Pose from Action: Unsupervised Learning of Pose Features based on Motion

Senthil Purushwalkam, Abhinav Gupta

Robotics Institute, Carnegie Mellon University

Motivation

- Human Actions are an ordered sequence of poses
- Can we leverage human action videos to learn a pose encoding representation?

A representation in which poses cluster together should be useful for Pose Estimation and Action Recognition
- Given two poses, it should be possible to predict the motion between them

Surrogate Task

- Given:
  - Two appearance representations A and A' (pose)
  - One motion representation T (transformation of pose)

- Predict:
  - If the transformation T could cause the change A→A'

Data and Implementation

- We use videos from UCF101, HMDB51 and ACT video datasets
- Sample two frames separated by Δn (=12) frames and extract optical flow for the Δn frames
- Use the VGG-M-2048 architecture for all CNNs

Experiments

- Nearest neighbor in the FC6 feature space of the Appearance ConvNet

Results

- Action Recognition:
  - Classification Accuracy

- Pose Estimation:
  - Strict PCP evaluation

- Static Image Action Recognition:
  - Classification Accuracies

Paper available on arxiv at:
https://arxiv.org/abs/1609.05420