SFU
Simon Fraser University

# Hierarchical Deep Temporal Models for Group Activity Recognition

Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, Greg Mori

Vision and Media Lab

IEEE 2016 Conference on Computer Vision and Pattern Recognition

CVPR 2016

## PROBLEM

Perform group activity recognition in two different contexts: Surveillance and Volleyball videos.

## CHALLENGES

1. Definition of group activity is **context dependent**. In surveillance video, it is based on what the majority of the people are doing, while in volleyball video, it is defined by what a few key players are doing.
2. It is a high level description of a video – Group activity is function of each individual activity.

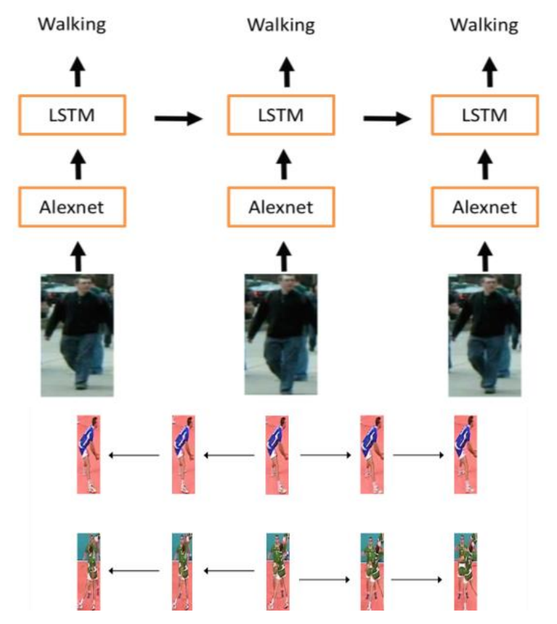
Group Activity = Majority's Activity
Group Activity = Key Player's Activity
Waiting
Group Activity – Right spike

## CONTRIBUTION

A novel deep temporal architecture that models group activities in a principled structured spatio-temporal framework based on modeling individual activities.

## STEP 1: BUILD PERSON REPRESENTATIONS

- **Target**: Build a spatio-temporal representation for every person's activity.
- Track a manually annotated bounding box for each person for a fixed temporal window around the target frame.
- 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

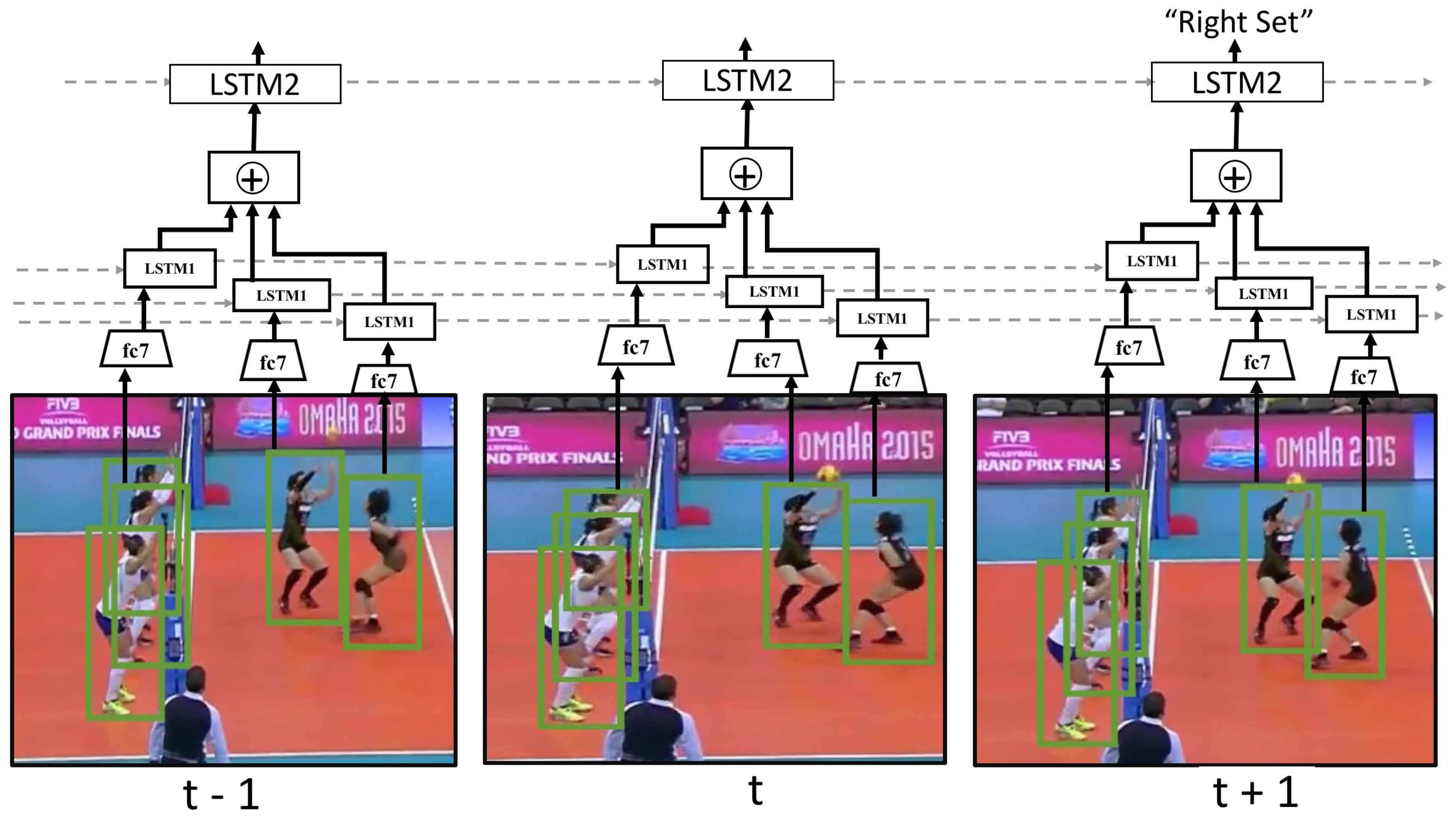


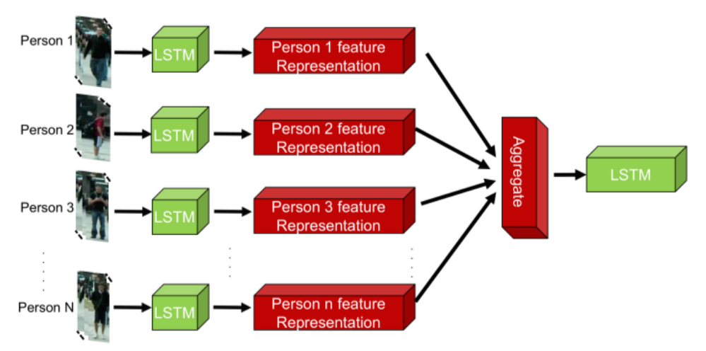
## STEP 2: BUILD GROUP ACTIVITY REPRESENTATION

- **Target**: Build a spatio-temporal representation for the group activity of a given frame
- Aggregate all individual person representations for every temporal step.
- For aggregation, 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 softmax classifier on top of the group activity representation.

## OVERALL MODELS

### Non-spatial Model




Person 1
Person 2
Person 3
Person N
LSTM
LSTM
LSTM
LSTM
Person 1 feature Representation
Person 2 feature Representation
Person 3 feature Representation
Person n feature Representation
Aggregate
LSTM

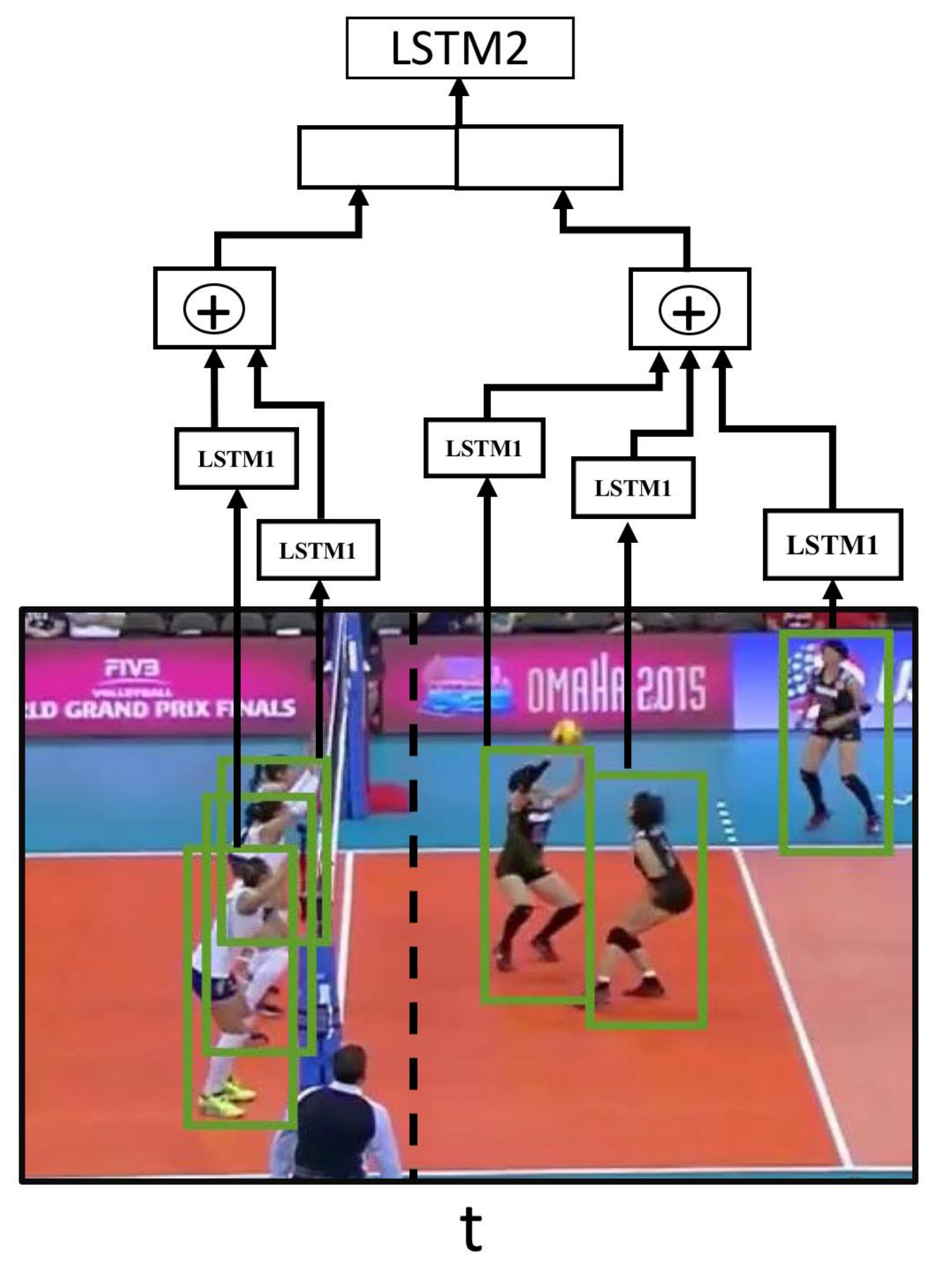
### PERSON LEVEL LSTMs

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

### Frame LEVEL LSTMs

$$
\begin{aligned}
P_{tk} &= x_{tk} \oplus h_{tk} \\
Z_t &= P_{t1} \diamond P_{t2} \ldots \diamond P_{tk}
\end{aligned}
$$

### Spatial Model



Frame LSTM

Concatenation

Team Pooling

Person LSTM

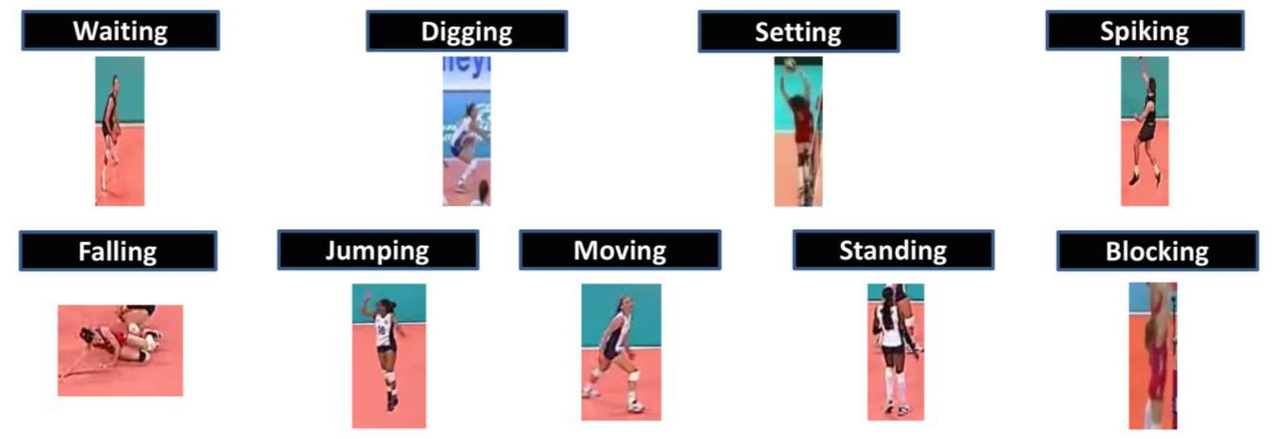
### Frame LEVEL LSTMs

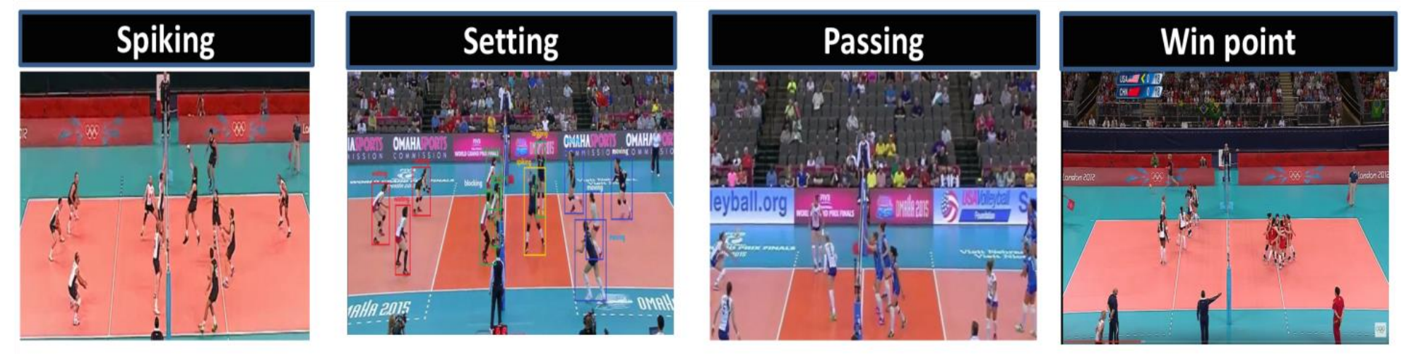
$$
\begin{aligned}
P_{tk} &= x_{tk} \oplus h_{tk} \\
S_m &= (m-1) * k/d + 1 \\
E_m &= m * k/d \\
G_{tm} &= P_{tSm} \diamond P_{t(Sm+1)} \ldots \diamond P_{tEm} \\
Z_t &= G_{t1} \oplus G_{t2} \ldots \oplus G_{td}
\end{aligned}
$$

- **Problem**: Pooling all players' representation in one representation reduces the model capabilities
- **Proposal**: players are split to several sub-groups and recognize the team activity based on the concatenation of each sub-group's representation.

## NEW VOLLEYBALL DATASET

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





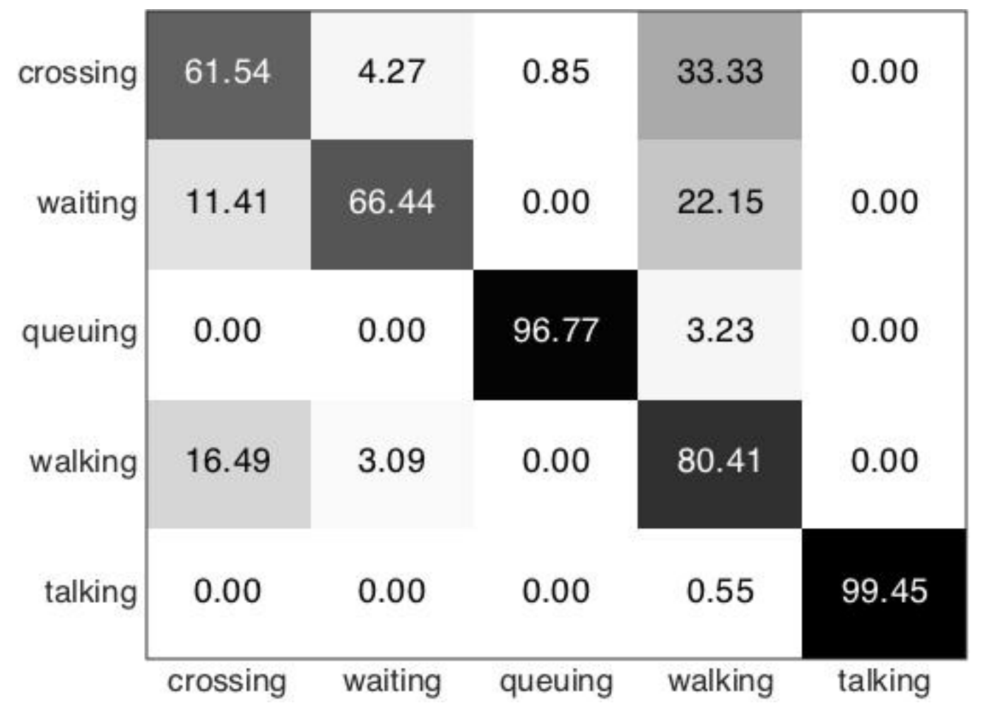
Left/right team variants

## EXPERIMENTS

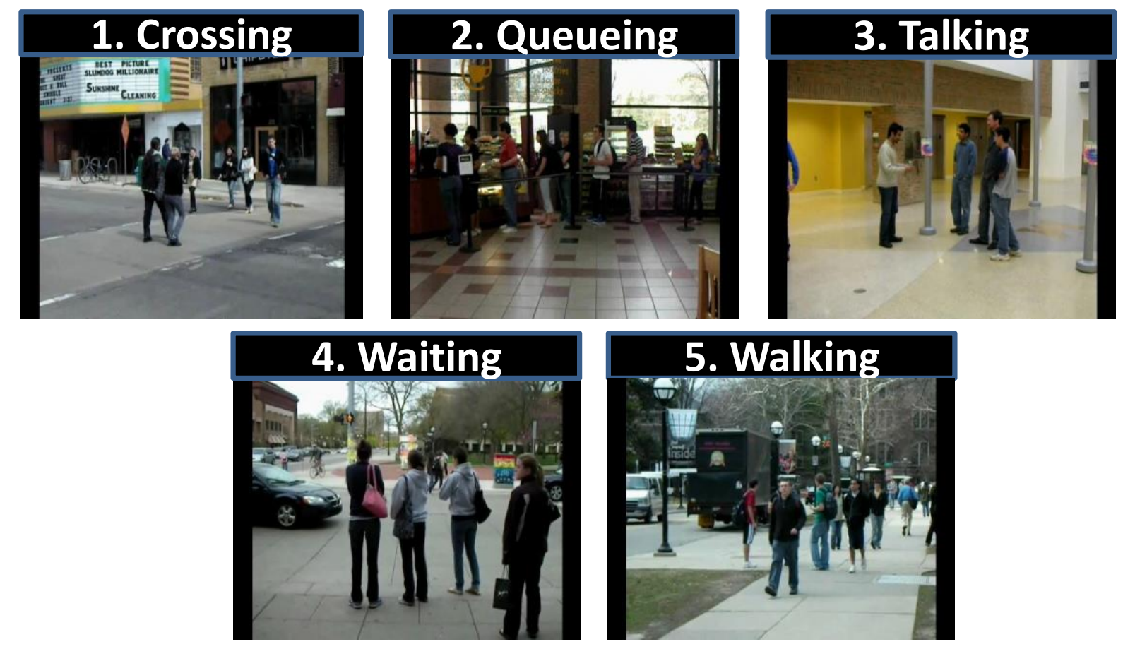
### Collective Activity Dataset

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

| 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** |

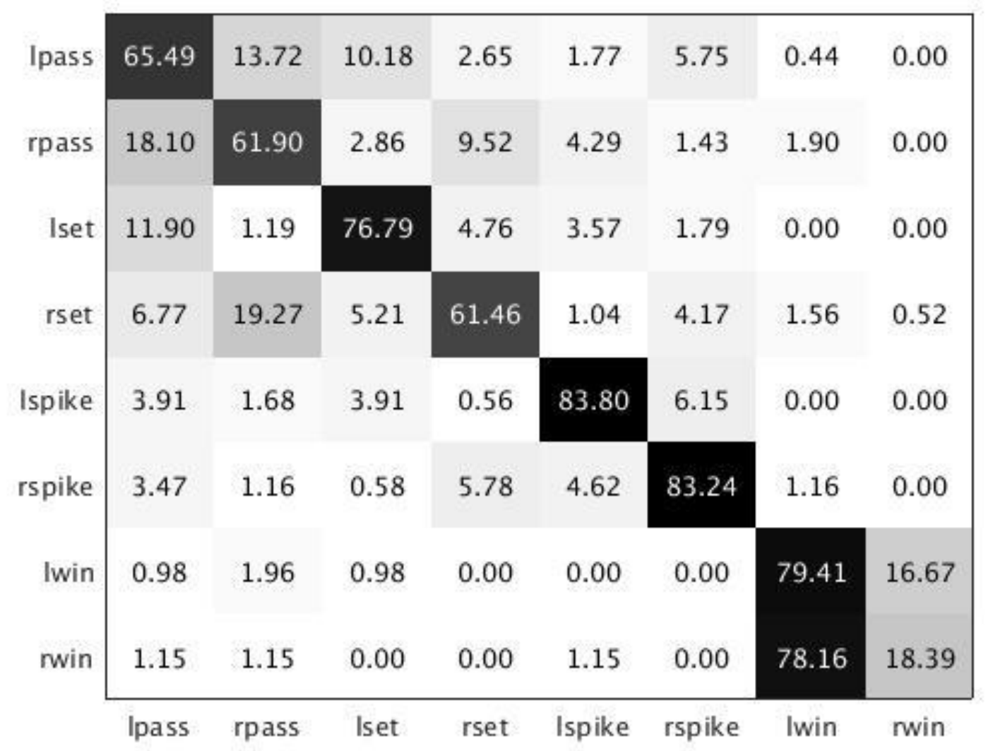


| | crossing | waiting | queuing | walking | talking |
|---|---|---|---|---|---|
| crossing | 61.54 | 4.27 | 0.85 | 33.33 | 0.00 |
| waiting | 11.41 | 66.44 | 0.00 | 22.15 | 0.00 |
| queuing | 0.00 | 0.00 | 96.77 | 3.23 | 0.00 |
| walking | 16.49 | 3.09 | 0.00 | 80.41 | 0.00 |
| talking | 0.00 | 0.00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

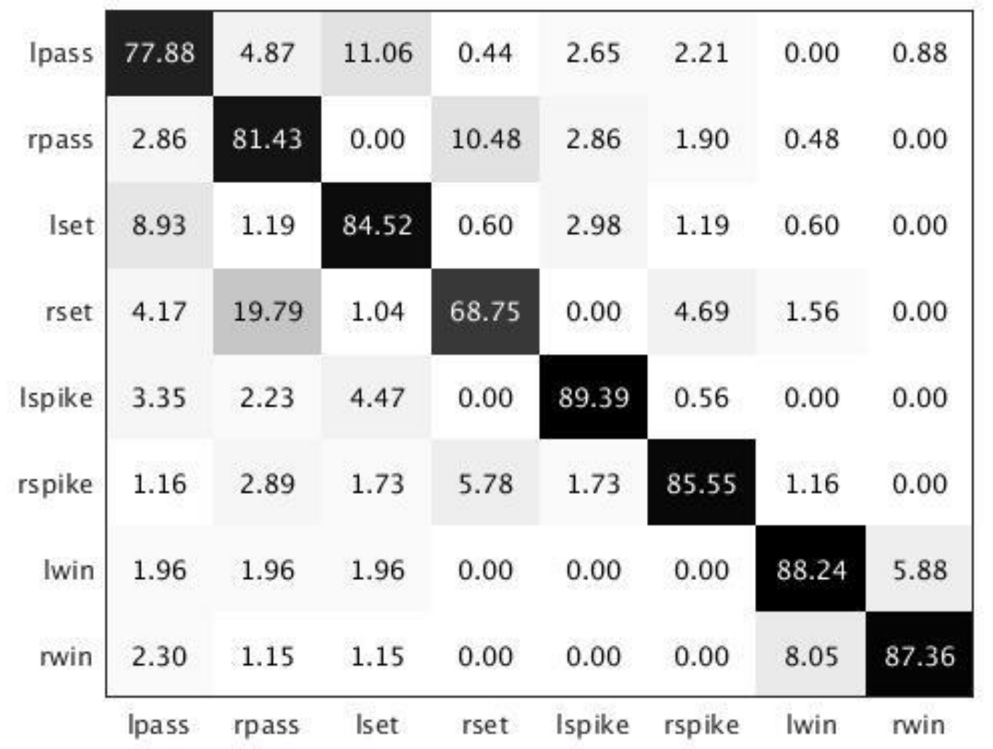


### Volleyball Dataset

| 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 |



| | lpass | rpass | lset | rset | lspike | rspike | lwin | rwin |
|---|---|---|---|---|---|---|---|---|
| lpass | 65.49 | 13.72 | 10.18 | 2.65 | 1.77 | 5.75 | 0.44 | 0.00 |
| rpass | 18.10 | 61.90 | 2.86 | 9.52 | 4.29 | 1.43 | 1.90 | 0.00 |
| lset | 11.90 | 1.19 | 76.79 | 4.76 | 3.57 | 1.79 | 0.00 | 0.00 |
| rset | 6.77 | 19.27 | 5.21 | 61.46 | 1.04 | 4.17 | 1.56 | 0.52 |
| lspike | 3.91 | 1.68 | 3.91 | 0.56 | 83.80 | 6.15 | 0.00 | 0.00 |
| rspike | 3.47 | 1.16 | 0.58 | 5.78 | 4.62 | 83.24 | 1.16 | 0.00 |
| lwin | 0.98 | 1.96 | 0.98 | 0.00 | 0.00 | 0.00 | 79.41 | 16.67 |
| rwin | 1.15 | 1.15 | 0.00 | 0.00 | 1.15 | 0.00 | 78.16 | 18.39 |



| | lpass | rpass | lset | rset | lspike | rspike | lwin | rwin |
|---|---|---|---|---|---|---|---|---|
| lpass | 77.88 | 4.87 | 11.06 | 0.44 | 2.65 | 2.21 | 0.00 | 0.88 |
| rpass | 2.86 | 81.43 | 0.00 | 10.48 | 2.86 | 1.90 | 0.48 | 0.00 |
| lset | 8.93 | 1.19 | 84.52 | 0.60 | 2.98 | 1.19 | 0.60 | 0.00 |
| rset | 4.17 | 19.79 | 1.04 | 68.75 | 0.00 | 4.69 | 1.56 | 0.00 |
| lspike | 3.35 | 2.23 | 4.47 | 0.00 | 89.39 | 0.56 | 0.00 | 0.00 |
| rspike | 1.16 | 2.89 | 1.73 | 5.78 | 1.73 | 85.55 | 1.16 | 0.00 |
| lwin | 1.96 | 1.96 | 1.96 | 0.00 | 0.00 | 0.00 | 88.24 | 5.88 |
| rwin | 2.30 | 1.15 | 1.15 | 0.00 | 0.00 | 0.00 | 8.05 | 87.36 |

## 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**: https://github.com/mostafa-saad/deep-activity-rec

## ACKNOWLEDGMENT

This work was supported by grants from NSERC and Disney Research.