Hierarchical Dynamic Parsing and Encoding for Action Recognition

Bing Su[1], Jiahuan Zhou[2], Xiaoqing Ding[3], Hao Wang[1], Ying Wu[2]

[1]Institute of Software Chinese Academy of Sciences. [2]Northwestern University. [3]Tsinghua University.

Motivation

- **Representation matters**:
  - The performance of action recognition methods depends heavily on the representation of video data.
- **Dynamics are inherent**:
  - Dynamics characterize the inherent temporal dependencies of actions.
  - Existing methods either cannot directly lead to vector representations with a fixed dimension, or treat the changes of all successive frames equally.
- **Dynamics are not uniform**:
  - The dynamic behind an action is time-varying, non-stationary and has some intuitive rhythms or regularities.
  - Humans can recognize an action from some ordered key frames or poses. These key poses segment the whole action into different divisions, and each division consists of the frames related to a key pose.

  - Therefore, the dynamics of an action can also be viewed as a hierarchy. The dynamics within each stage are relatively stable, and the dynamics of the sequence of the stages represent the essential evolution of the action.

Dynamic Parsing

- **Learn the parse of an action sequence from the sequence itself**
  - Capture the temporal structures w.r.t. relatively-uniform local dynamics.
- **Unsupervised Temporal Clustering**
  - Represent the video with a sequence of frame-wide features
    $$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$$
  - A partition of $\boldsymbol{X}$ is defined by a segmentation path $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_L]$ $\mathbf{p}_t = [s_t, e_t]^T$ denotes the start and end frames of the t-th division; f controls the maximum extent of wrapping
  - Define the essential sequence of $\boldsymbol{X}$: $\mathbf{U} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_L]$ which can be viewed as the sequence of key poses.

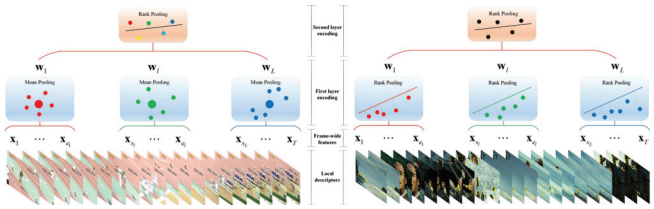
- **Objective:**
$$\min_{\mathbf{P},\mathbf{U}} \sum_{j=1}^{L} \sum_{i=s_j}^{e_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2$$
- **Optimization:**
  - Given **P**, optimize **U**: compute the means of all the divisions
$$\boldsymbol{\mu}_j = \frac{1}{l_j} \sum_{k=s_j}^{e_j} \mathbf{x}_k, j = 1, \cdots, L$$
  - Given **U**, optimize **P**: dynamic temporal wrapping

repeat iteratively

$$d(i,j,l) = \begin{cases} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2, l = 1, i = j = 1 \\ \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2 + \min_{k=1}^{f \cdot l_{ave}} d(i-1, j-1, k), l = 1 \\ \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2 + d(i-1, j, l-1), l \le f \cdot l_{ave} \\ Inf, otherwise \end{cases}$$

Hierarchical Dynamic Encoding

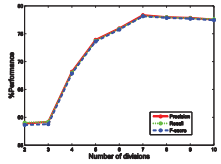
- **Incorporate the dynamics in the hierarchy of two layers into a joint representation**



- **The first layer modeling**
  - The action sequence is parsed into several smooth-changing divisions corresponding to different key poses or temporal structures by unsupervised temporal clustering
  - The dynamics within each stage are encoded by mean-pooling (M-HDPE) or rank pooling (R-HDPE).
- **The second layer modeling**
  - The dynamics of the ordered representations extracted from the previous layer is encoded again by rank pooling to form the overall representation

Experimental Results

- **Influence of parameters**
  - At first performances improve with the increase of number of divisions, because more temporal structures can be captured.
  - Performances stop increasing when L is large enough, indicating redundant divisions exist.



- **Comparison of pooling in the first layer**
  - M-HDPE outperforms R-HDPE on ChaLearn dataset. For fine-grained actions, since the evolution within each division is quite uniform, the local appearance information is enhanced by mean-pooling.
  - R-HDPE outperforms M-HDPE on Hollywood2 dataset. For complex actions, the complex dynamics within divisions contain important discriminative information of the action and hence cannot be eliminated.
- **Comparison with state-of-the-arts**
  - If *L=1*, HDPE boils down to "Local + BoW"; If *L=T*, HDPE boils down to "rank pooling".
  - When the same frame-wide features are used, HDPE outperforms both methods on all the three datasets. The superior performances come from the hierarchical parsing and modeling.
  - On the Olympic and Hollywood2 datasets, the state-of-the-arts are achieved by using Fisher Vector based features and data augmentation. Applying these techniques can also benefit HDPE.

Method	Precision	Recall	F-score
Wu et al. [31]	59.9	59.3	59.6
Yao et al. [32]	-	-	56.0
Pfister et al. [33]	61.2	62.3	61.7
Fernando et al. [5]	75.3	75.1	75.2
Rank pooling [5]	74.0	73.8	73.9
HDPE	**78.34**	**78.18**	**78.15**

Method	Olympic Sports
Brendel et al. [22]	77.3
Gaidon et al. [34]	82.7
Jain et al. [10]	83.2
Wang et al. [6]	**91.1**
Local+BoW [6]	83.3
HDPE	87.66
HDPE+Rank pooling	**89.09**

Method	Hollywood2
Jain et al. [10]	62.5
Wang et al. [6]	64.3
Hoai et al. [35]	73.6
Fernando et al. [5]	**73.7**
Local+BoW [6]	62.2
Rank pooling+BoW [5]*	62.19
HDPE	**63.51**

Method	Accuracy
Laptev et al. [29]	62.0
Niebles et al. [28]	72.1
Tang et al. [36]	66.8
Wang et al. [20]	73.8
HDPE	**81.34**
HDPE+Rank Pooling+Local	**83.58**

Contribution

- The proposed HDPE is a new unsupervised representation learning method. It hierarchically abstracts the prominent dynamic and generates a representation that is robust to speed and local variations.
- We propose an unsupervised method for temporal clustering to achieve efficient dynamic parsing.