking Wine (b) Storing Food

Fig. 3. Figure 3(a) illustrates the weights for the drinking wine activity, where the glass and bottle are very important. Figure 3(b) shows the weights for the activity of storing food, where the fridge is the most relevant object.

Using this setup, our approach is able to recognize 96.67% of home activities correctly. Additionally, this setup allowed our approach to identify discriminative objects, as each column of the $\mathbf{U}$ matrix contained only five values, each relating one object to that column's associated human activity. Figure 3 displays two sample columns from the $\mathbf{U}$ weight matrix. In Figure 3(a), we see that the bottle and glass are the only objects receiving weights, as these are very indicative of the drinking wine activity. Similarly in Figure 3(b), we see that the fridge receives nearly 90% of the total column weight, identifying it as being very indicative to recognize the storing food activity.

References

[1] G.-J. M. Kruijff, M. Janíček, S. Keshavdas, B. Larochelle, H. Zender, N. J. Smets, T. Mioch, M. A. Neerincx, J. V. Diggelen, and F. Colas, "Experience in system design for human-robot teaming in urban search and rescue," in *Field and Service Robotics*, pp. 111–125, Springer, 2014.

[2] R. Schulz, P. Kratzer, and M. Toussaint, "Preferred interaction styles for human-robot collaboration vary over tasks with different action types," *Frontiers in Neurorobotics*, vol. 12, p. 36, 2018.