i-VisionGroup@Tsinghua

Mining Semantics-Preserving Attention for Group Activity Recognition

Yansong Tang[1], Zian Wang[1], Peiyang Li[1], Jiwen Lu[1], Ming Yang[2], Jie Zhou[1]

[1]Department of Automation, Tsinghua University, China

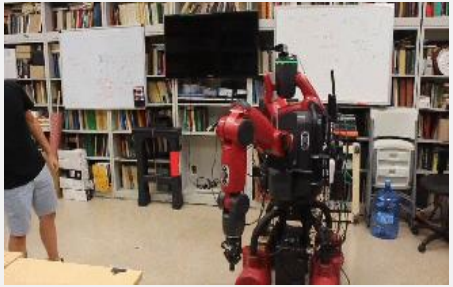
[2]Horizon Robotics Inc., China

Human Activity Analytics

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

Sign Language Recognition

Gesture



[1]

Human-robot Interaction

Interaction

Sports Video Analysis

Action

[2]

Sports Video Analysis

Group Activity

[1] Shu et al. ICRA2017

[2] Ibrahim et al. CVPR2016

Group Activity Recognition

Input Video

Tracklets of different people

"What are the people doing in this video?"

"Where are the people?"

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

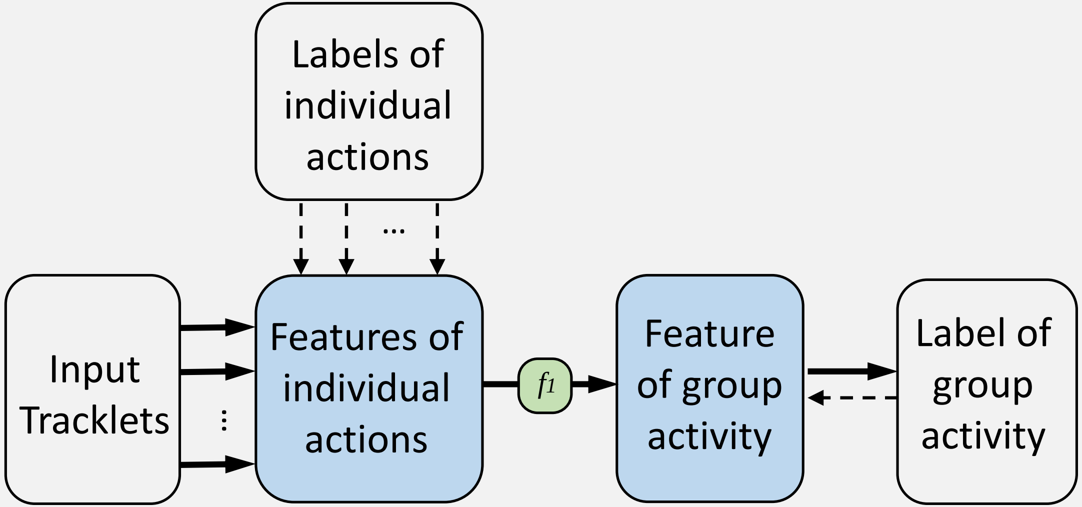
"What is the action of each person?"

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

Problem setting in this work

	Training	Testing
Video Frames	√	√
Tracklets	√	√
Individual Action	√	?
Group Activity	√	?

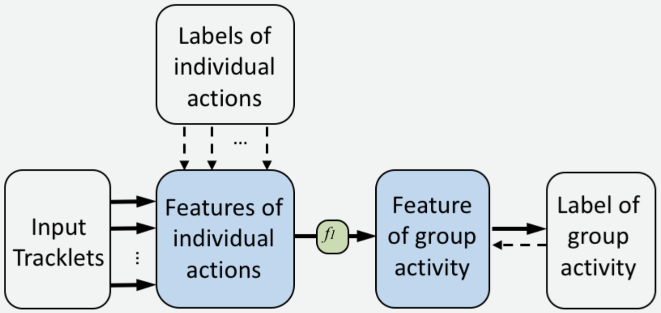
Related Works – Group Activity Recognition


Labels of individual actions
...
Input Tracklets
⋮
Features of individual actions
f_1
Feature of group activity
Label of group activity

(a) HDTM [Ibrahim et al. CVPR2016]

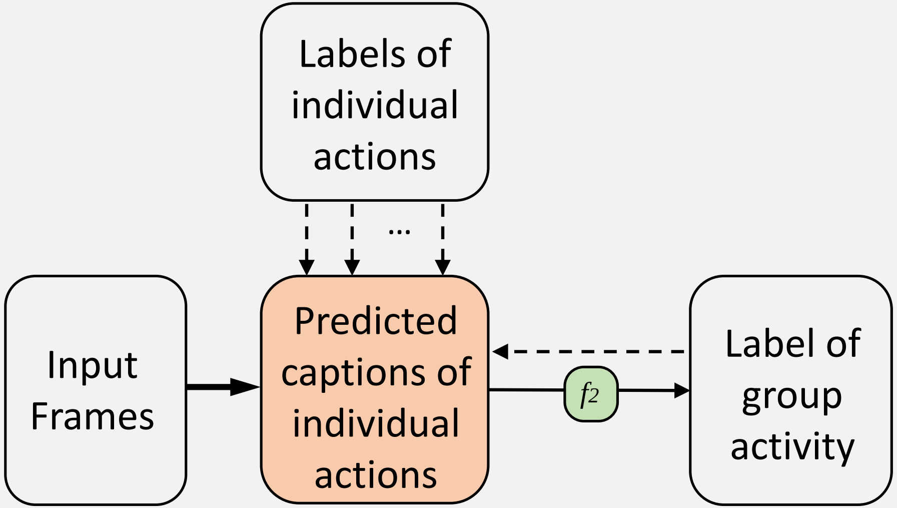
Appearance Domain

Related Works – Group Activity Recognition


Labels of individual actions
...
Input Tracklets
⋮
Features of individual actions
f1
Feature of group activity
Label of group activity

(a) HDTM [Ibrahim et al. CVPR2016]

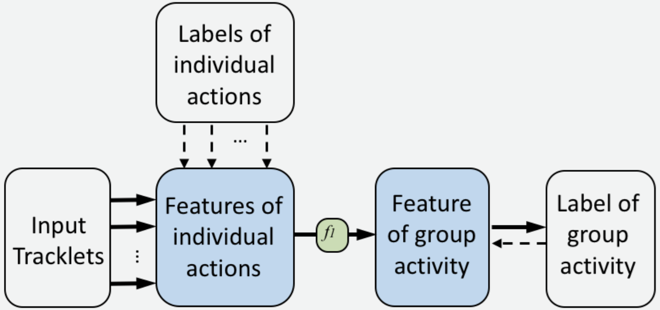
Appearance Domain


Labels of individual actions
...
Input Frames
Predicted captions of individual actions
f2
Label of group activity

(b) SBGAR [Li et al. ICCV2017]

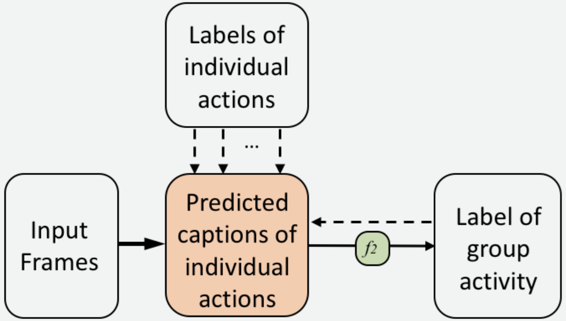
Semantic Domain

Related Works – Group Activity Recognition


Labels of individual actions
...
Input Tracklets
⋮
Features of individual actions
f_1
Feature of group activity
Label of group activity

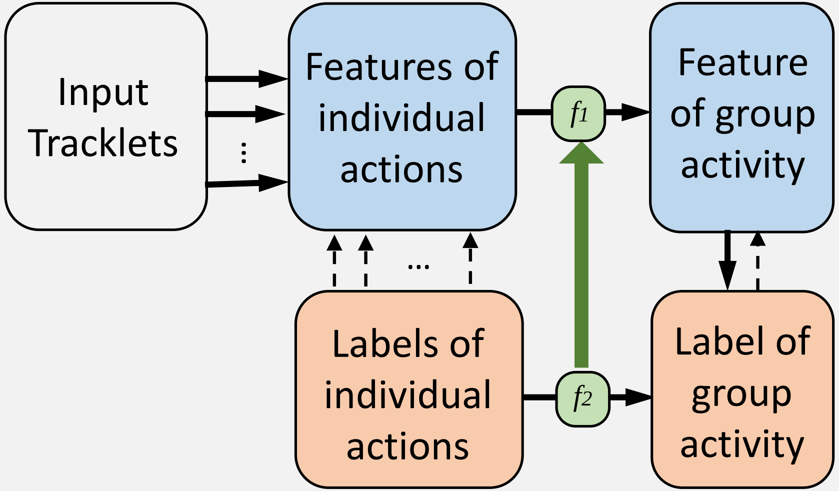
(a) HDTM [Ibrahim et al. CVPR2016]

Appearance Domain


Labels of individual actions
...
Input Frames
Predicted captions of individual actions
f_2
Label of group activity

(b) SBGAR [Li et al. ICCV2017]

Semantic Domain


Input Tracklets
⋮
Features of individual actions
f_1
Feature of group activity
...
Labels of individual actions
f_2
Label of group activity

(c) Our method

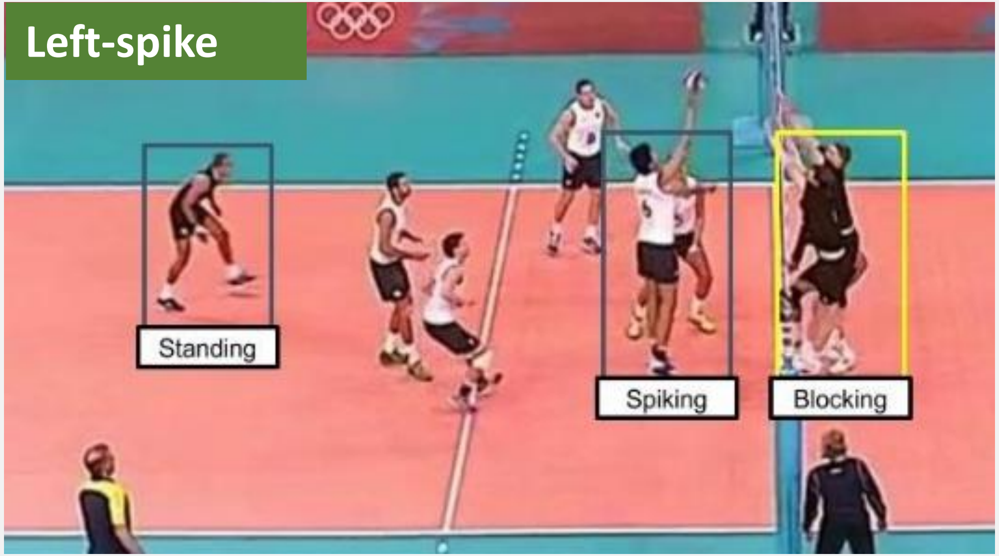
Appearance Domain

+

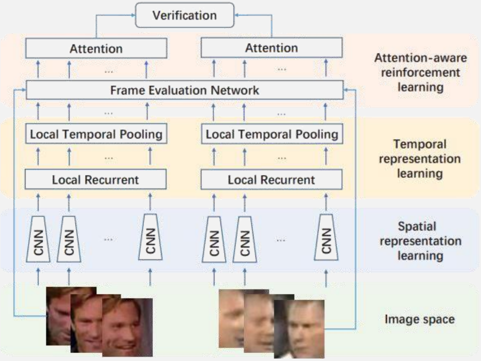
Semantic Domain

Related Works – Attention Model (AM)

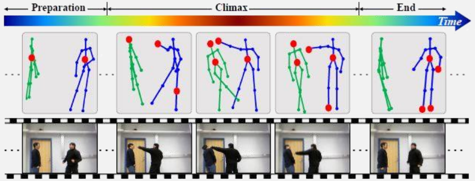
Attention model (AM): selecting the most informative parts from the global field.



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

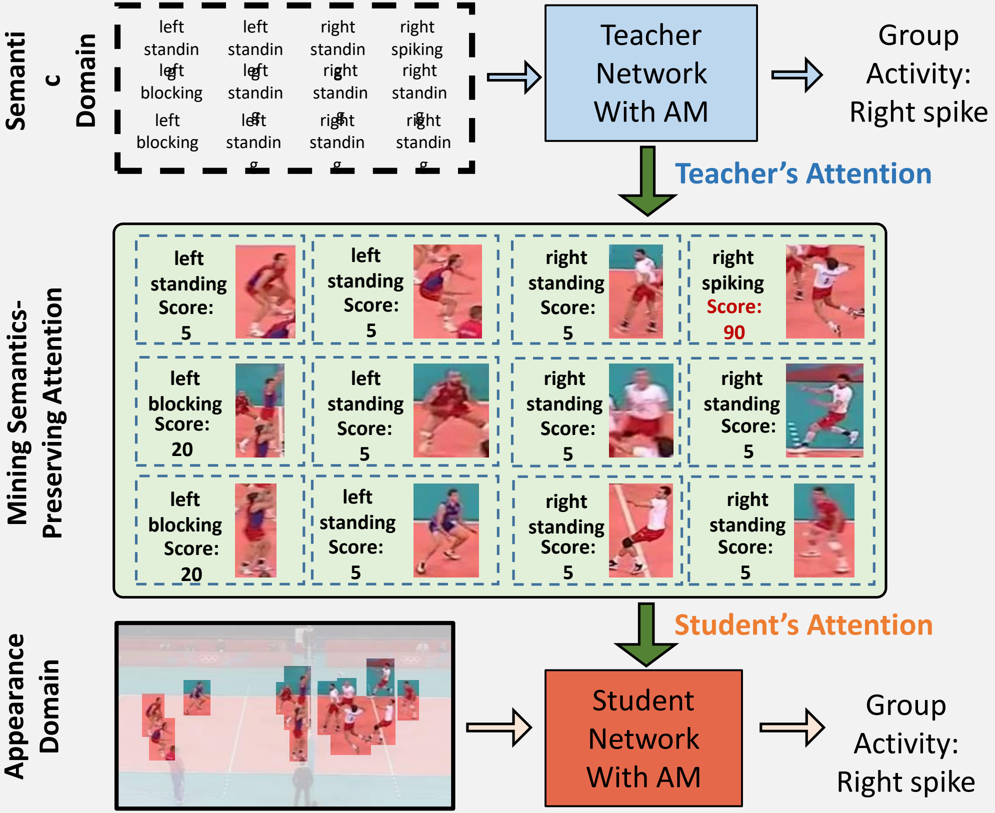


[Rao et al. ICCV 2017]



[Song et al. AAAI 2017]

Related Works – Attention Model (AM)



Semantic Domain

left standing	left standing	right standing	right spiking
left blocking	left standing	right standing	right standing
left blocking	left standing	right standing	right standing

Teacher Network With AM

Group Activity: Right spike

Teacher's Attention

Mining Semantics-Preserving Attention

left standing Score: 5	left standing Score: 5	right standing Score: 5	right spiking Score: 90
left blocking Score: 20	left standing Score: 5	right standing Score: 5	right standing Score: 5
left blocking Score: 20	left standing Score: 5	right standing Score: 5	right standing Score: 5

Appearance Domain

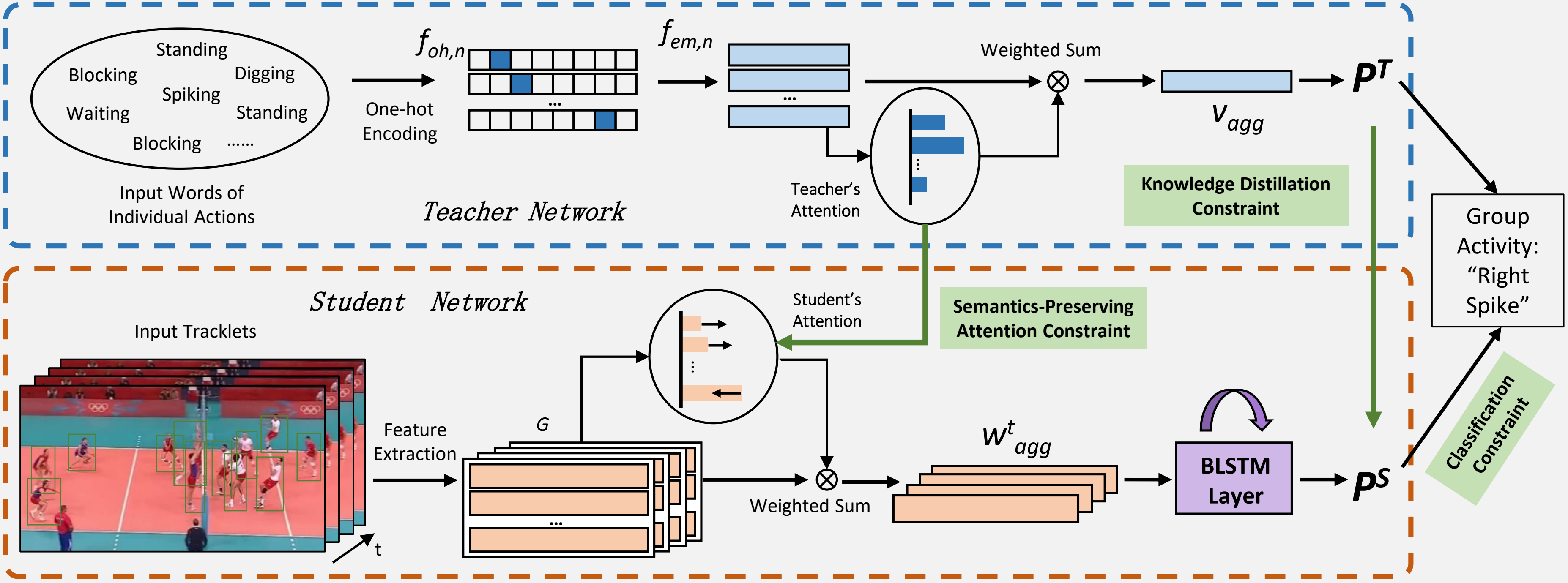
Student's Attention

Student Network With AM

Group Activity: Right spike

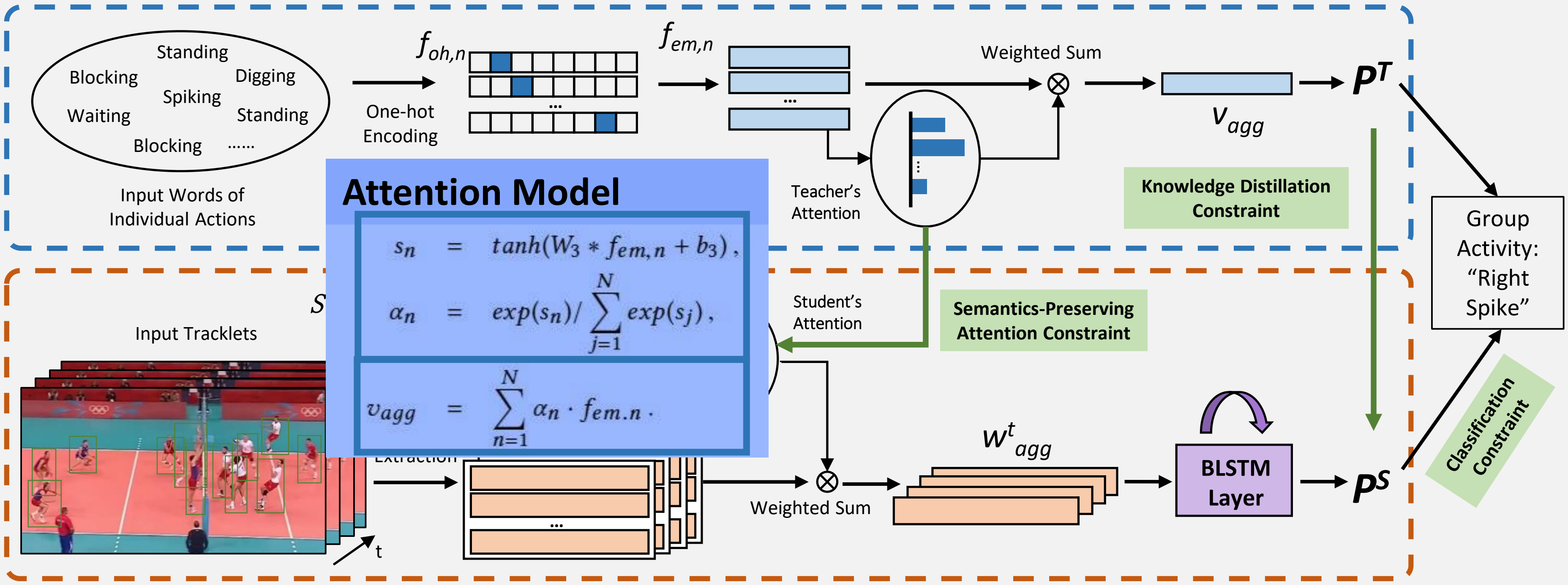
Our Main Idea:
Employ the learned attention information by a Teacher Network in the semantics domain, to guide a Student Network in the appearance domain.

Approach



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

Approach



Attention Model

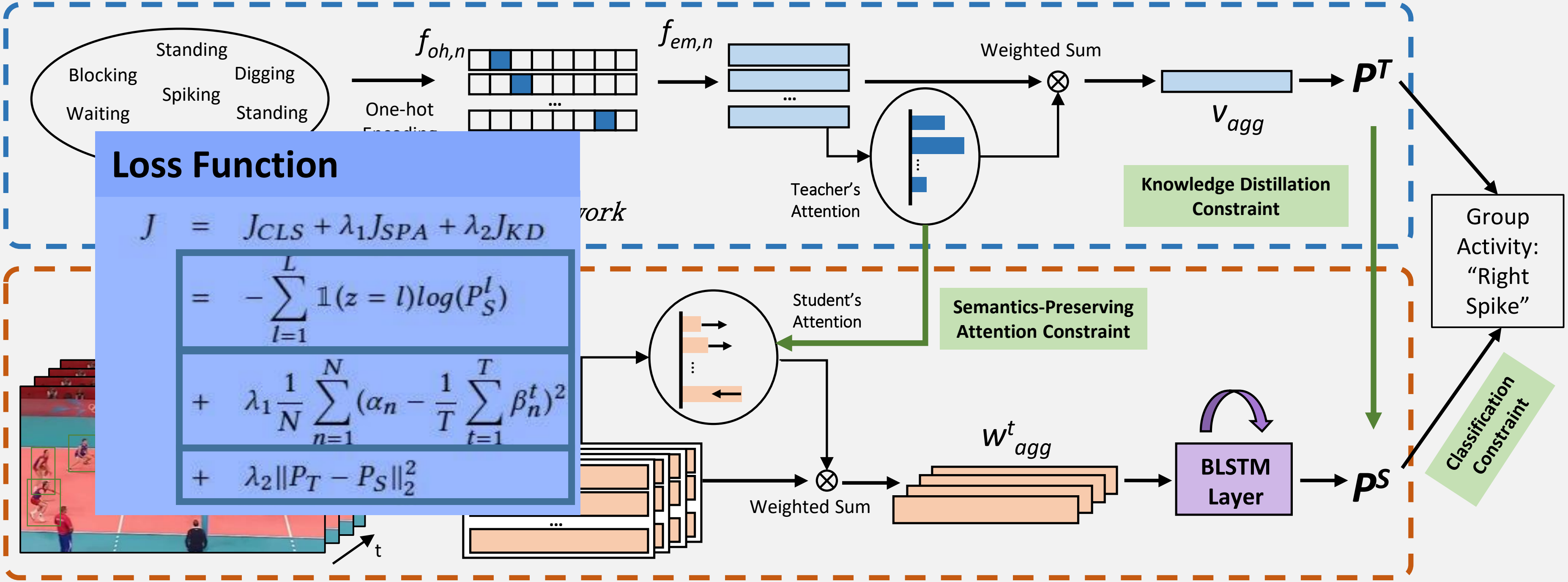
$$s_n = tanh(W_3 * f_{em,n} + b_3)\,,$$

$$\alpha_n = exp(s_n)/\sum_{j=1}^{N} exp(s_j)\,,$$

$$v_{agg} = \sum_{n=1}^{N} \alpha_n \cdot f_{em.n}\,.$$

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

Approach



Loss Function

$$
\begin{aligned}
J &= J_{CLS} + \lambda_1 J_{SPA} + \lambda_2 J_{KD} \\
&= -\sum_{l=1}^{L} \mathbb{1}(z = l) log(P_S^l) \\
&+ \lambda_1 \frac{1}{N} \sum_{n=1}^{N} (\alpha_n - \frac{1}{T} \sum_{t=1}^{T} \beta_n^t)^2 \\
&+ \lambda_2 ||P_T - P_S||_2^2
\end{aligned}
$$

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

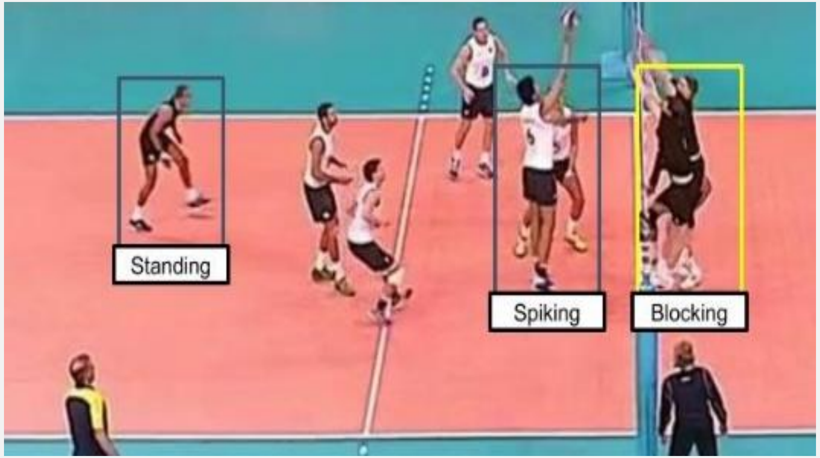
Datasets and Experiment Settings

Collective Activity (CA) dataset [1]

2420 video clips,
4 group activities, 6 individual actions

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

Volleyball dataset [2]



4830 video clips,
8 group activities, 9 individual actions

[1]Choi et al. ICCVW2009 [2]Ibrahim et al. CVPR2016 [3]Wang et al. CVPR2017

Experimental Results

Comparison of the group activity recognition accuracy on the volleyball dataset

Method	MPCA	Year
HDTM	82.9	CVPR’16
CERN-2	83.6	CVPR’17
stagNet	84.4	ECCV’18

Method	MPCA	Gain
Ours-SA	86.1	-
+OF	87.0	0.9
+SPA	89.5	2.5
+KD	90.0	0.5

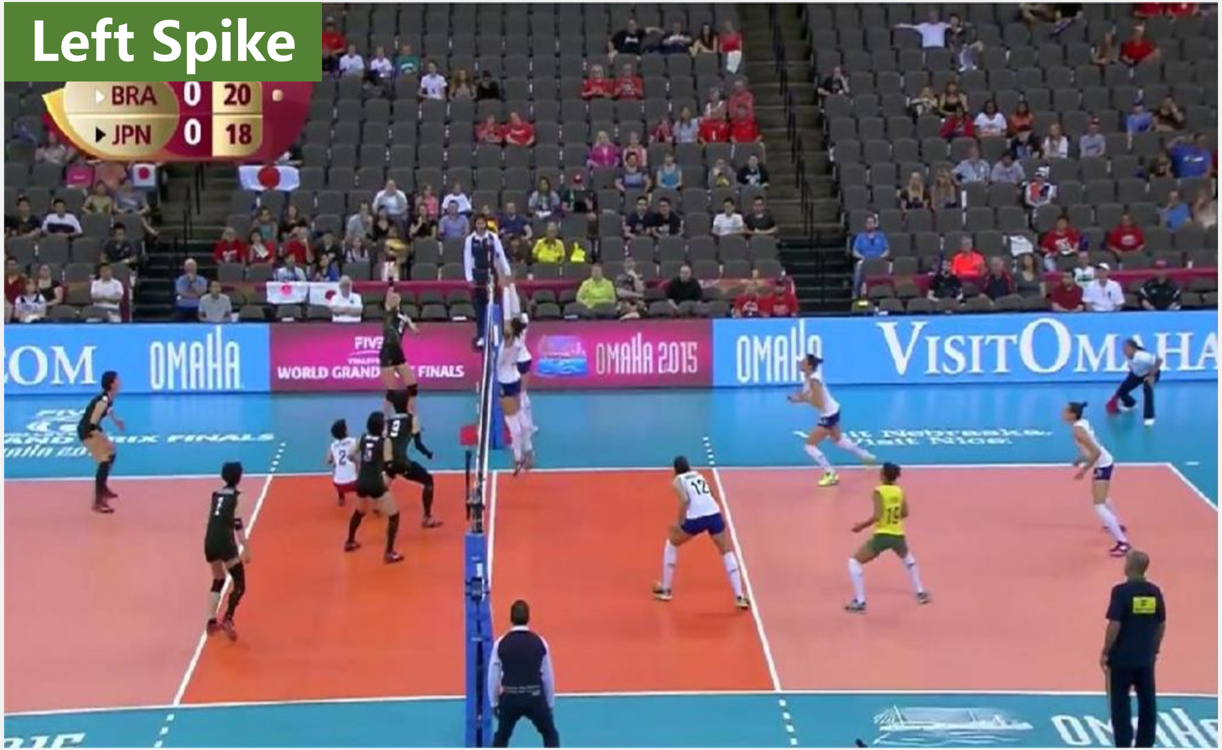
Comparison of the group activity recognition accuracy on the CA dataset

Method	MPCA	Year
Cardinality Kernel	88.3	CVPR’15
CERN-2	88.3	CVPR’17
RMIC	89.4	CVPR’17
HDTM	89.6	CVPR’16
stagNet	91.3	ECCV’18

Method	MPCA	Gain
Ours-SA	91.5	-
+OF	94.3	2.8
+SPA	95.6	1.3
+KD	95.7	0.1

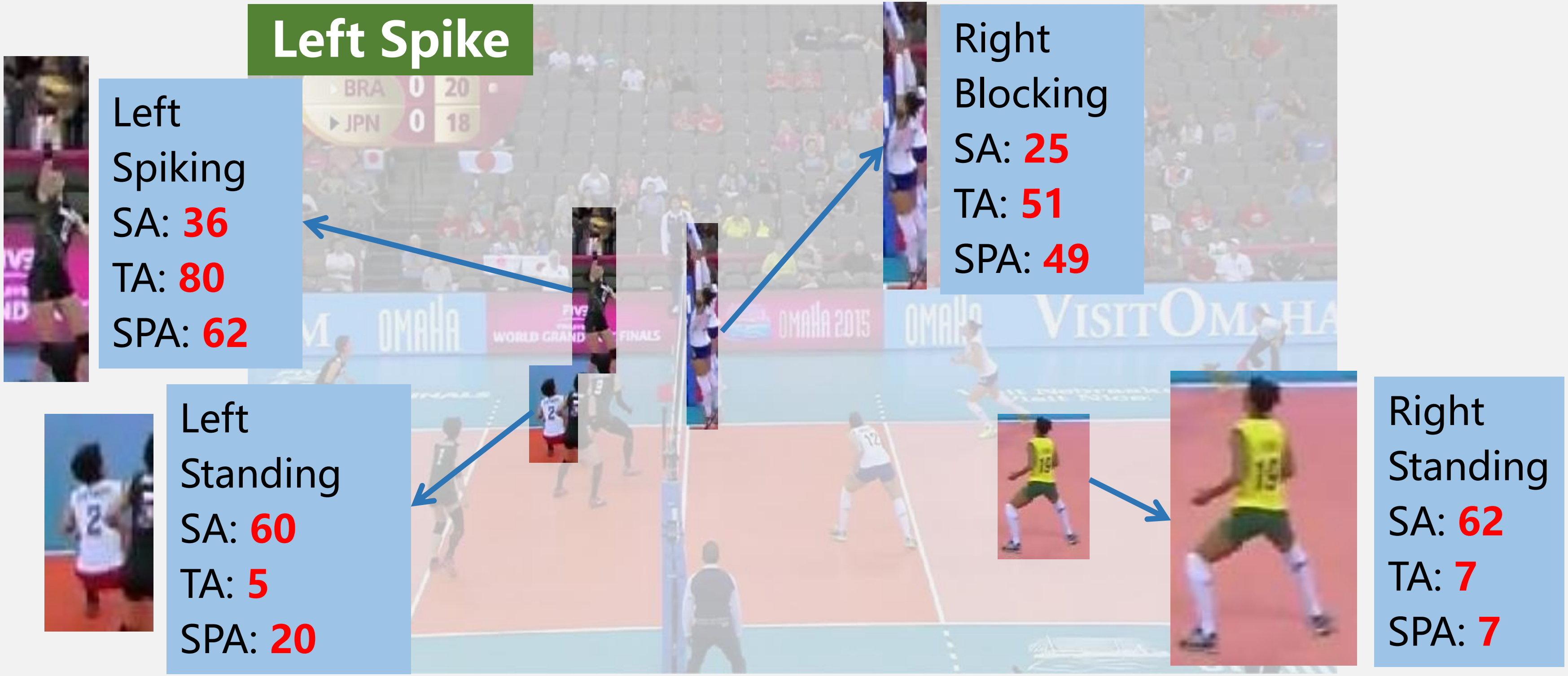
(+OF): combining optical flow   SA: self-attention   SPA: semantics-preserving attention   KD: knowledge distillation loss

Experimental Results



Left Spike

Experimental Results



Left Spike

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

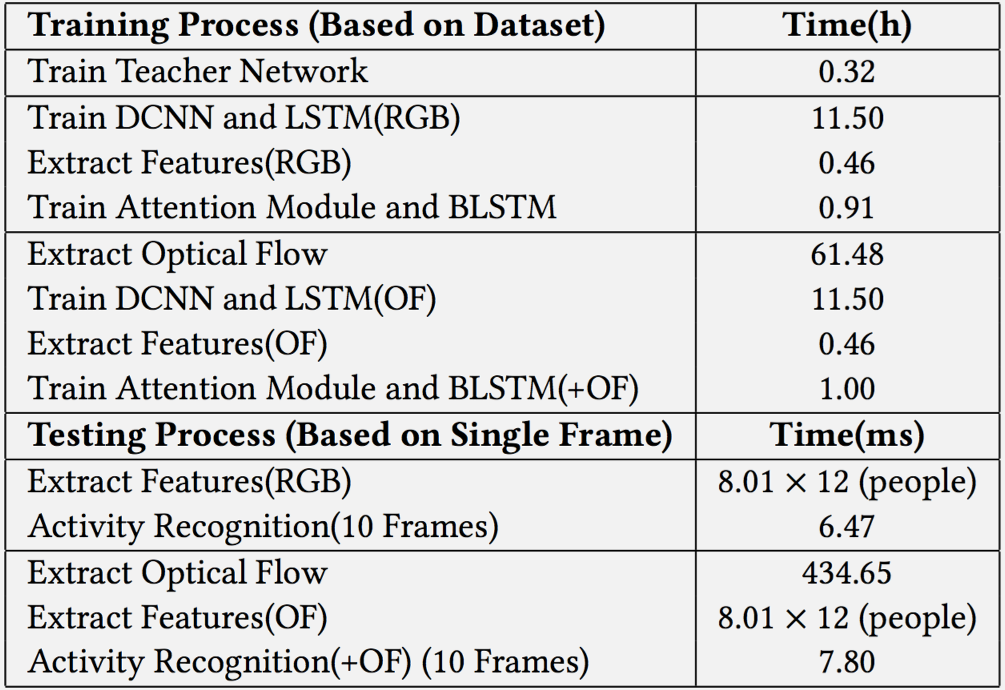
Right Blocking
SA: **25**
TA: **51**
SPA: **49**

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

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



Training Process (Based on Dataset)	**Time(h)**
Train Teacher Network	0.32
Train DCNN and LSTM(RGB)	11.50
Extract Features(RGB)	0.46
Train Attention Module and BLSTM	0.91
Extract Optical Flow	61.48
Train DCNN and LSTM(OF)	11.50
Extract Features(OF)	0.46
Train Attention Module and BLSTM(+OF)	1.00
Testing Process (Based on Single Frame)	**Time(ms)**
Extract Features(RGB)	8.01 × 12 (people)
Activity Recognition(10 Frames)	6.47
Extract Optical Flow	434.65
Extract Features(OF)	8.01 × 12 (people)
Activity Recognition(+OF) (10 Frames)	7.80

- **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 i-VisionGroup@Tsinghua

Thanks and Questions?

Poster on P3-03

Yansong Tang 2018-10-24

tys15@mails.tsinghua.edu.cn