Two Stream Self-Supervised Learning for Action Recognition
Ahmed Taha  Moustafa Meshry  Xitong Yang  Yi-Ting Chen  Larry Davis

Goals
• Tangle spatial and temporal representation for action recognition.
• Boost performance on small and imbalanced video datasets
• HMDB and UCF 130 samples per class
• Honda Driving Dataset is imbalanced

Approach
1. Sequence Verification  [Shuffle & learn, O3N, OPN]
2. Spatial Temporal Verification

Architecture
Given a tuple of a RGB frame and a stack of difference (SOD), the network reasons about frame ordering and spatio-temporal correspondence.


Lack of big annotated datasets can limit supervised learning techniques for many domains like the medical, and autonomous car driving. For these applications, there exists massive data. However, annotating such datasets can be costly for various reasons; it can be time-consuming as in semantic segmentation or expensive due to medical field expertise requirement.

Motion Tower
Learned Temporal Visualization
Temporal filter visualization is uninformative. Input motion reconstruction is utilized for qualitative evaluation. Self-supervised reconstructions are less noisy compared to supervised learned features.

Our approach supports various motion encodings like stack of difference, dynamic images or optical flow.

Temporal representation learning can benefit from ImageNet Spatial representation.