Zero-Shot Learning for IMU-Based Activity Recognition Using Video Embeddings

Catherine Tong[†*], Jinchen Ge[§*], and Nicholas D. Lane[§]

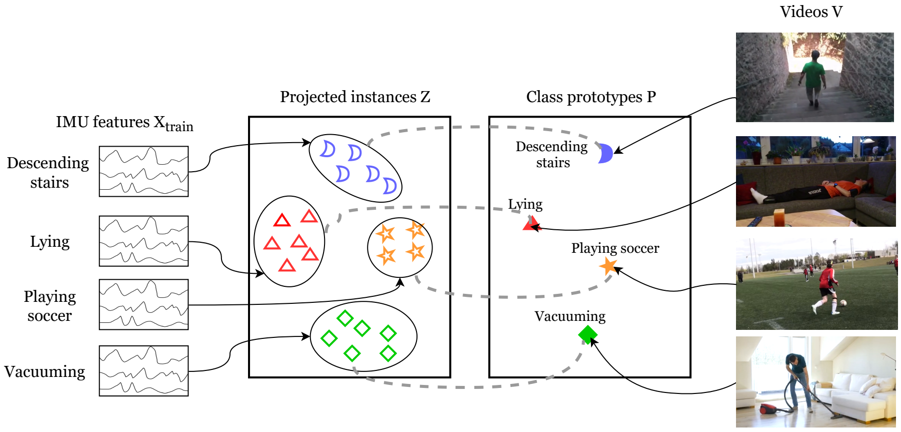
[†]Department of Computer Science, University of Oxford
[§]Department of Computer Science and Technology, University of Cambridge

The Human Activity Recognition (HAR) chain generally precludes the challenging scenario of recognizing new activities that were unseen during training, despite this scenario being a practical and common one as users perform diverse activities at test time. A few prior works have adopted zero-shot learning methods for IMU-based activity recognition, which work by relating seen and unseen classes through an auxiliary semantic space. However, these methods usually rely heavily on a hand-crafted attribute space which is costly to define, or a learnt semantic space based on word embeddings, which lack motion-related information crucial for distinguishing IMU features.

In observing that video-based activity recognition is a well-researched area with ample data samples, and that videos of human activities contain much richer domain-specific information compared to word embeddings, we propose a novel strategy to exploit videos of activities to construct an informative semantic space for zero-shot learning. Specifically, as demonstrated in Figure 1, we first use a pretrained video HAR model, namely I3D [Carreira and Zisserman, 2017], to extract feature representations of videos of activities, which we refer to as *video embeddings*. These rich video embeddings are then used to construct the semantic space that relates seen and unseen classes during zero-shot learning. Not only does this video-based semantic space circumvent the need to manually define attributes, it also manifests the transfer of knowledge from video-based HAR models to an IMU-based HAR problem.

We use three publicly available IMU datasets that cover various human activities or actions for evaluation. The experiment results show that our video semantic space is consistently superior to word-embedding-based methods, and comparable to and sometimes even better than manual-attribute-based methods. Moreover, our video semantic space also demonstrates an additional desirable feature of scalability, as recognition performance is seen to scale with the amount of data used. More generally, our results indicate that exploiting information from the video domain for IMU-based tasks is a promising direction, with tangible returns in a zero-shot learning scenario.

Figure 1: Overview of our projection-based method for zero-shot learning. The right side illustrates the construction of a video semantic space, which is done by passing video data through a pretrained I3D model to compute class prototypes. The left side shows the projection of IMU features into the video semantic space, done using a 4-layer MLP.



References

[Carreira and Zisserman, 2017] Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

[*]Both authors contributed equally to this research.