COLORADOSCHOOLOFMINES
engineering the way

Simultaneous Learning from Human Pose and Object Cues for Real-Time Activity Recognition

2020 International Conference on Robotics and Automation

Brian Reily[1], Qingzhao Zhu[1], Christopher Reardon[2], and Hao Zhang[1]
1: Human-Centered Robotics Lab, Colorado School of Mines (http://hcr.mines.edu)
2: U.S. Army Research Laboratory

Overview

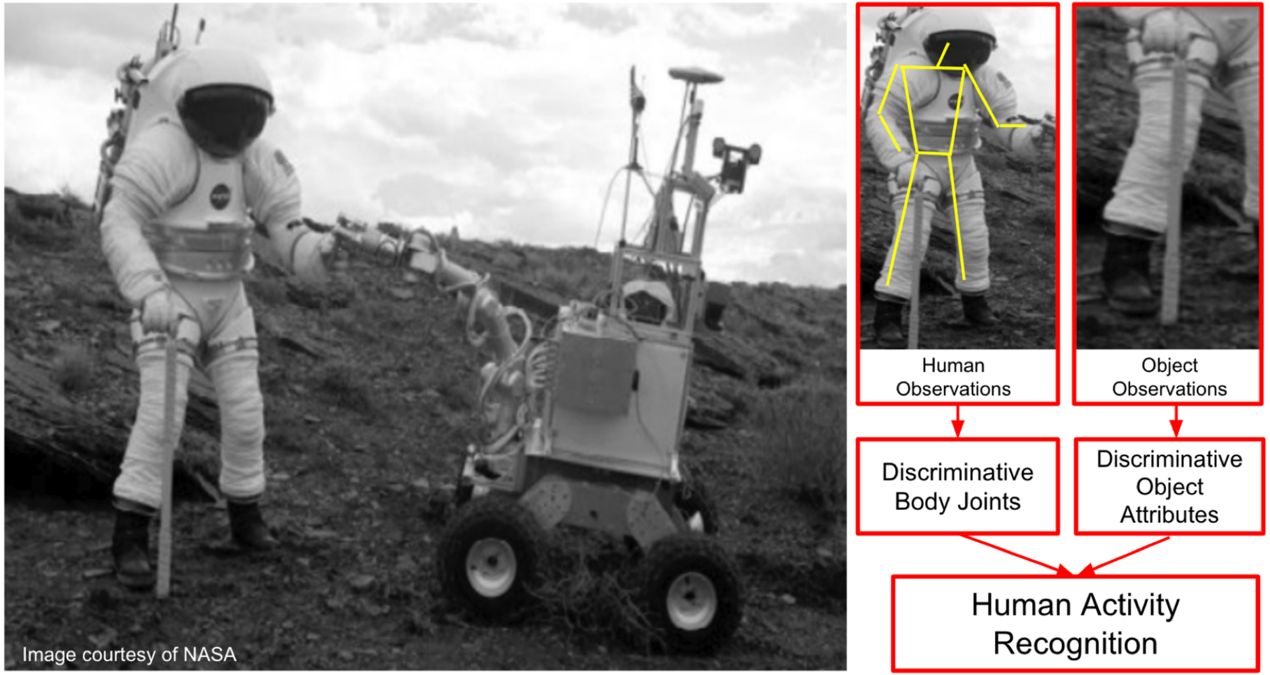


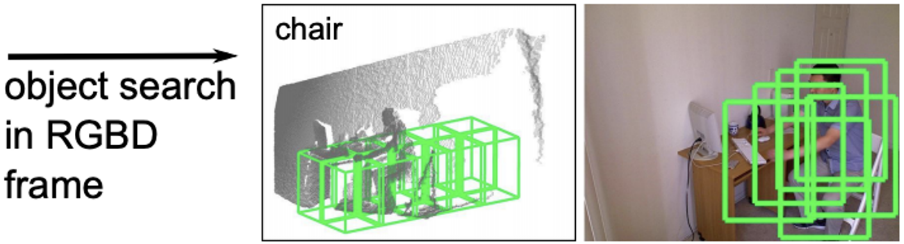
Image courtesy of NASA
Human Observations
Object Observations
Discriminative Body Joints
Discriminative Object Attributes
Human Activity Recognition

Motivation

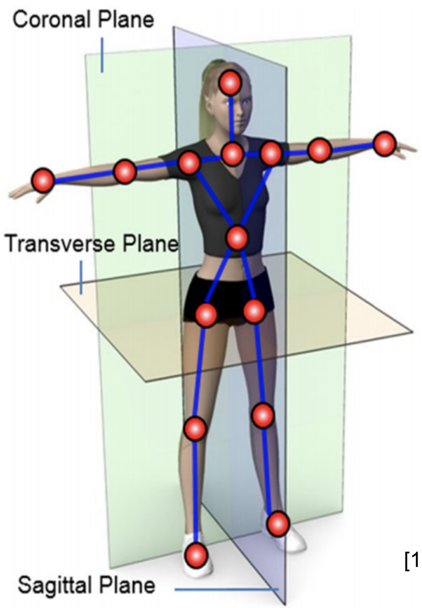
- Humans prefer when robots can interact implicitly without continuous commands - robots need real-time understanding of a human’s actions.
- Real-time activity recognition is difficult - activities can occur indoors or outdoors, at night or during the day; humans can vary widely in appearance (e.g., adults vs. children); etc.
- And so robots must extract as much information as possible from their observations - in this case, objects that their teammate is interacting with.

Existing Research

- Most limited to using information *only* from the pose/body of the human teammate.
- While a small number do utilize object information, they rely on a predetermined sets of possible objects.



[2]

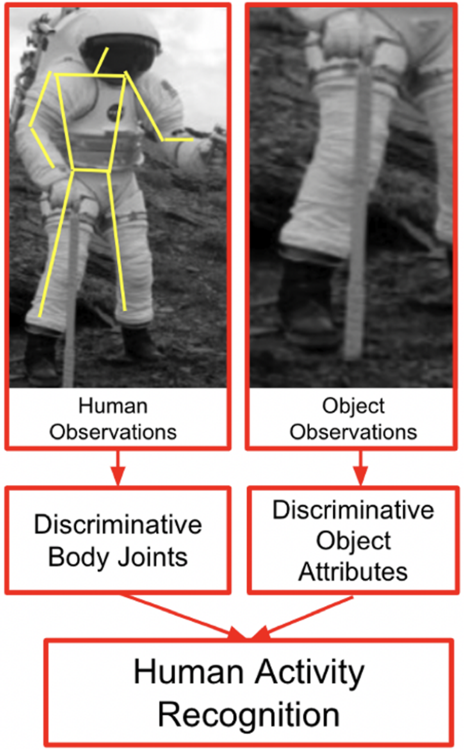


[1]

[1]: Brian Reily, Fei Han, Lynne E Parker, and Hao Zhang. "Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction." In Autonomous Robots, 42(6):1281– 1298, 2018.
[2]: Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu, "Modeling 4d human- object interactions for event and object recognition." In International Conference on Computer Vision (ICCV), 2013.

Our Contribution

- We formulate human activity recognition as simultaneously learning from human and object observations.
- The method identifies both discriminative skeletal joints and discriminative object attributes.



Problem Formulation

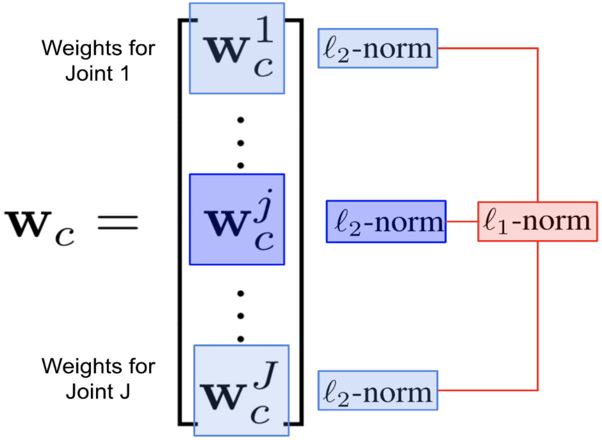
We consider observations of the teammate $\mathbf{T}$, which can represent features such as joint positions, and observations of the objects $\mathbf{O}$, which can represent various attributes such as size, shape, and color.

We define a single loss function that describes the relationship between a linear combination of $\mathbf{T}$ and $\mathbf{O}$ with ground truth activity labels in $\mathbf{Y}$.

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

Problem Formulation

We introduce a sparsity inducing norm to identify discriminative joints, termed the *skeletal* norm.



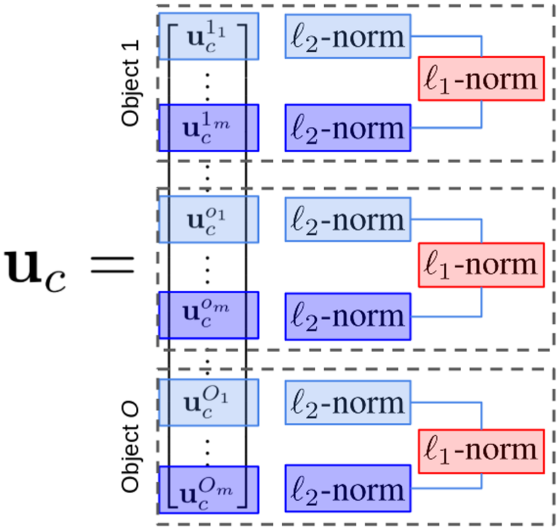
$$\|\mathbf{W}\|_S = \sum_{c=1}^{C} \sum_{j=1}^{J} \|\mathbf{w}_c^j\|_2$$

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

Problem Formulation

We introduce a second sparsity inducing norm to identify discriminative objects and object attributes, termed the *attribute* norm.

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



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

Problem Formulation

We solve this using an iterative algorithm, updating **W** and **U** at each step until convergence.

We can then use the optimal **W** and **U** to classify new observations.

Algorithm 1: An iterative algorithm to solve the formulated optimization problem in Eq. (5).

Input : $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_N] \in \mathbb{R}^{d_T \times N}$, $\mathbf{O} = [\mathbf{o}_1, \dots, \mathbf{o}_N] \in \mathbb{R}^{d_O \times N}$ and $\mathbf{Y} = [\mathbf{y}^1; \dots; \mathbf{y}^N] \in \mathbb{R}^{N \times C}$.

Output : $\mathbf{W}^* = \mathbf{W}(i) \in \mathbb{R}^{d_T \times C}$ and $\mathbf{U}^* = \mathbf{U}(i) \in \mathbb{R}^{d_O \times C}$.

Let $i = 1$. Initialize $\mathbf{W}$ and $\mathbf{U}$ randomly.
repeat
  Calculate $\mathbf{D}_S^c(i+1)$ for $c \in 1, \dots, C$.
  Calculate $\mathbf{D}_A^c(i+1)$ for $c \in 1, \dots, C$.
  Calculate $\mathbf{w}_c(i+1)$ via Eq. (9) for each $c \in 1, \dots, C$.
  Calculate $\mathbf{u}_c(i+1)$ via Eq. (11) for each $c \in 1, \dots, C$.
  $i = i + 1$.
until *convergence*;
return $\mathbf{W}^*$ and $\mathbf{U}^*$

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

Problem Formulation

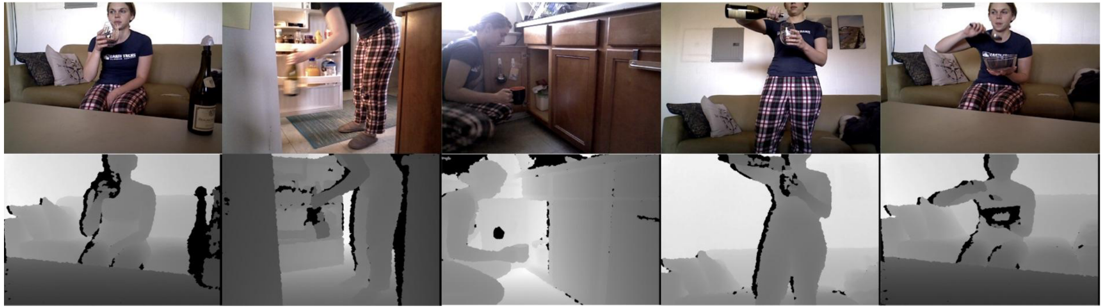
To classify a new scene with observations $\mathbf{t}$ and $\mathbf{o}$, we find the category c that maximizes the category indicator y.

$$y\left(\mathbf{t},\mathbf{o}\right) = \max_{c} \mathbf{t}^{\top}\mathbf{w}_c^* + \mathbf{o}^{\top}\mathbf{u}_c^*$$

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

Case Study Evaluation

- We conducted a study using a Turtlebot robot running a small netbook.
- We recorded 5 different activities involving a common set of objects, with objects appearing in multiple different activities.



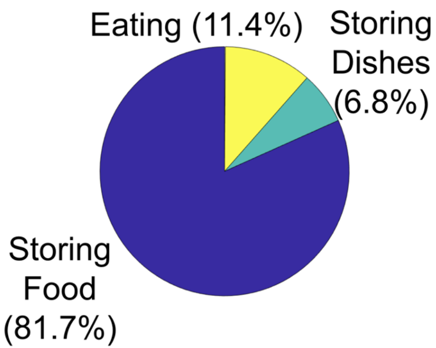
Case Study Evaluation

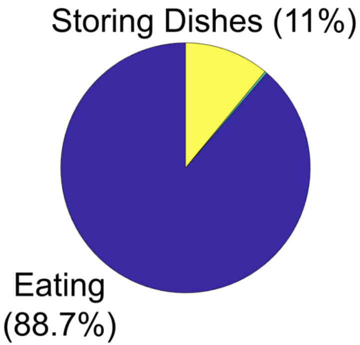
- For this evaluation, we utilized the YOLO object detection system. Each object was described by 5 modalities, each of which is the probability of it being a certain object.
- 5 possible objects were used: glass, bottle, fridge, bowl, and spoon.

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%**

Effects of Introduced Norms

For the *fridge*, our introduced attribute norm identified that it was most closely associated with *storing food*.

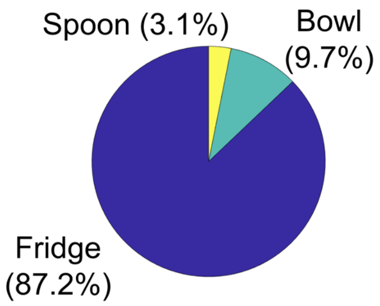


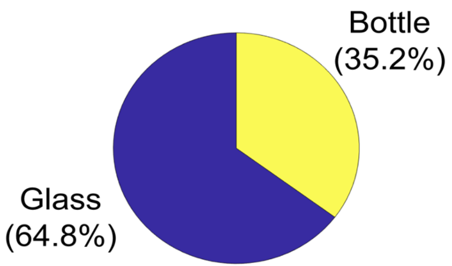


For the *bowl*, this norm associated it most with *eating*, while *storing dishes* is the only other activity associated with it.

Effects of Introduced Norms

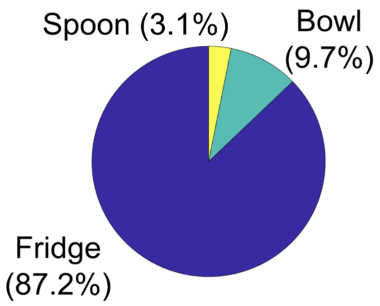
For *storing food*, our introduced attribute norm identified the *fridge* as being very discriminative.



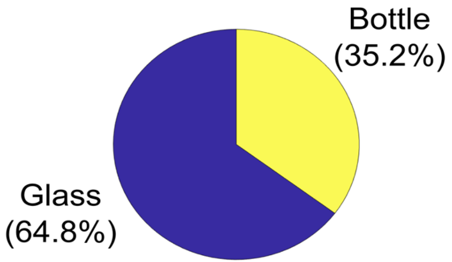


For *drinking wine*, this norm assigned weights to only the *glass* and *bottle*.

Effects of Introduced Norms



For *storing food*, our introduced attribute norm identified the *fridge* as being very discriminative.

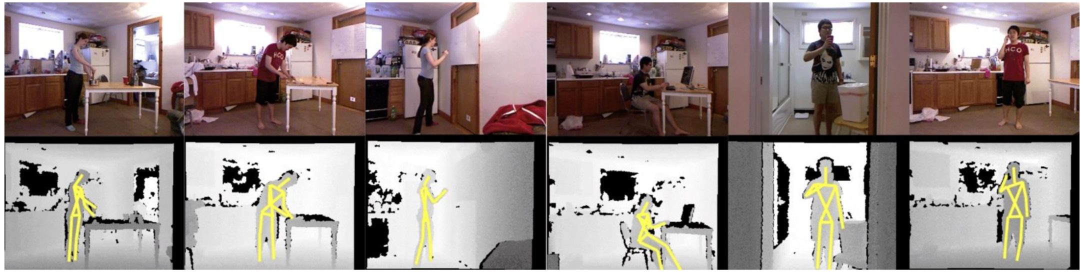


For *drinking wine*, this norm assigned weights to only the *glass* and *bottle*.

This provides *interpretability* that black box methods cannot: if we wonder why the robot teammate thinks we are *storing food*, we can see that this is because it has observed a *fridge* and a *bowl*.

Dataset Evaluation

We evaluated our approach on the Cornell Activity Dataset (CAD-60), testing our full approach as well as variations with only a single introduced norm.



Approach	Accuracy
Our Approach (only *skeletal norm*)	86.86%
Our Approach (only *attribute norm*)	96.18%
Our Approach	**98.11%**

Dataset Evaluation

We evaluated our approach on the MSR Daily Activity 3D dataset, testing our full approach as well as variations with only a single introduced norm.

Approach	Accuracy
Our Approach (only *skeletal norm*)	82.00%
Our Approach (only *attribute norm*)	95.71%
Our Approach	**97.71%**

Thanks!

- We formulate activity recognition as learning simultaneously from observations of the teammate and the objects in a scene.
- Our approach outperforms existing state-of-the-art approaches, with sparsity-inducing norms increasing accuracy and providing explainability about how a robot classifies actions.

Thanks!

- We formulate activity recognition as learning simultaneously from observations of the teammate and the objects in a scene.
- Our approach outperforms existing state-of-the-art approaches, with sparsity-inducing norms increasing accuracy and providing explainability about how a robot classifies actions.

Questions?

Contact Us: hzhang@mines.edu / breily@mines.edu
http://hcr.mines.edu

This work was partially supported by NSF IIS-1942056, ARL DCIST CRA W911NF-17-2-0181, USAFA FA7000-18-2-0016, and ARO W911NF- 17-1-0447.