PAN: Towards Fast Action Recognition via Learning Persistence of Appearance

Can Zhang, Yuexian Zou, Senior Member, IEEE, Guang Chen, and Lei Gan

Introduction

The main contribution of this paper is the replacement of Optical Flow as a motion cue in Action recognition tasks. The authors made this possible by designing a novel motion cue called Persistence of Appearance, based on the observation that small displacements of motion boundaries are the most critical ingredients for distinguishing actions. They argue that the PA module focuses more on distilling the motion information at boundaries, in contrast to optical flow.

As a simple summary, a simple difference in feature space (after a convolutional layer) is done between consecutive frames, in order to extract motion cues.
They also design a module that is able to model the long-term dynamics that is called Various-timescale Aggregation Pool.

Persistence of Appearance

As explained before, the PA module is a difference between feature maps coming from consecutive frames. The intuition comes from the known fact that the first layers of a network are able to extract cues such as edges, so a difference between them would result in the displacements of motion boundaries. The difference is made channel with channel, and the result is summed up, resulting in a “motion map” formed of only one channel.
Following experiments, the paper states that the best architecture is using only one convolutional layer with 7x7 kernel size, stride=1 and padding=3, so that the spatial resolutions of the obtained feature maps are not reduced.

Persistence of Appearance Network

The paper proposes two network architectures, one that treats the PA input as a separate branch of the network, and one that takes the PA maps and concatenates them with the spatial information, creating only one network.

PAN Full:

The authors describe it as separate and accurate, composed of two branches, capturing the spatial and temporal features separately. The spatial branch takes N frames, selected using the TSN method and processes them through the selected backbone network. Then, these features are aggregated as video-level features using the VAP model:

$$\boldsymbol{y_s} = \mathcal{H}_{VAP}(\mathcal{H}_B(\{\boldsymbol{I_t}\}_{t=1}^{N}))$$

The second branch takes as input N stacks of m adjacent frames and transforms them to motion cues. Then, the motion features are fed to the backbone network and to the VAP module. In the end, the two results are merged using the score fusion strategy (calculating the weighted average of the scores).

PAN Lite:

It is described as unified and light-weighted. This architecture, as described previously, stacks the two input modalities tighter and feeds them through the backbone, allowing it to decide how to use the spatiotemporal information. The stacking is done through a channel concatenation operation:

$$\boldsymbol{y_{st}} = \mathcal{H}_{VAP}(\mathcal{H}_B(\boldsymbol{I_t}, e(\mathcal{H}_{PA}(\{\boldsymbol{I_{t[m]}}\}_{t=1}^{N}))))$$

Various-timescale Aggregation Pooling (VAP)

This module is composed of two main steps:

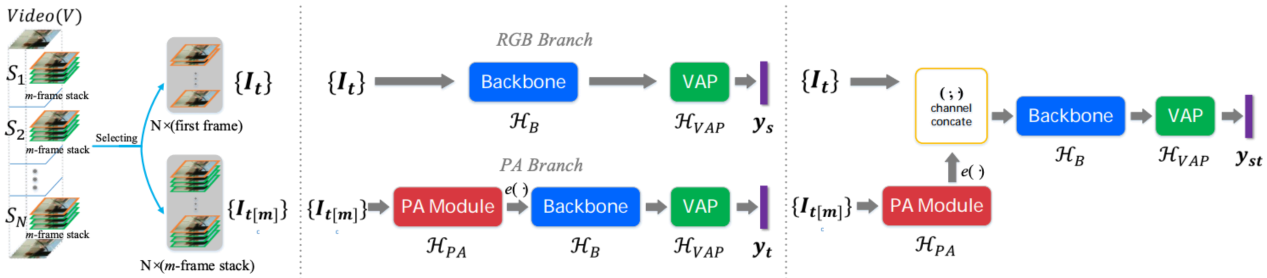
- Specific-timescale Pooling:

In this step, a dilated max pooling operation is conducted over the time dimension, where the dilation controls the spacing between the kernel points.

- Various-timescale Aggregation

The basic idea is to fuse temporal information at each timescale by weighted timescale-wise aggregation. First, after obtaining a vector v that represents the total pooled features, they shrink global spatial semantics in each feature into a temporal descriptor that contains the temporal statistics. The result of this shrinking is fed to a networks composed of two FC layers and a softmax function used for weight perception. The final global video-level representation of the proposed module is obtained by firstly rescaling the feature vector and then aggregate the recalibrated features along the temporal dimension.

Architecture



(A) Sampling Strategy (B) PANFull Architecture (C) PANLite Architecture