Mining Motion Atoms and Phrases for Complex Action Recognition

LiMin Wang[1,2], Yu Qiao[2], and Xiaoou Tang[1,2]

[1]Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong
[2]Shenzhen Institutes of Advanced Technology, CAS, China

Introduction

- **Goal**: Mine mid-level motion units (i.e. motion atoms and phrases) for representing and classifying complex actions, such as sports actions.
- **Key insights**:

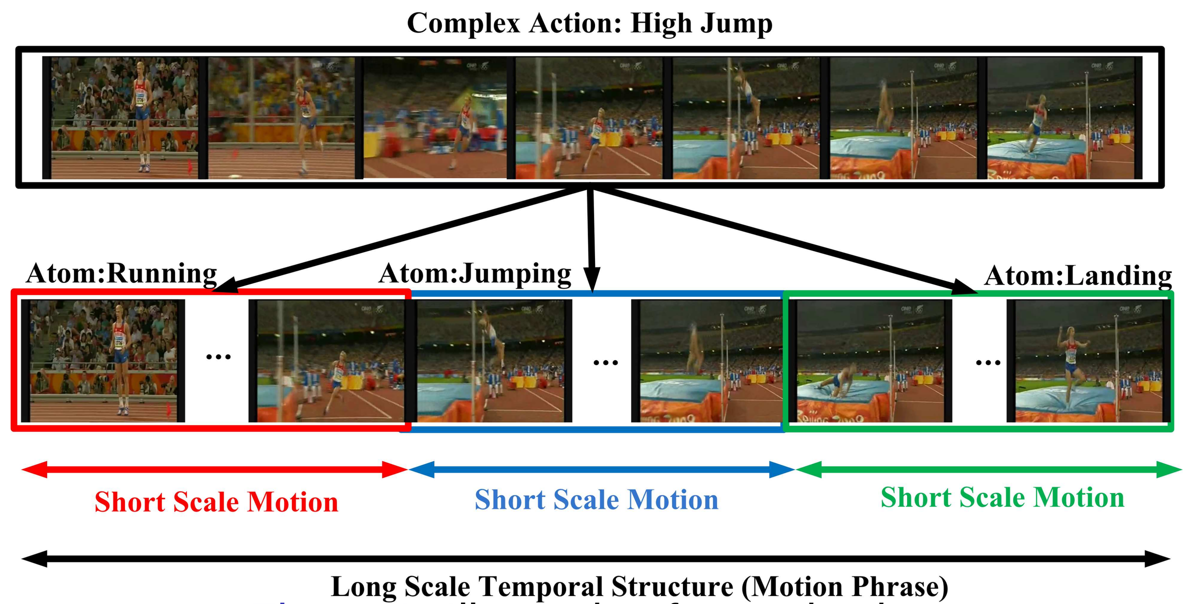

Figure 1: Illustration for motivation.

  - From a long temporal scale, a complex action is a sequence of atomic motions, and there is temporal structure among them.
  - From a short temporal scale, each motion atom corresponds to a certain and simple motion pattern, and can be shared by different complex action classes.
- **Our method**:
  - Unsupervised discovery of a set of motion atoms for the whole dataset.
  - Supervised mining discriminative motion phrases as temporal composite of motion atoms for each action class.
- **Compared with other works**:
  - Existing works focus on using Sequential State Model to capture the temporal structure such as HMMs, HCRFs, and DBNs. Others use latent variable to model the temporal decomposition and formulate the problem by Latent SVM.
  - Motion atom and phrase provides a mid-level temporal representation for action video, which encodes the motion, appearance, and temporal structure. Our representation is flexible with classifiers and easily combined with other methods.

Discovery of Motion Atoms

- **Iterative discriminative clustering processing:**
  - *Step0:* We divide each video clip into k segments and cluster these segments.
  - *Step1:* For each cluster, we train a kernel SVM using segments belonging to it as positive examples and hard negative examples from other clusters.
  - *Step2:* Using the discriminative classifier, we update the cluster by the segments with top scores.
- **Some details:**
  - For each segment, we extract dense trajectories with four types of descriptors as HOG, HOF, MBHX, and MBHY, and then use Bag of Visual Words to obtain a global representation.
  - For clustering algorithm, we choose the Affinity Propagation algorithm with segment similarity defined as:

$$\mathbf{Sim}(S_i, S_j) = \sum_{m=1}^{4} \exp(-\mathcal{D}(\mathbf{h}_i^m, \mathbf{h}_j^m)) \qquad \mathcal{D}(\mathbf{h}_i^m, \mathbf{h}_j^m) = \frac{1}{2M_m} \sum_{k=1}^{K} \frac{(\mathbf{h}_{i,k}^m - \mathbf{h}_{j,k}^m)^2}{\mathbf{h}_{i,k}^m + \mathbf{h}_{j,k}^m}.$$

Mining Motion Phrases

- **Definition:** motion phrase is a temporal composite of multiple motion atoms organized with an *AND-OR structure*, whose size equals to the number of OR operations.

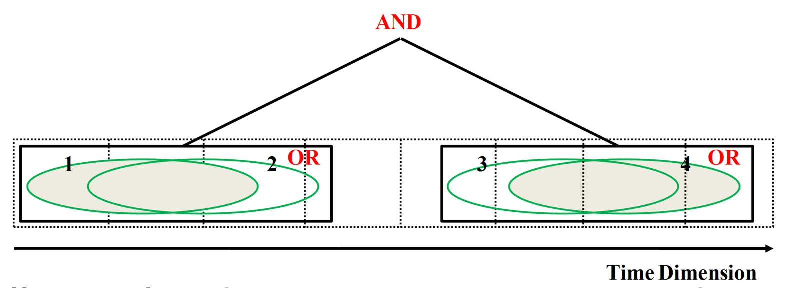

Figure 2: Illustration for AND-OR structure of motion phrase.

  - Atom unit $\Pi = (A, t, \sigma)$: $v(V, \Pi) = \max_{t' \in \Omega(t)} \mathbf{Score}(\Phi(V, t'), A) \cdot \mathcal{N}(t' | t, \sigma)$.
  - Atom phrase $P = \{OR_i\}$: $r(V, P) = \min_{OR_i \in P} \max_{\Pi_j \in OR_i} v(V, \Pi_j)$.
- **Three key properties:**
  - Descriptive property: It should capture the temporal structure and deal with motion speed variations.
  - Discriminative property: A motion phrase is expected to be highly related with certain action class.
  - Representative property: A single motion phrase can support a subset of videos while the whole set of motion phrases should cover the various patterns in different action videos.
- **Measures of discriminative and representative ability:**
  - For a single motion phrase P:

$$\mathbf{Rep}(P, c) = \frac{\sum_{i \in S(P,c)} r(V_i, P)}{|S(P, c)|}, \qquad \mathbf{Dis}(P, c) = \mathbf{Rep}(P, c) - \max_{c_j \in C - c} \mathbf{Rep}(P, c_j)$$

  where $S(P, c)$ denotes a set of videos: $S(P, c) = \{i | c(V_i) = c \wedge V_i \in top(P)\}$.
  - For a set of motion phrases $\mathcal{P} = \{P_i\}_{i=1}^{K}$:

$$\mathbf{RepSet}(\mathcal{P}, c) = \frac{1}{N_c} | \cup_{P_i \in \mathcal{P}} S(P_i, c)|.$$

- **Algorithms for Mining motion phrases:**

Algorithm 2: Mining motion phrases

Data: videos: $\mathcal{V} = \{V_i, y_i\}_{i=1}^{N}$, motion atoms: $\mathcal{A} = \{A_i\}_{i=1}^{M}$.
Result: Motion phrases: $\mathcal{P} = \{P_i\}_{i=1}^{K}$.
- Compute response value for each atom unit on all videos $v(V, \Pi)$ defined by Equation (3).
foreach *class c* **do**
  1. Select a subset of atom units (see Algorithm 3).
  2. Merge continuous atom units into 1-motion phrase $\mathcal{P}_1^c$.
  while $maxsize < MAX$ **do**
    a. Generate candidate s-motion phrase based on $(s-1)$-motion phrase.
    b. Select a subset of motion phrases $\mathcal{P}_s^c$ (see Algorithm 3).
  end
  3. Remove the motion phrase whose $\mathrm{Dis}(P, c) < \tau$.
end
- Return motion phrases: $\mathcal{P} = \cup_{c,s} \mathcal{P}_s^c$.

Algorithm 3: Selecting a subset of motion phrases.

Data: motion phrases candidates $\mathcal{P} = \{P_i\}_{i=1}^{L}$, class: c, number: K_c.
Result: selected motion phrases: $\mathcal{P}^* = \{P_i\}_{i=1}^{K_c}$.
- Compute the representative ability of each motion phrase $\mathrm{Rep}(P, c)$ defined in Equation (6).
- Initialization: $n \leftarrow 0, \mathcal{P}^* \leftarrow \emptyset$.
while $n < K_c$ **do**
  1. For each remaining motion phrase P, compute: $\triangle\mathrm{RepSet}(P, c) = \mathrm{RepSet}(\mathcal{P} \cup P, c) - \mathrm{RepSet}(\mathcal{P}, c)$, where $\mathrm{RepSet}(\mathcal{P}, c)$ is defined in Equation (8).
  2. Choose the motion phrase: $P^* \leftarrow \arg\max_P [\mathrm{Rep}(P, c) + \triangle\mathrm{RepSet}(P, c)]$.
  3. Update: $n \leftarrow n + 1, \mathcal{P}^* \leftarrow \mathcal{P}^* \cup \{P^*\}$
end
- Return motion phrases: $\mathcal{P}^*$.

References

1. S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In ECCV, 2012.
2. H.Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In CVPR, 2011.
3. B. Yao and F.-F. Li. Grouplet: A structured image representation for recognizing human and object interactions. In CVPR, 2010.

Experiment Results on OSD and UCF50

- **Examples of motion atoms and phrases:**

Figure 3: Examples of motion atoms and phrases.

- **Exploration of different sizes for motion phrases on OSD:**

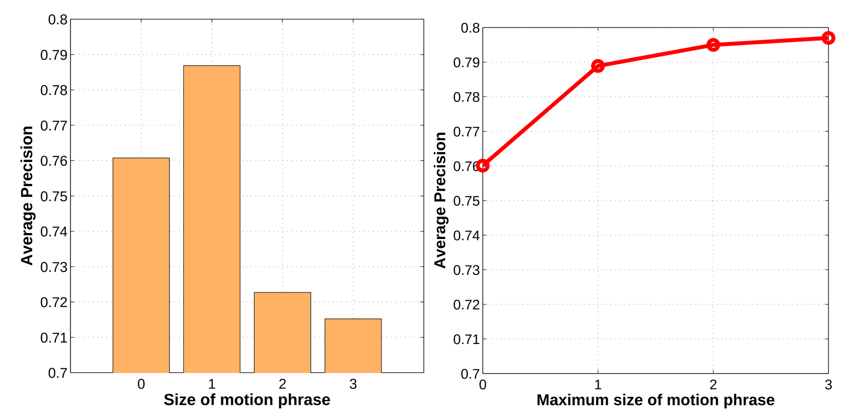

Figure 4: Recognition performance on OSD with different phrase sizes.

- **Effectiveness of motion atoms and phrases:**

Dataset	Olympic Sports	UCF50
Low-level Features (linear)	58.1 %	66.6 %
Low-level Features (kernel)	70.1 %	77.4 %
Motion Atoms	76.1%	82.5%
Motion Atoms and Phrases	79.5%	84.0%
Combine All	**84.9%**	**85.7%**

Table 1: Performance comparison of different representations.

- **Comparison with the state-of-the-art methods:**

Method	Performance
Laptev (CVPR 2008)	58.2%
Niebels (ECCV 2010)	62.5%
Liu (CVPR 2011)	74.4%
Tang (CVPR 2012)	66.8%
Wang (IJCV 2013)	74.1%
Our Best	84.9%

Method	Performance
Sadanand (CVPR 2012)	57.9%
Kliper (ECCV 2012)	72.6%
Reddy (MVAP 2012)	76.9%
Wang (CVPR 2013)	78.4%
Wang (IJCV 2013)	84.5% / 85.6%
Our Best	85.7%

Table 2: Compare with state-of-the-art methods (Left: OSD, Right: UCF50).