Static Image Action Recognition with Hallucinated Fine-grained Motion Information

Shengyuan Huang, Xing Zhao, Li Niu, Liqing Zhang

MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
目录 Contents

1 Introduction

2 Our method

3 Experiments
Introduction

- Static image action recognition
  - recognize human action in a single image
  - Lack of motion information
  - Leveraging unlabeled videos to hallucinate the motion information

- Related work
  - Using fine-grained low-level motion information
  - Using coarse-grained high-level motion information

- Our work
  - Using fine-grained high-level motion information
  - Attention mechanism
1 Introduction
2 Our method
3 Experiments
Our method: some pre-works

- Video Trajectory
- Region Proposal Extraction
- Region-level Pseudo Motion Label
- Motion Attention Map
Our method: some pre-works

- **Video Trajectory**
  - Extracting dense trajectories
  - Obtaining four local descriptors: Trajectory coordinates, HOG, HOF and MBH.
  - Encoding the local descriptor into Fisher vector.
  - Using PCA to reduce the dimension.
  - Obtaining hand-crafted feature.
Our method: some pre-works

- Region Proposal Extraction
  - Using keypoint detector to get 16 keypoints
  - Obtaining 7 region proposals (head, torso, left arm, left hand, right arm, right hand, and lower body)
Our method: some pre-works

- Region-level Pseudo Motion Label
  - Finding all trajectories passing through this region proposal
  - Averaging hand-crafted motion features of these trajectories
  - Using K-Means to obtain K clusters.
Our method: some pre-works

- Motion Attention Map
  - Obtaining motion attention map based on trajectories.
Our method

\[ L = \lambda_m (L_{mf} + L_{mi}) + \lambda_2 L_2 + L_{cls} \]

Fig. 2. The illustration of our proposed motion hallucination network. \( X_i^s \) (resp., \( X_i^v \)) and \( \hat{M}_i^s \) (resp., \( \hat{M}_i^v \)) are the hallucinated motion feature map and motion attention map of \( I_i^s \) (resp., \( I_i^v \)) respectively. The blue (resp., green) arrow represents the data flow of a static image (resp., video frame).
目录 Contents

1 Introduction
2 Our method
3 Experiments
Experiments

- Datasets
  - Two unlabeled video datasets
    - Stanford40
    - PASCAL VOC 2012 Actions
  - Two action image datasets
    - UCF101
    - HMDB
## Experiment results

<table>
<thead>
<tr>
<th>Video dataset</th>
<th>UCF101</th>
<th>HMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image dataset</td>
<td>S40</td>
<td>S40</td>
</tr>
<tr>
<td>visual only</td>
<td>79.91</td>
<td>79.91</td>
</tr>
<tr>
<td>Walker et al. [21]</td>
<td>80.08</td>
<td>80.95</td>
</tr>
<tr>
<td>Walker et al. [6]</td>
<td>81.20</td>
<td>80.44</td>
</tr>
<tr>
<td>Im2Flow [3]</td>
<td>81.04</td>
<td>80.26</td>
</tr>
<tr>
<td>HVM [4]</td>
<td>81.87</td>
<td>81.22</td>
</tr>
<tr>
<td>MRA [5]</td>
<td>81.76</td>
<td>81.07</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>83.71</strong></td>
<td><strong>83.77</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video dataset</th>
<th>UCF101</th>
<th>HMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image dataset</td>
<td>VOC</td>
<td>VOC</td>
</tr>
<tr>
<td>visual only</td>
<td>78.01</td>
<td>78.01</td>
</tr>
<tr>
<td>Walker et al. [21]</td>
<td>78.25</td>
<td>78.86</td>
</tr>
<tr>
<td>Walker et al. [6]</td>
<td>78.97</td>
<td>78.22</td>
</tr>
<tr>
<td>Im2Flow [3]</td>
<td>78.89</td>
<td>79.23</td>
</tr>
<tr>
<td>MRA [5]</td>
<td>79.87</td>
<td>79.19</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>81.08</strong></td>
<td><strong>80.88</strong></td>
</tr>
</tbody>
</table>

**Table 1.** Accuracies(%) of different methods in four settings. The best results are denoted in boldface.
Experiment results

- Ablation study

<table>
<thead>
<tr>
<th>Setting</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>visual only</td>
<td>79.91</td>
</tr>
<tr>
<td>motion only</td>
<td>77.60</td>
</tr>
<tr>
<td>motion only (coarse-grained)</td>
<td>68.60</td>
</tr>
<tr>
<td>visual+motion (coarse-grained)</td>
<td>82.47</td>
</tr>
<tr>
<td>w/o $L_{mf}$</td>
<td>81.11</td>
</tr>
<tr>
<td>w/o $L_{mi}$</td>
<td>83.08</td>
</tr>
<tr>
<td>w/o $L_2$</td>
<td>83.13</td>
</tr>
<tr>
<td>visual+motion (w/o att)</td>
<td>82.95</td>
</tr>
<tr>
<td>visual+motion (full)</td>
<td><strong>83.71</strong></td>
</tr>
</tbody>
</table>

**Table 2.** Ablation studies of our method in the setting of UCF101 → Stanford40.
Experiment results

- Hyper-parameter Analyses

Fig. 3. Analyses of the number of clusters $K$ and the hyper-parameters $\lambda_m, \lambda_2$. The default values are indicated by vertical dashed lines.
Experiment results

- Visualization
  - motion attention maps

**Fig. 4.** Visualization of hallucinated motion attention maps of static images.
Experiment results

- Visualization
  - region-level motion features

**Fig. 5.** Visualization of region-level motion features. The top (resp., bottom) sequence of sub-frames on the right represents the head (resp., arm) movement in the left static image.
Thanks!