A Hierarchical Deep Temporal Model for Group Activity Recognition

CVPR 16 - Code

Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, *Greg Mori*
Simon Fraser University

Mostafa Saad Ibrahim
PhD Student @ Simon Fraser University

Human Activities

Gestures

Actions

Interactions

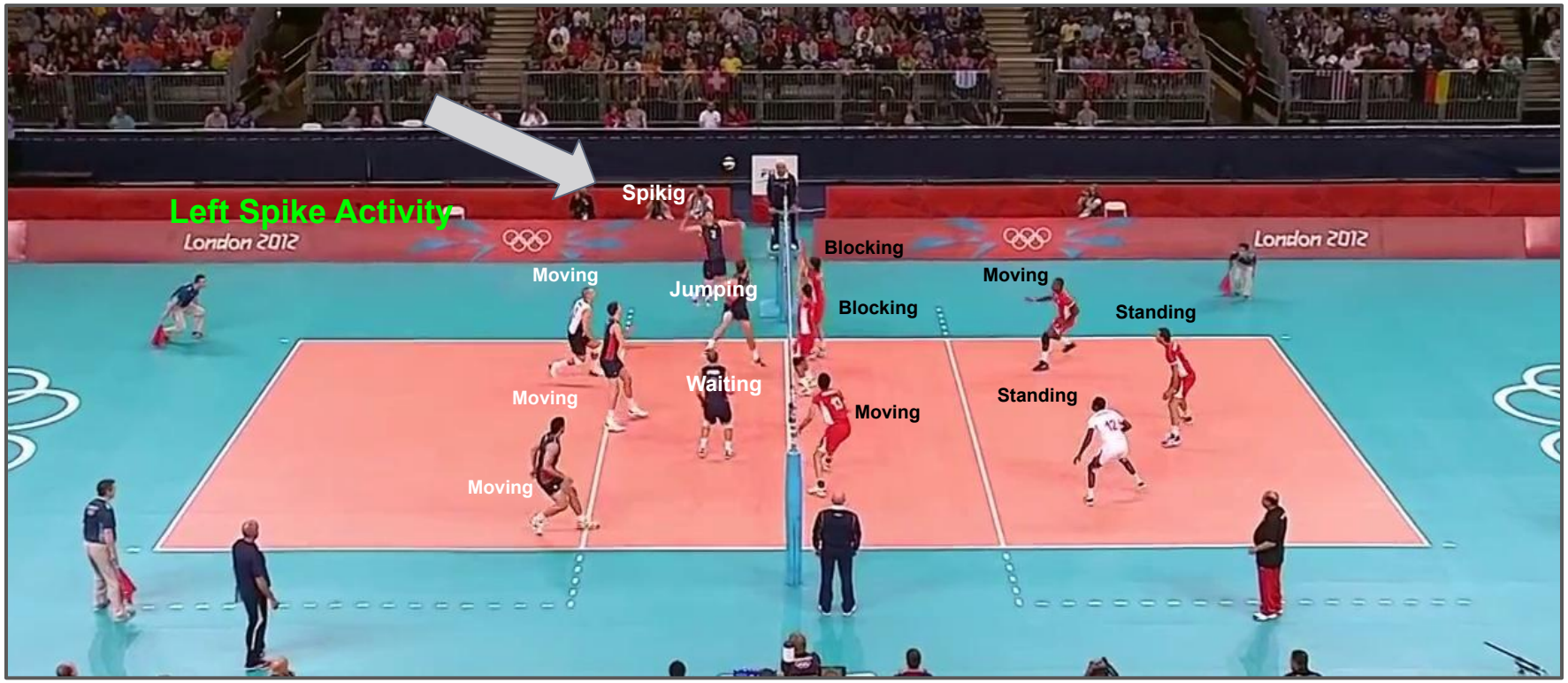
Group Activity

Group Activity Recognition

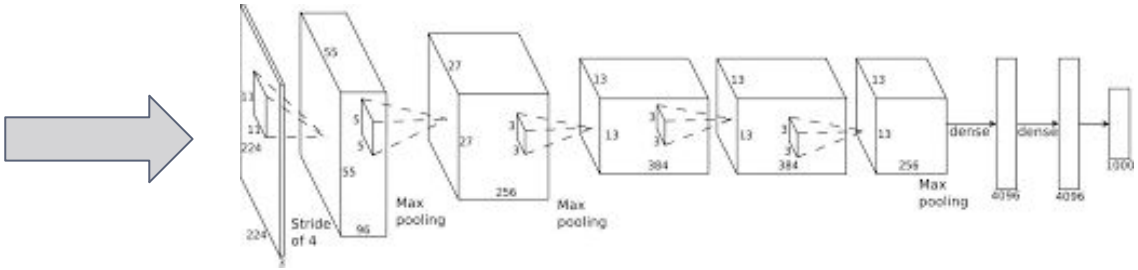
Major Activity = Walking scene.

Group Activity Recognition

Main Activity = Left Spike



Naive Approach: Image Classifier



Background?!
Person's Actions?!

Walking

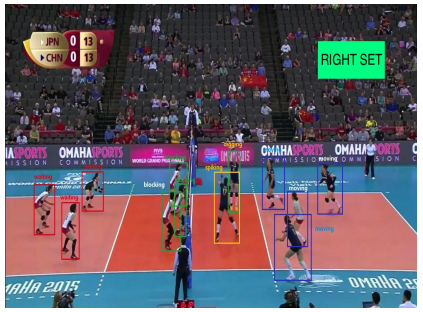
Hierarchical Modeling Approach

Understand
person action

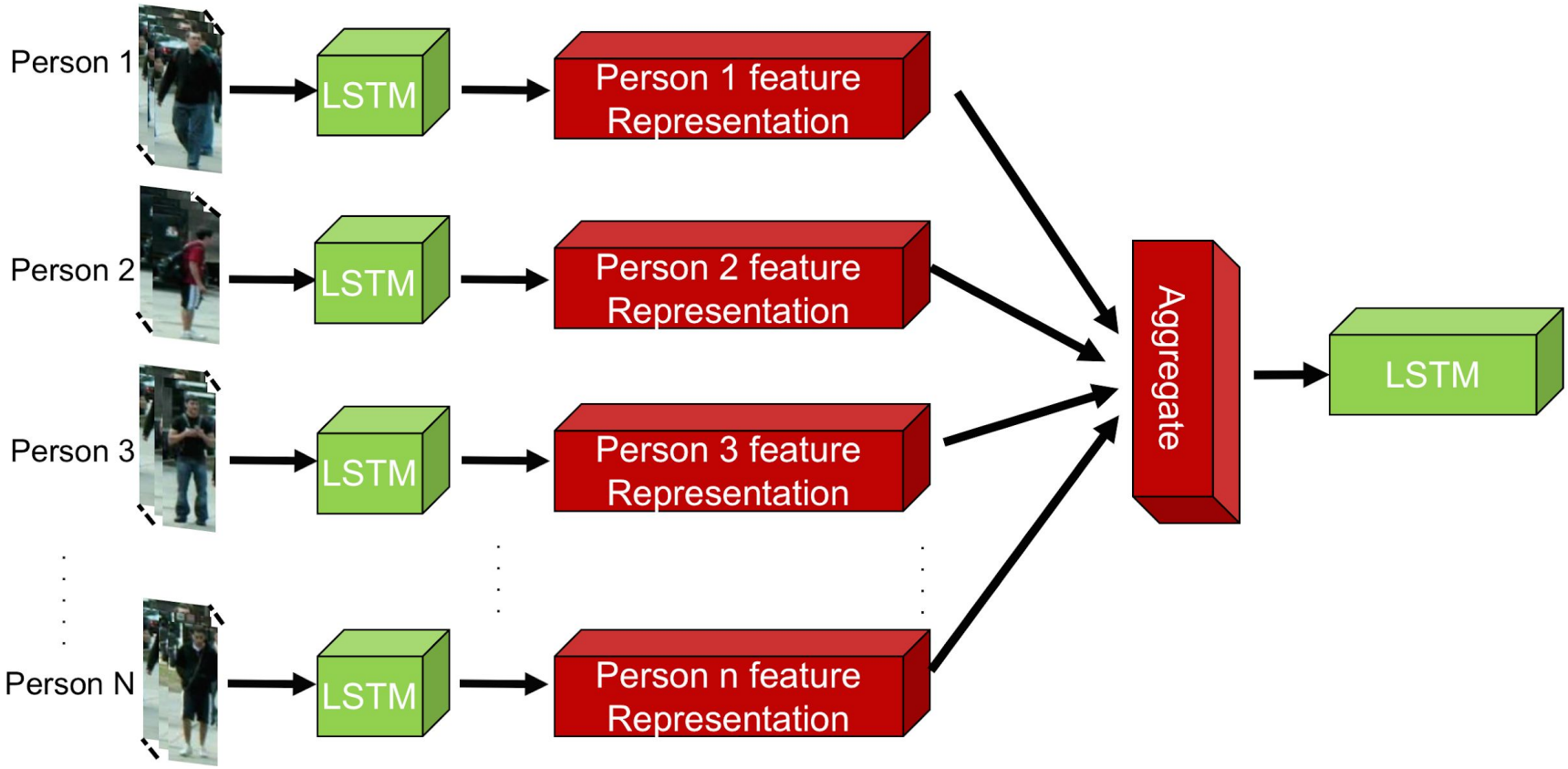
- Track people in the scene
- Learn **action** classifier
  - <Bounding Box, Person Action> inputs
- Extract person's representation

Understand
group activity

- Aggregate the people's representations
- Learn **activity** classifier
  - <Scene Representation, Scene Activity>



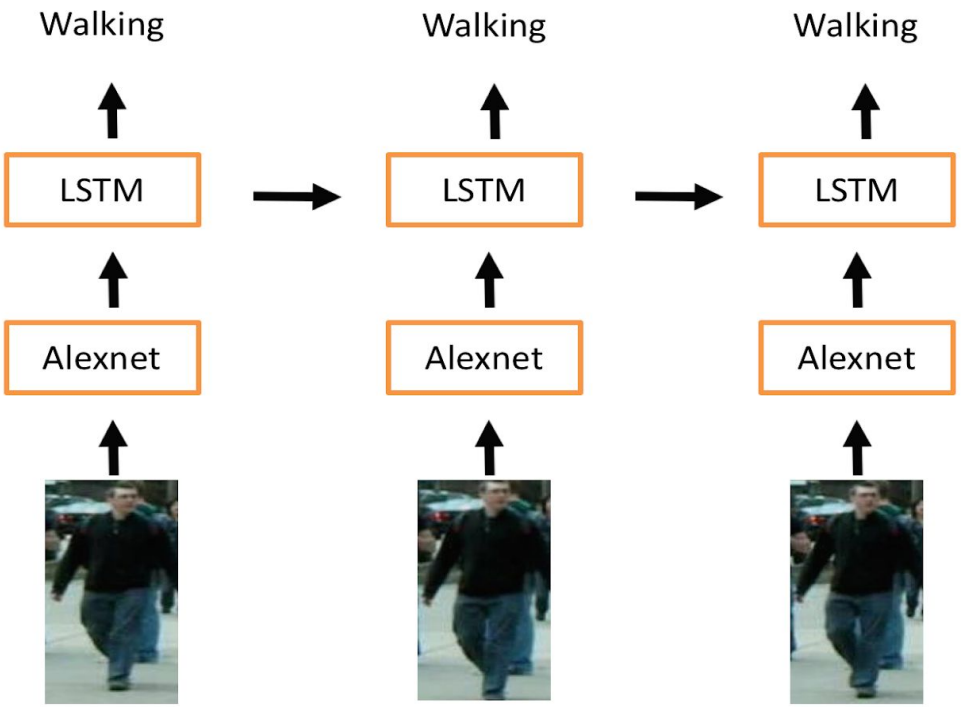
Hierarchical Deep Temporal Model



Person Action Classifier

- Build a **spatio-temporal** representation for **person's actions**.
- **Track** a manually annotated bounding box for each person for a fixed temporal window
- **Extract** deep visual representation for each tracked person using **Alexnet's fc7** features,
- Feed fc7 to a person **LSTM** to model the **temporal** dimension.
- **Extract spatio-temporal** features per person from its LSTM

Person Action Classifier



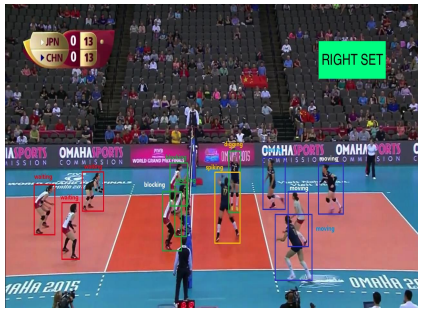
$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
g_t &= \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \phi(c_t)
\end{aligned}
$$

h_t is the extracted features from lstm layer representing spatio-temporal features of a person at time t

Group Activity Classifier

- Build a **spatio-temporal** representation for the **group activity** of a given frame
- **Aggregate** all individual person representations for every temporal step.
  - Standard pooling operators (e.g. max/avg pooling) are experimented
- Feed aggregated representation to a **group level LSTM**
- Extract spatio-temporal representation for the group activity from the top-level LSTM
- Learn a soft-max **classifier** on top of the **group activity** representation.

Pooling Persons' representations



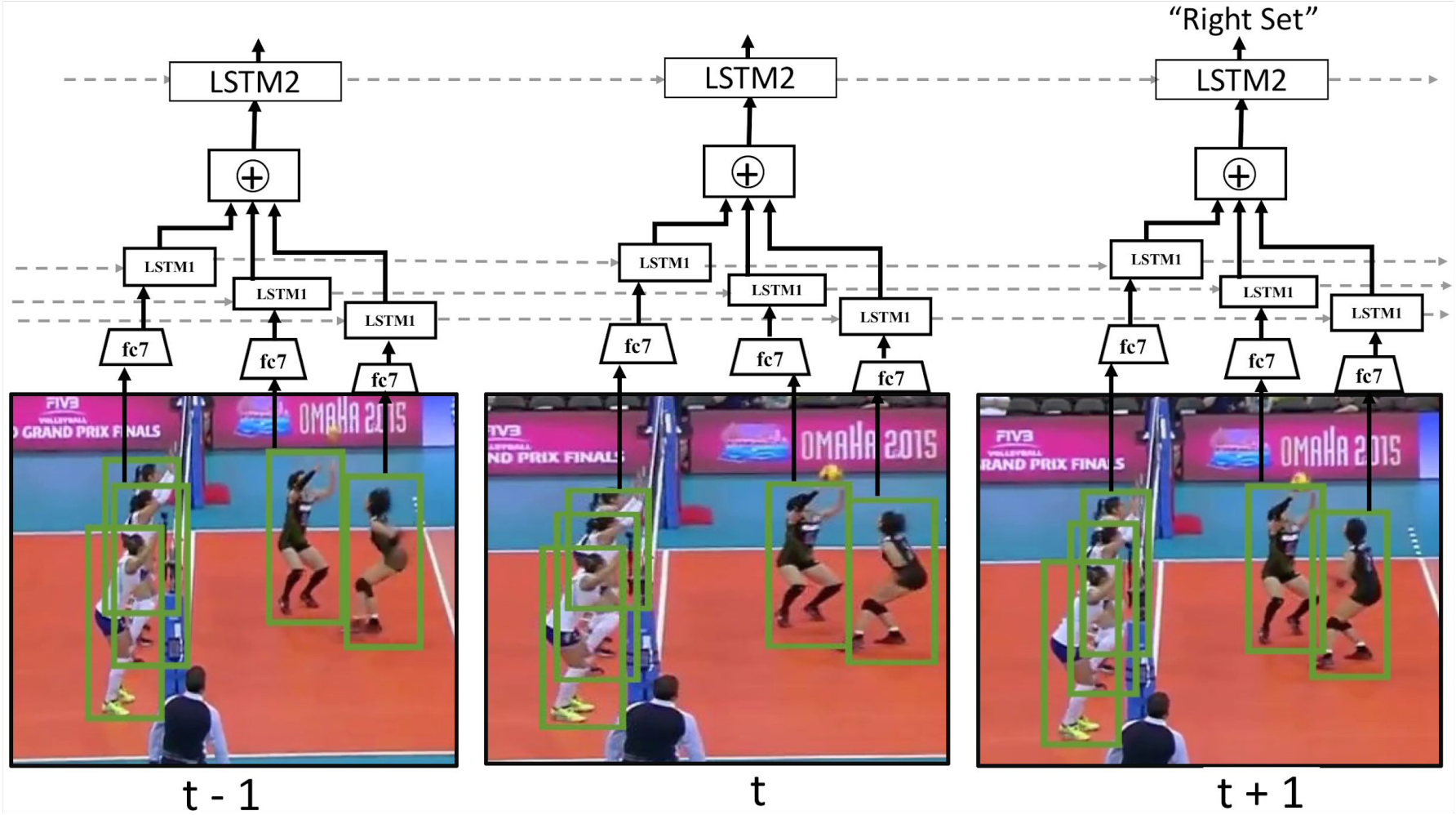
. . . .

$$P_{tk} = x_{tk} \oplus h_{tk}$$
$$Z_t = P_{t1} \diamond P_{t2} \ldots \diamond P_{tk}$$

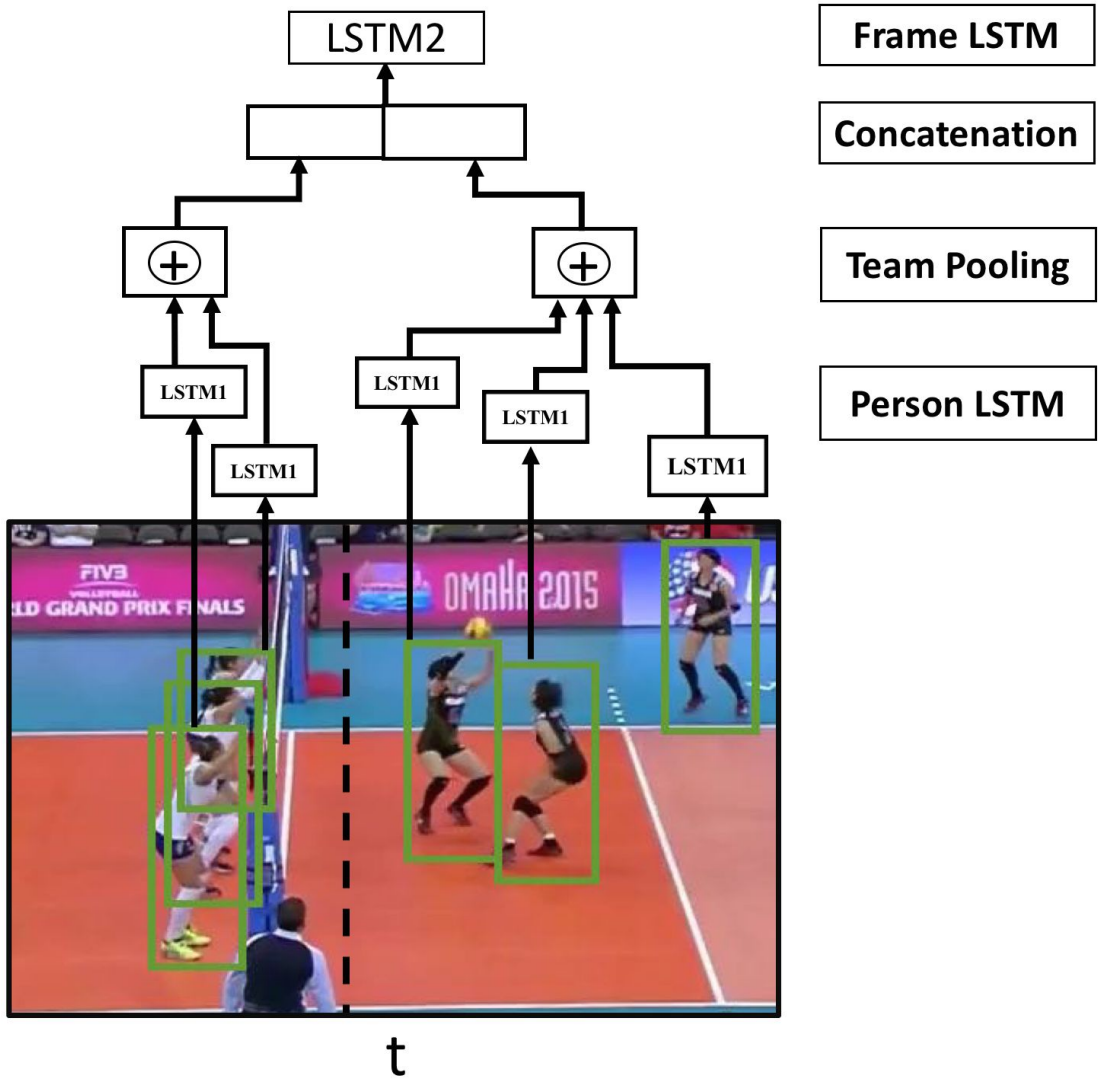
$\mathbf{P_{tk}}$ = kth Person representation

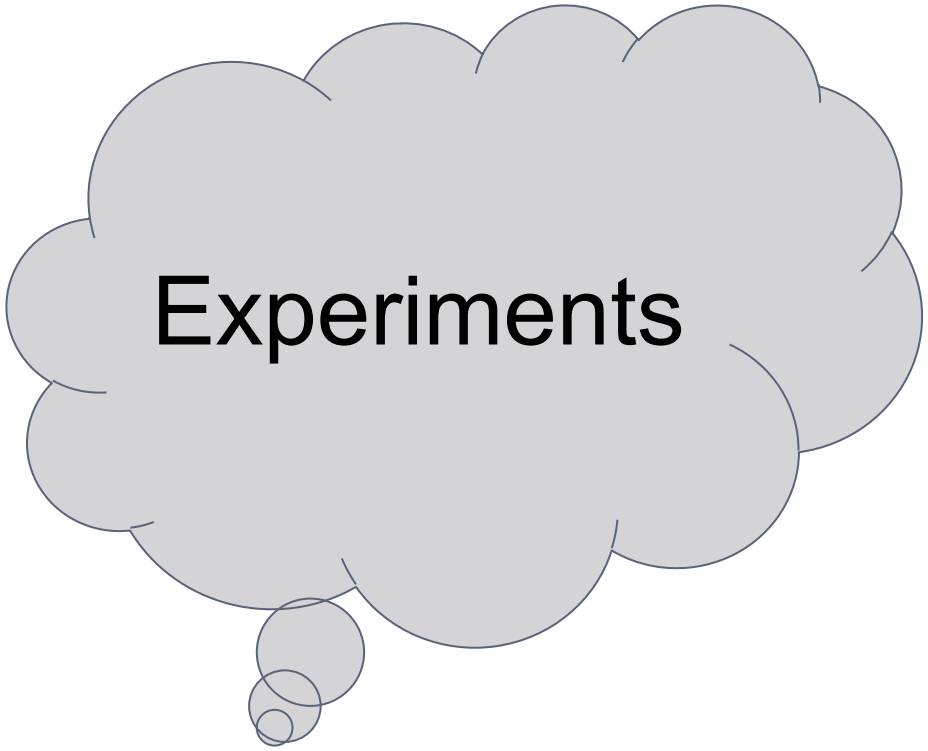
$\mathbf{Z_t}$ = Scene representation at time t

Overall model



Overall model - more spatial



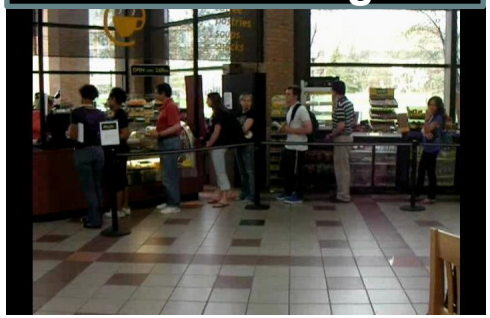

Experiments

Collective Activity Dataset

- Same label set for people and group activities
- 1925 video clips for training, 638 clips for test

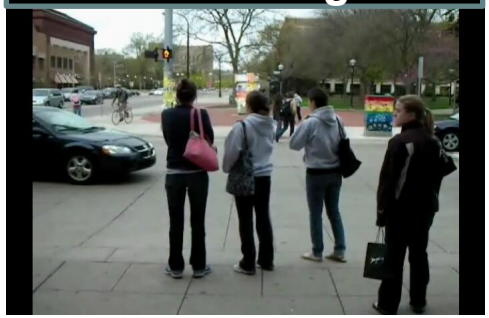
1. Crossing

2. Queueing



3. Talking

4. Waiting



5. Walking

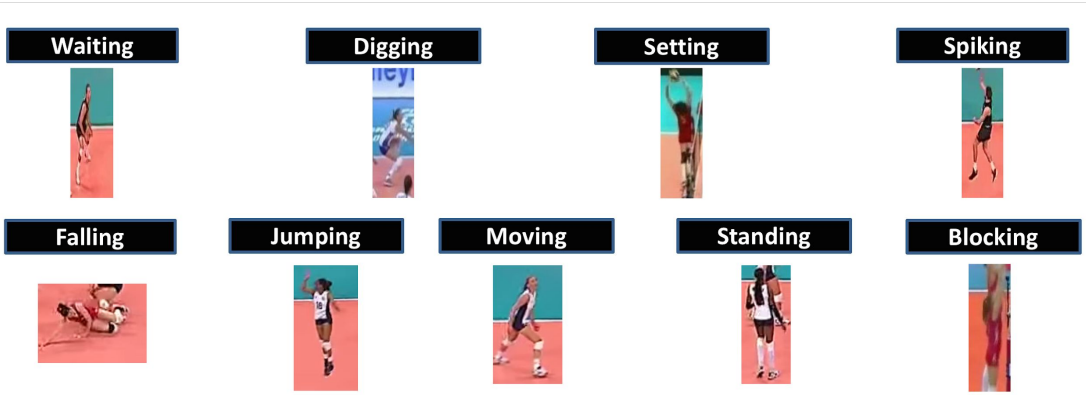
Collective Activity Dataset

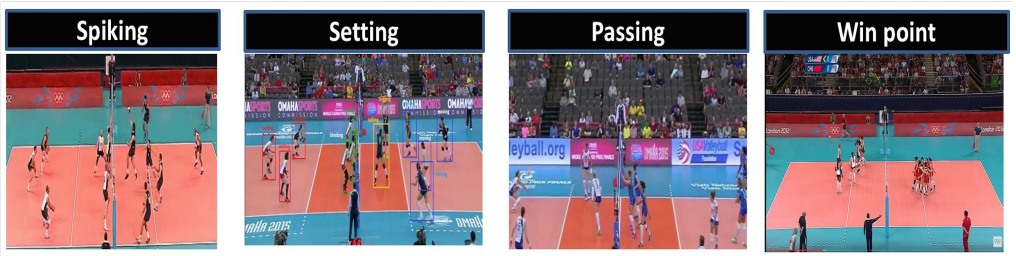
Method	Accuracy
B1-Image Classification	63.0
B2-Person Classification	61.8
B3-Fine-tuned Person Classification	66.3
B4-Temporal Model with Image Features	64.2
B5-Temporal Model with Person Features	64.0
B6-Two-stage Model without LSTM 1	70.1
B7-Two-stage Model without LSTM 2	76.8
Two-stage Hierarchical Model	**81.5**

Method	**Accuracy**
Contextual Model [Lan NIPS'10]	79.1
Deep Structured Model [Deng BMVC'15]	80.6
Our Model	81.5
Cardinality Kernel [Hajimirsadeghi CVPR'15]	**83.4**

New Volleyball Dataset

- 4830 annotated frames from 55 YouTube videos.
- 9 person level labels, and 8 group activity labels.





Left/right team variants

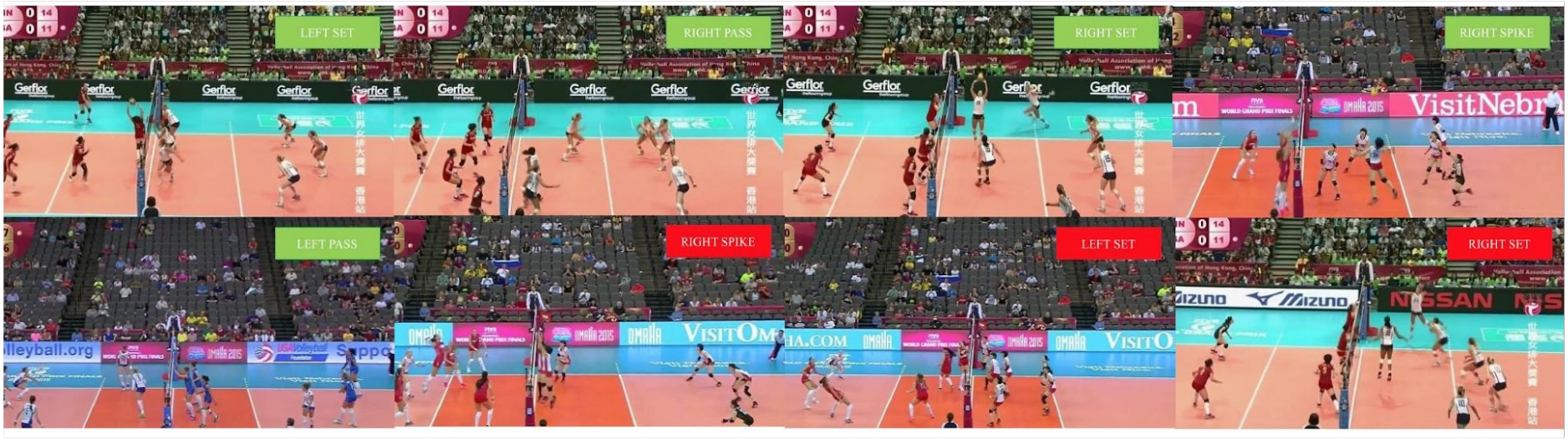
Volleyball Dataset

Method	Accuracy
B1-Image Classification	66.7
B2-Person Classification	64.6
B3-Fine-tuned Person Classification	68.1
B4-Temporal Model with Image Features	63.1
B5-Temporal Model with Person Features	67.6
B6-Two-stage Model without LSTM 1	74.7
B7-Two-stage Model without LSTM 2	80.2
Our Two-stage Hierarchical Model	**81.9**
IDTF (Improved Dense Trajectories)	73.4
IDTF - 1 group-box trajectories	71.7
IDTF - 2 groups-box trajectories	78.7

Spatial Model

Non Spatial Model

Volleyball Dataset: Success/Failure



Summary

- A two stage hierarchical model for group activity recognition
- LSTMs as a highly effective temporal model and temporal feature source
- Decent people-relation modeling with simple pooling
- Code & Dataset Link