LEHIGH UNIVERSITY

SBGAR: Semantics Based Group Activity Recognition

Xin Li, Mooi Choo Chuah
Lehigh University

Motivation

- Smart City typically involves large population participating in crowded events e.g. watching baseball games, NFL games
- Law personnel may want to monitor the crowd to quickly identify some suspicious behaviors
- Sport coaches may want to monitor a game and be alerted about game highlights.
- Group activity recognition is important in above application scenarios and hence having efficient schemes for identify group activity is critically important.

Existing Work

Existing approach in CVPR 2016 paper [7]:

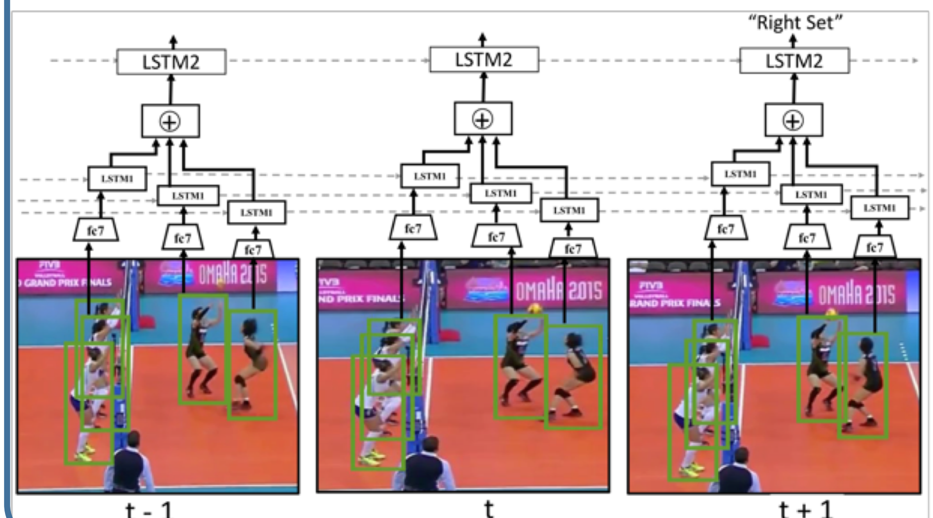
1. Detect all players from each frame
2. Employ a LSTM for each player
3. Output a corresponding group activity label

Our Approach:

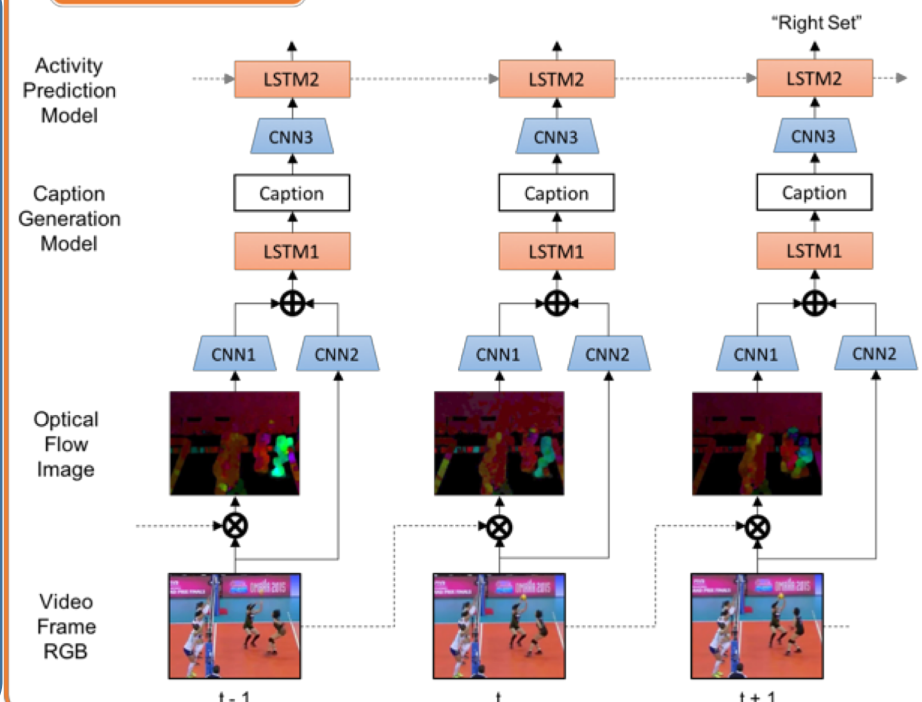
1. One LSTM to generate a sentence for each video frame. Generating sentences for frames allows users to:
    a. Search videos with similar content.
    b. Search videos by typing some sentences.
2. Also generate a group activity label. Can also group video frames into several sub-events of the same category e.g. spiking.

Group Activity Recognition

Scheme in [1]



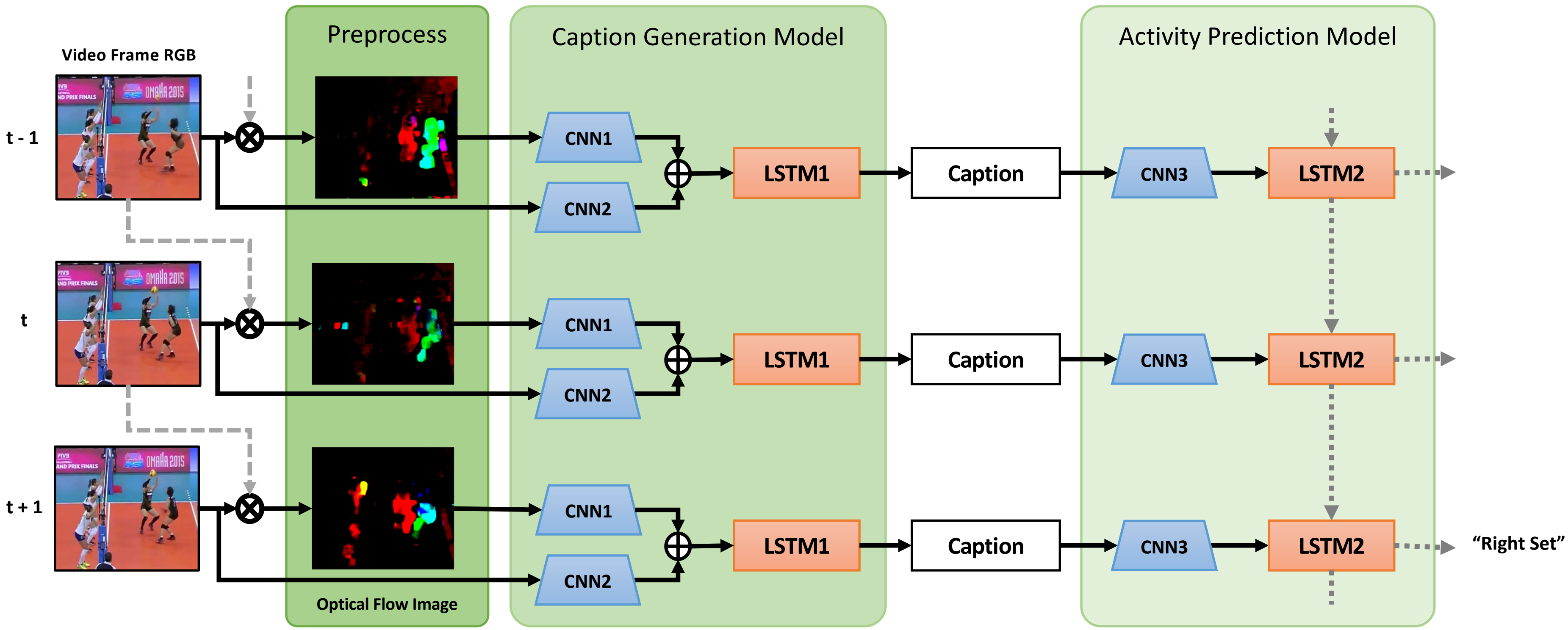
Our Scheme



[1] Ibrahim, Moustafa, et al. "A Hierarchical Deep Temporal Model for Group Activity Recognition." Computer Vision and Pattern Recognition. 2016

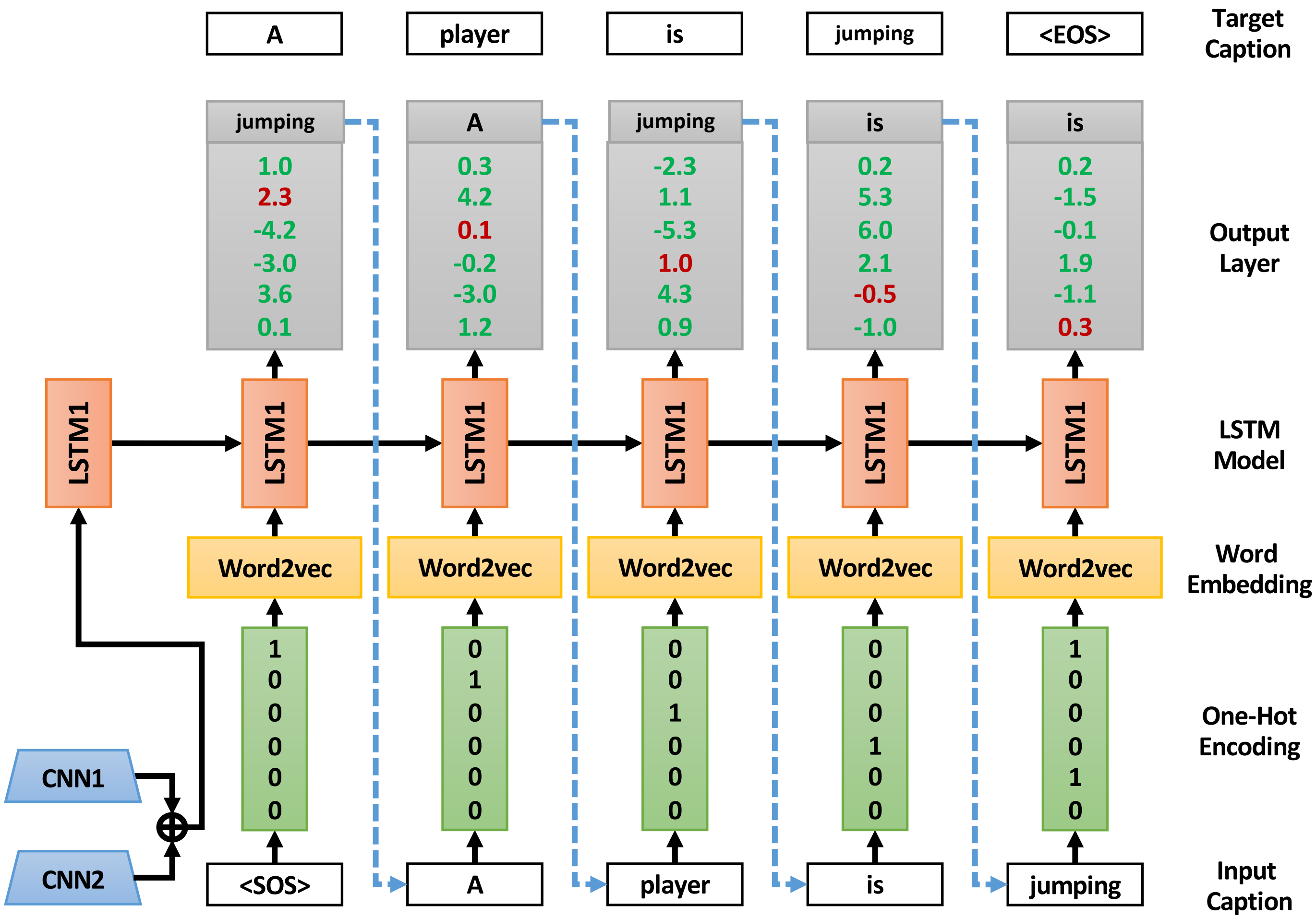
Group Activity Recognition

Our Solution



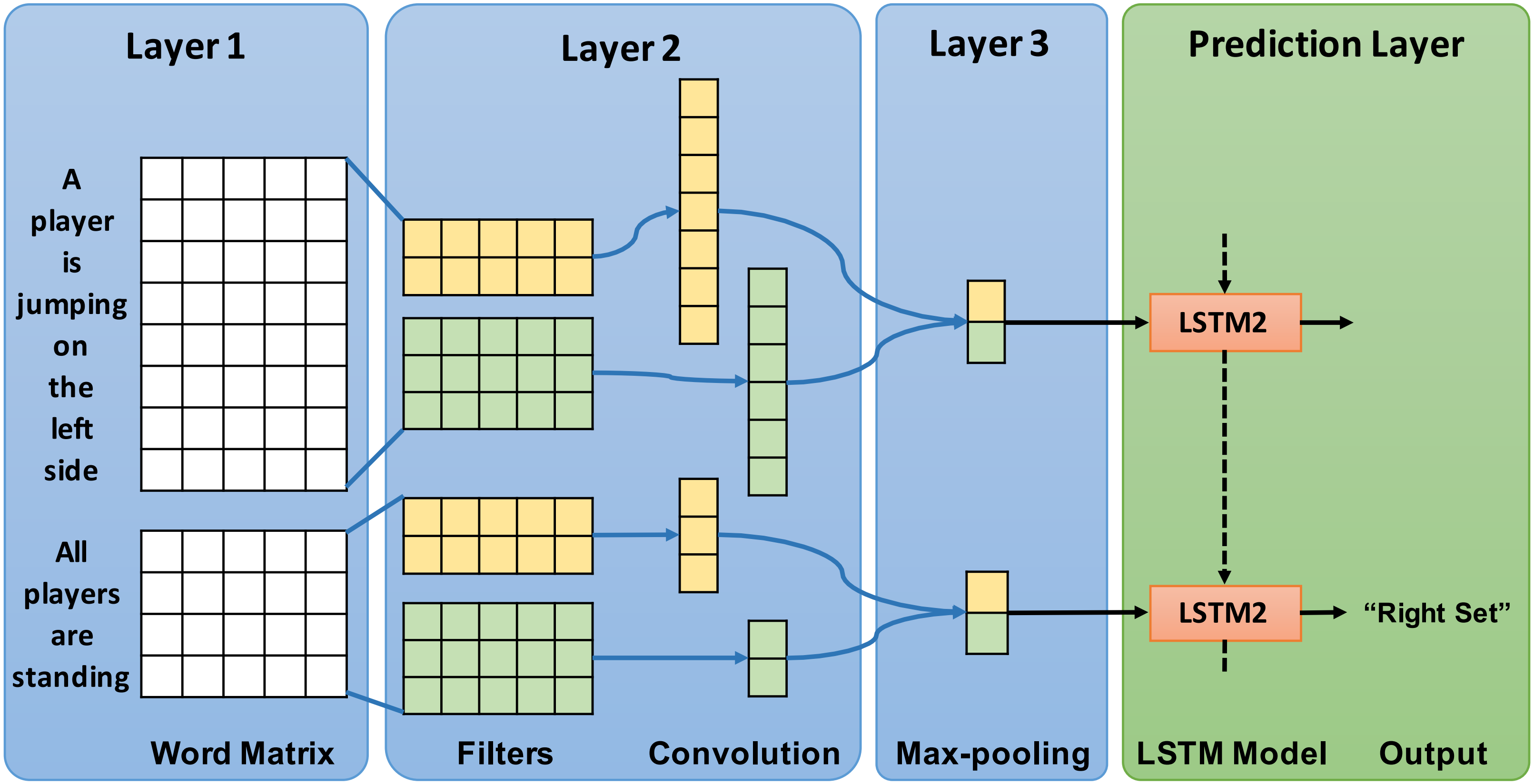
Group Activity Recognition

Caption Generation Model



Group Activity Recognition

Activity Prediction Model



Dataset1: VolleyBall

YouTube Volleyball (http://vml.cs.sfu.ca/wp-content/uploads/volleyballdataset/volleyball.zip):
4830 frames from 55 videos are annotated with 9 player action labels and 6 team activity labels.

Group Activity Class	No. of Instances
Right set	644
Right spike	623
Right pass	801
Left pass	826
Left spike	642
Left set	633

Action Classes	No. of Instances
Waiting	3601
Setting	1332
Digging	2333
Falling	1241
Spiking	1216
Blocking	2458
Jumping	341
Moving	5121
Standing	38696

Intermediate Results from Our Caption Generation Model

Left: standing blocking Right: standing setting moving

Left: standing waiting blocking Right: standing moving waiting spiking

Test Result using Volleyball Dataset

Result from [1]

Accuracy: 51.1%

	lset	rset	rspike	lspike	lpass	rpass
lset	56.94	16.67	4.17	2.78	12.50	6.94
rset	12.82	43.59	12.82	2.56	7.69	20.51
rspike	5.56	3.70	62.96	11.11	9.26	7.41
lspike	5.13	5.13	17.95	51.28	12.82	7.69
lpass	4.67	5.61	2.80	1.87	56.07	28.97
rpass	2.25	8.99	1.12	1.12	47.19	39.33

Our Result

Accuracy:66.9%

	lset	rset	rspike	lspike	lpass	rpass
lset	67.26	1.19	5.36	6.55	13.69	5.95
rset	3.13	52.08	11.98	1.56	6.77	24.48
rspike	0.00	6.36	79.19	0.00	8.67	5.78
lspike	7.26	0.00	1.12	82.12	3.35	6.15
lpass	11.06	1.33	8.85	2.65	55.75	20.35
rpass	3.33	8.10	3.81	5.24	10.48	69.05

[1] Ibrahim, Moustafa, et al. "A Hierarchical Deep Temporal Model for Group Activity Recognition." Computer Vision and Pattern Recognition. 2016

Test Result using Volleyball Dataset

Methods	**Accuracy (%)**
Two-stage Hierarchical Model [1] *	51.1
SBGAR (RGB Frame Only) SBGAR (Optical Flow Image Only)	38.7 54.3
SBGAR (RGB & Optical Flow)	**66.9**

Additional Test Results:

- Dataset: Collective Activity Dataset
- 44 short video sequences
- **5 different collective activities** :
  - crossing
  - walking
  - waiting
  - talking
  - queueing

Test Result using Collective Activity Dataset

Result from [1]

Accuracy: 81.5%

	crossing	waiting	queuing	walking	talking
crossing	61.54	4.27	0.85	33.33	0.00
waiting	11.41	66.44	0.00	22.15	0.00
queuing	0.00	0.00	96.77	3.23	0.00
walking	16.49	3.09	0.00	80.41	0.00
talking	0.00	0.00	0.00	0.55	99.45

Our Result

Accuracy:86.1%

	crossing	waiting	queuing	walking	talking
crossing	78.03	16.76	0.00	5.20	0.00
waiting	18.63	81.37	0.00	0.00	0.00
queuing	0.84	0.00	99.16	0.00	0.00
walking	10.74	0.67	1.01	87.58	0.00
talking	0.00	0.00	0.00	15.38	84.62

[1] Ibrahim, Moustafa, et al. "A Hierarchical Deep Temporal Model for Group Activity Recognition." Computer Vision and Pattern Recognition. 2016

Test Result using Collective Activity Dataset

Methods	Accuracy (%)
Contextual Model [2] *	79.1
Deep Structured Model [3] *	80.6
Two-stage Hierarchical Model [1] *	81.5
Cardinality kernel [4] *	83.4
SBGAR (RGB Frame Only)	83.7
SBGAR (Optical Flow Image Only)	70.1
SBGAR (RGB & Optical Flow)	**86.1**

Test Result: Computation Time

Testing on a desktop:

CPU: Intel i7 6700K, 4.2GHz

Memory:16GB

Graphic: GTX 1080

Our Scheme (Based On Single Frame)

Process	Computation time (ms)
De-shake	2.42
Optical Flow Image	19.77
Extract CNN Feature (Inceptionv3)	27.78
Caption generation	28.63
Activity Recognition	0.057
Total	78.657

Our Scheme (Based On 10 Frames)

Process	Computation time (ms)
De-shake	2.42 (* 10)
Optical Flow Image	19.77 (* 10)
Extract CNN Feature (Inceptionv3)	27.78 (* 10)
Caption generation	28.63 (* 10)
Activity Recognition(10 frames)	2.15
Total	80.75

* The input size of Inception-v3 is (299*299*3). Thus, we first resize the image into (299*299*3) and then collect the computation time.

Reference

- [1] Ibrahim, Moustafa, et al. "A Hierarchical Deep Temporal Model for Group Activity Recognition.” Computer Vision and Pattern Recognition. 2016
- [2] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, “discriminativeminative latent models for recognizing contextual group activities,” *IEEE Trans- actions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549–1562, 2012.
- [3] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidha- ran, M. J. Roshtkhari, and G. Mori, “Deep structured models for group activity recognition,” *arXiv preprint arXiv:1506.04191*, 2015.
- [4] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, “Visual recognition by counting instances: A multi- instance cardinality potential kernel,” in *Proceedings of the IEEE Conference on Computer Vision and Pat- tern Recognition*, 2015, pp. 2596–2605.