Mining Semantics-Preserving Attention for Group Activity Recognition

Yansong Tang\textsuperscript{1}, Zian Wang\textsuperscript{1}, Peiyang Li\textsuperscript{1}, Jiwen Lu\textsuperscript{1}, Ming Yang\textsuperscript{2}, Jie Zhou\textsuperscript{1}

\textsuperscript{1}Department of Automation, Tsinghua University, China
\textsuperscript{2}Horizon Robotics Inc., China
Human Activity Analytics

- Wide real-world applications
- Different levels of human activities

Sign Language Recognition
Gesture

Human-robot Interaction

Sports Video Analysis
Action

Sports Video Analysis
Group Activity

[1] Shu et al. ICRA2017
Group Activity Recognition

“What are the people doing in this video?”

“What are the people?”

Tracklets of different people provided by [Choi et al. ECCV 2012]

“What is the action of each person?”

Labels are available during training, but not available at testing.

Problem setting in this work

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Frames</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Tracklets</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Individual Action</td>
<td>√</td>
<td>?</td>
</tr>
<tr>
<td>Group Activity</td>
<td>√</td>
<td>?</td>
</tr>
</tbody>
</table>

Input Video

Tracklets of different people

Right Spike
Related Works – Group Activity Recognition

(a) HDTM [Ibrahim et al. CVPR2016]
Related Works – Group Activity Recognition

Input Tracklets → Features of individual actions → Feature of group activity → Label of group activity

(a) HDTM [Ibrahim et al. CVPR2016]

Semantic Domain

Input Frames → Predicted captions of individual actions → Labels of individual actions → Label of group activity

(b) SBGAR [Li et al. ICCV2017]
Related Works – Group Activity Recognition

(a) HDTM [Ibrahim et al. CVPR2016]

(b) SBGAR [Li et al. ICCV2017]

(c) Our method

Input Tracklets

Features of individual actions

Feature of group activity

Label of group activity

Features of individual actions

Label of group activity

Features of group activity

(c) Our method

Appearance Domain

Semantic Domain

Labels of individual actions

Features of individual actions

Predicted captions of individual actions

Label of group activity

Labels of individual actions

Input Frames

Semantic Domain
Related Works – Attention Model (AM)

Attention model (AM): selecting the most informative parts from the global field.

- The group activity is usually sensitive to a few key persons
- Other people may bring ambiguous information and mislead the recognition process

[Rao et al. ICCV 2017]

[Song et al. AAAI 2017]
Related Works – Attention Model (AM)

Our Main Idea: Employ the learned attention information by a Teacher Network in the semantics domain, to guide a Student Network in the appearance domain.
**Approach**

### Teacher Network

- **Input Words of Individual Actions**
  - Blocking
  - Spiking
  - Standing
  - Waiting

- **One-hot Encoding**

- **Teacher Network**
  - **$f_{oh,n}$**
  - **$f_{em,n}$**
  - **Teacher's Attention**
  - **Weighted Sum**
  - **$V_{agg}$**
  - **Knowledge Distillation Constraint**

### Student Network

- **Input Tracklets**

- **Feature Extraction**

- **$G$**

- **Weighted Sum**

- **$W^T_{agg}$**

- **BLSTM Layer**

### Knowledge Distillation Constraint

### Group Activity: “Right Spike”

### Classification Constraint

### Extract Features [Donahue et al. CVPR2015]

### Compute Optical Flow [Ilg et al. CVPR2017]
Approach

Input Tracklets

Attention Model

- $s_n = \tanh(W_3 \cdot f_{em,n} + b_3)$
- $\alpha_n = \exp(s_n)/\sum_{j=1}^{N} \exp(s_j)$
- $v_{agg} = \sum_{n=1}^{N} \alpha_n \cdot f_{em,n}$

Extract Features [Donahue et al. CVPR2015]  Compute Optical Flow [Ilg et al. CVPR2017]
Loss Function

\[ J = J_{CLS} + \lambda_1 J_{SPA} + \lambda_2 J_{KD} \]

\[ = - \sum_{l=1}^{L} \mathbb{1}(z = l) \log(P^l_S) \]

\[ + \lambda_1 \frac{1}{N} \sum_{n=1}^{N} (\mathbf{\alpha}_n - \frac{1}{T} \sum_{i=1}^{T} \beta^t_n)^2 \]

\[ + \lambda_2 \| P_T - P_S \|_2^2 \]

Approach

- Blockig
- Standing
- Waiting
- Spiking

Teacher's Attention

Student's Attention

Teacher's Network

Student Network

One-hot Encoding

Weighted Sum

Teacher's Attention

Student's Attention

Semantics-Preserving Attention Constraint

Knowledge Distillation Constraint

Classification Constraint

Group Activity: “Right Spike”

Extract Features [Donahue et al. CVPR2015]

Compute Optical Flow [Ilg et al. CVPR2017]
Datasets and Experiment Settings

Collective Activity (CA) dataset [1]

- 2420 video clips,
- 4 group activities, 6 individual actions

Volleyball dataset [2]

- 4830 video clips,
- 8 group activities, 9 individual actions

We follow the experimental setup in [3], to merge the class of “walking” and “crossing” as a new class of “moving”.

# Experimental Results

Comparison of the group activity recognition accuracy on the volleyball dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MPCA</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDTM</td>
<td>82.9</td>
<td>CVPR’16</td>
</tr>
<tr>
<td>CERN-2</td>
<td>83.6</td>
<td>CVPR’17</td>
</tr>
<tr>
<td>stagNet</td>
<td>84.4</td>
<td>ECCV’18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>MPCA</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours-SA</td>
<td>86.1</td>
<td>-</td>
</tr>
<tr>
<td>+OF</td>
<td>87.0</td>
<td>0.9</td>
</tr>
<tr>
<td>+SPA</td>
<td>89.5</td>
<td>2.5</td>
</tr>
<tr>
<td>+KD</td>
<td>90.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

(+OF): combining optical flow  
SA: self-attention  
SPA: semantics-preserving attention  
KD: knowledge distillation loss

Comparison of the group activity recognition accuracy on the CA dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MPCA</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinality Kernel</td>
<td>88.3</td>
<td>CVPR’15</td>
</tr>
<tr>
<td>CERN-2</td>
<td>88.3</td>
<td>CVPR’17</td>
</tr>
<tr>
<td>RMIC</td>
<td>89.4</td>
<td>CVPR’17</td>
</tr>
<tr>
<td>HDTM</td>
<td>89.6</td>
<td>CVPR’16</td>
</tr>
<tr>
<td>stagNet</td>
<td>91.3</td>
<td>ECCV’18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>MPCA</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours-SA</td>
<td>91.5</td>
<td>-</td>
</tr>
<tr>
<td>+OF</td>
<td>94.3</td>
<td>2.8</td>
</tr>
<tr>
<td>+SPA</td>
<td>95.6</td>
<td>1.3</td>
</tr>
<tr>
<td>+KD</td>
<td>95.7</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Experimental Results

Left Spike

BRA 0 20
JPN 0 18
Experimental Results

Left Spike

- Left Spiking
  - SA: 36
  - TA: 80
  - SPA: 62

- Left Standing
  - SA: 60
  - TA: 5
  - SPA: 20

Right Blocking

- SA: 25
- TA: 51
- SPA: 49

Right Standing

- SA: 62
- TA: 7
- SPA: 7

SA (Student’s Attention w/o SPA), TA (Teacher’s Attention), SPA (Semantics-preserving attention)
Analysis on Computational Time

- **Without utilizing optical flow:**
  
  Train SPTS: **13.19h** = 0.32 + 11.50 + 0.46 + 0.91

  Train the Teacher Network: **0.32h**, **2.43%** of the entire training time

  Testing (a video clip with 10 frames):
  
  967.67ms = 10 × (8.01 × 12) + 6.47 = 967.67ms

- **Combining optical flow:**

  Train SPTS: **86.72h**
  
  = 0.32 + 61.48 + 2 × (11.5 + 0.46) + 1.00

  Testing (a video clip with 10 frames):
  
  6276.70ms = 10 × (434.65 + 8.01 × 12 × 2) + 7.80
Summary

**Teacher Network (semantic domain):**
- Taking additional 2.43% computational time cost to train

**Student Network (appearance domain):**
- Guided by *semantics-preserving attention* learned by the Teacher Network

Original efforts leveraging attention in multimedia clues, both semantic and vision clues, performing group activity recognition
Thanks and Questions?

Poster on  P3-03

Yansong Tang  2018-10-24

tys15@mails.tsinghua.edu.cn