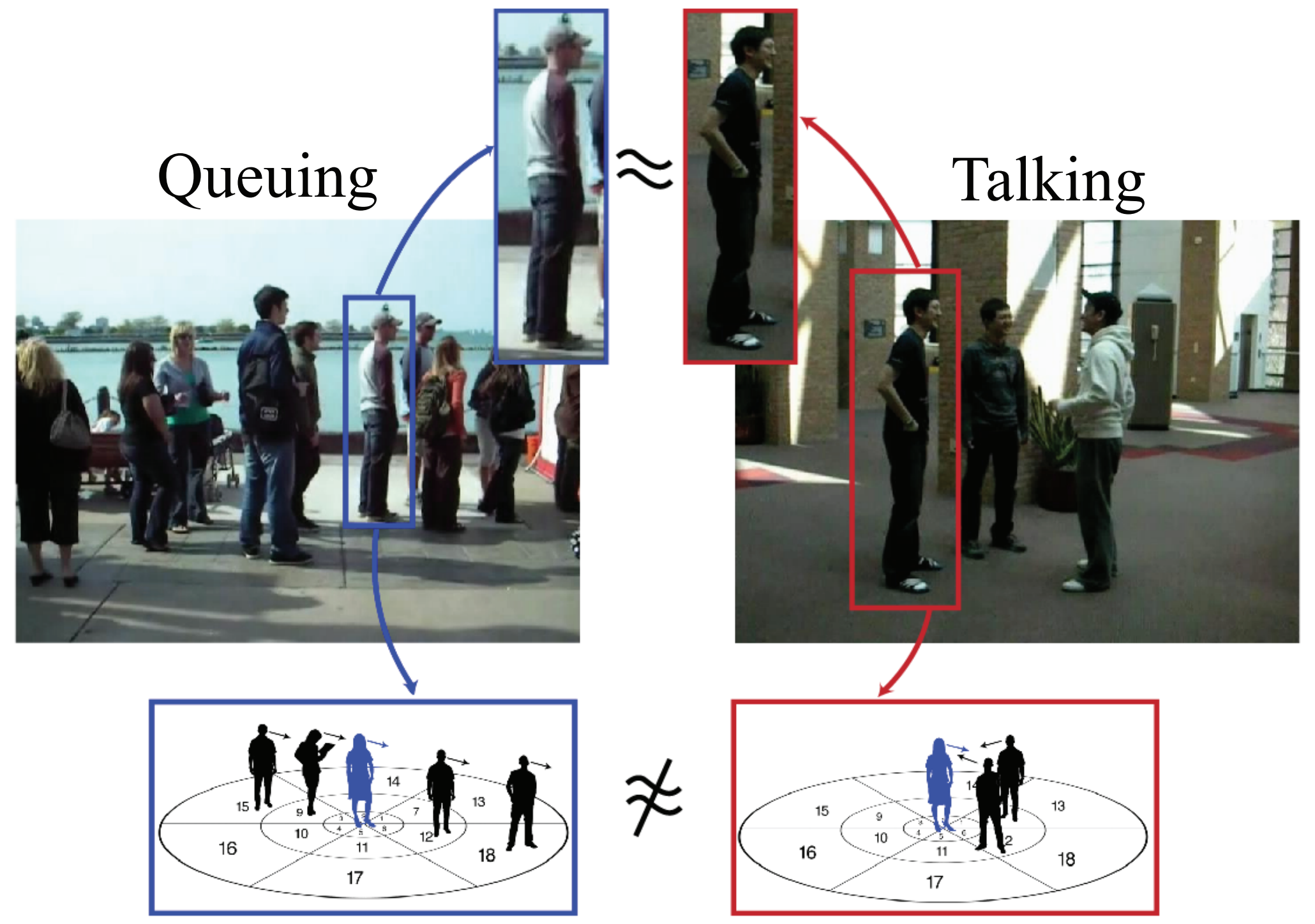
Learning Context for Collective Activity Recognition

Wongun Choi Khuram Shahid Silvio Savarese

Department of Electirical and Computer Engineering, University of Michigan, Ann Arbor MI

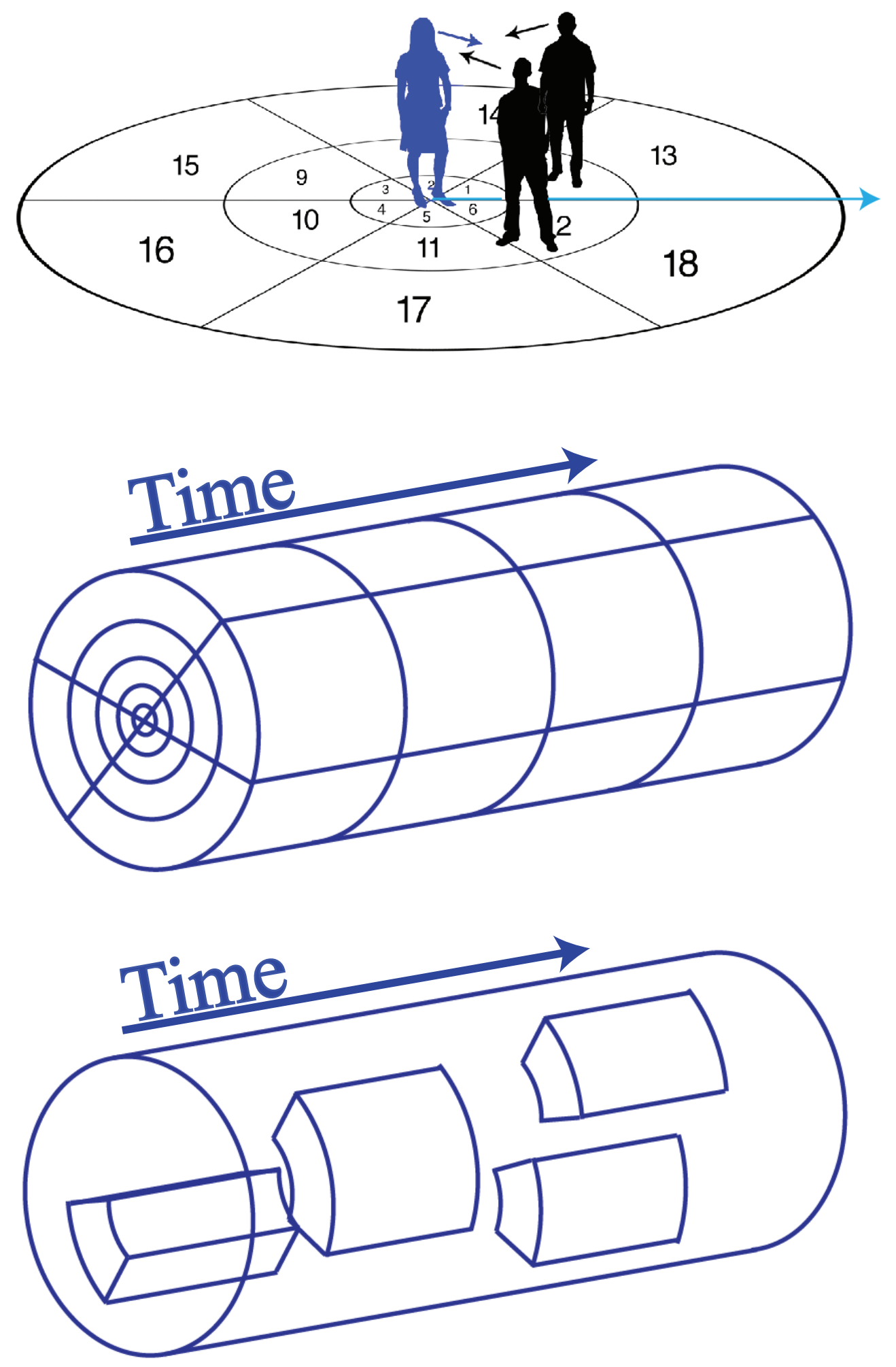
Introduction

- Collective activity
  - Activities that are defined by the interaction among people.
  - Cannot be characterized by single person's appearance.
- Crowd Context
  - Spatio-temporal context around one person.
  - Data-driven approach to learn the crowd context.
- Regularization by spatio-temporal consistency.



Crowd Context

- STV [1]
  - Appropriate for capturing spatio-temporal relationship.
  - Rigid structured descriptor
  - Susceptible to clutter
  - Several parameters to be tuned
- RSTV
  - Randomize the discretization in feature space.
  - Parameter free.
  - Structure learned from training data.
  - More flexible structure.
  - Higher scalablility.
  - Robustness under clutter.

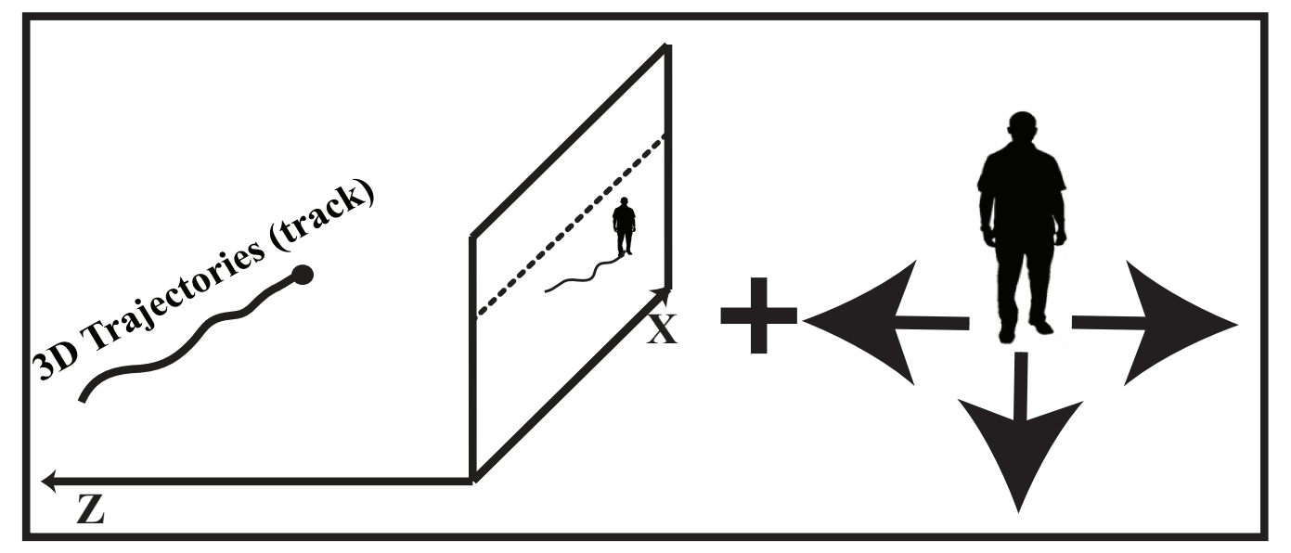


Acknowledgement

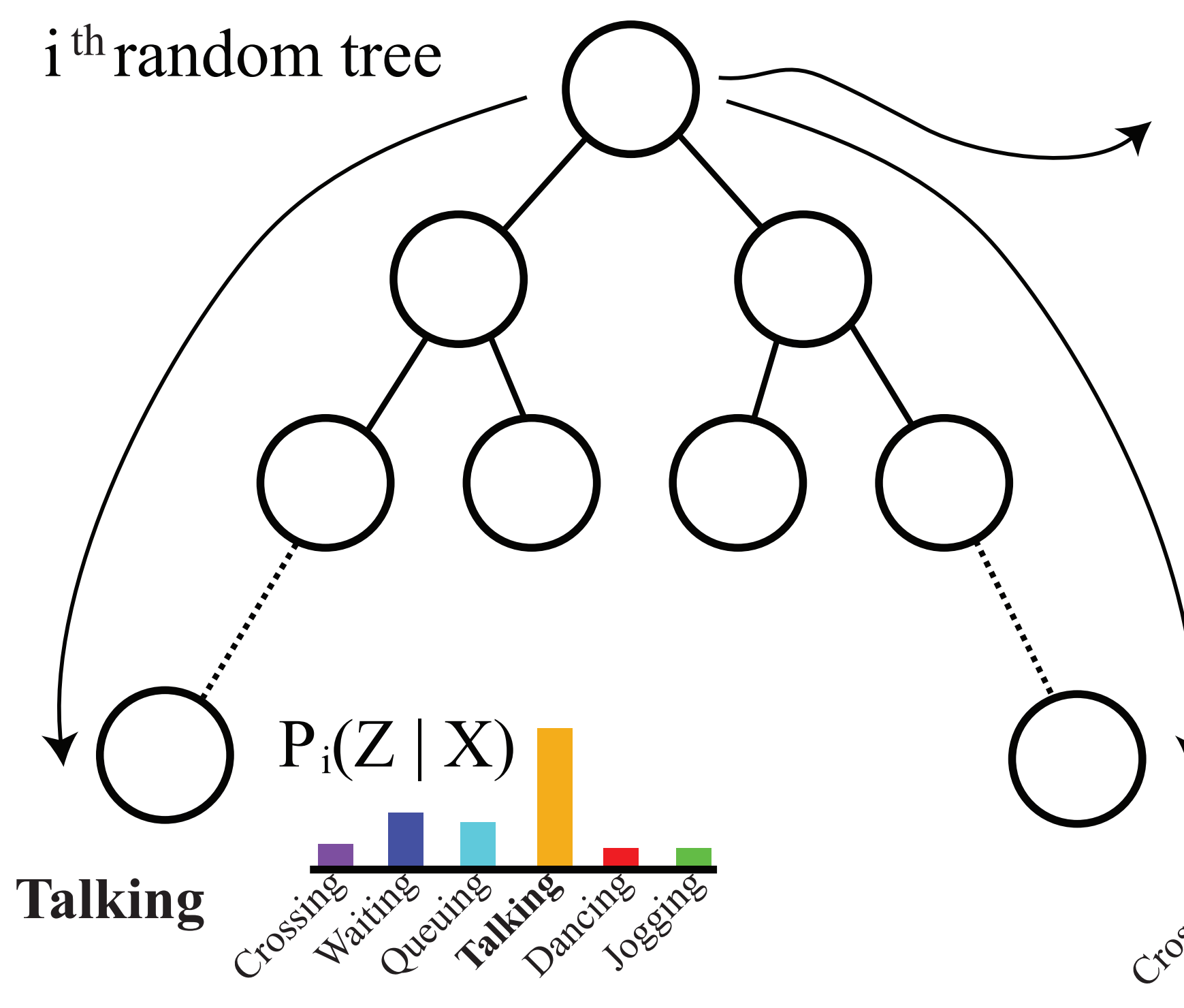
This work is supported by NSF EAGER (award #1052762)

Method

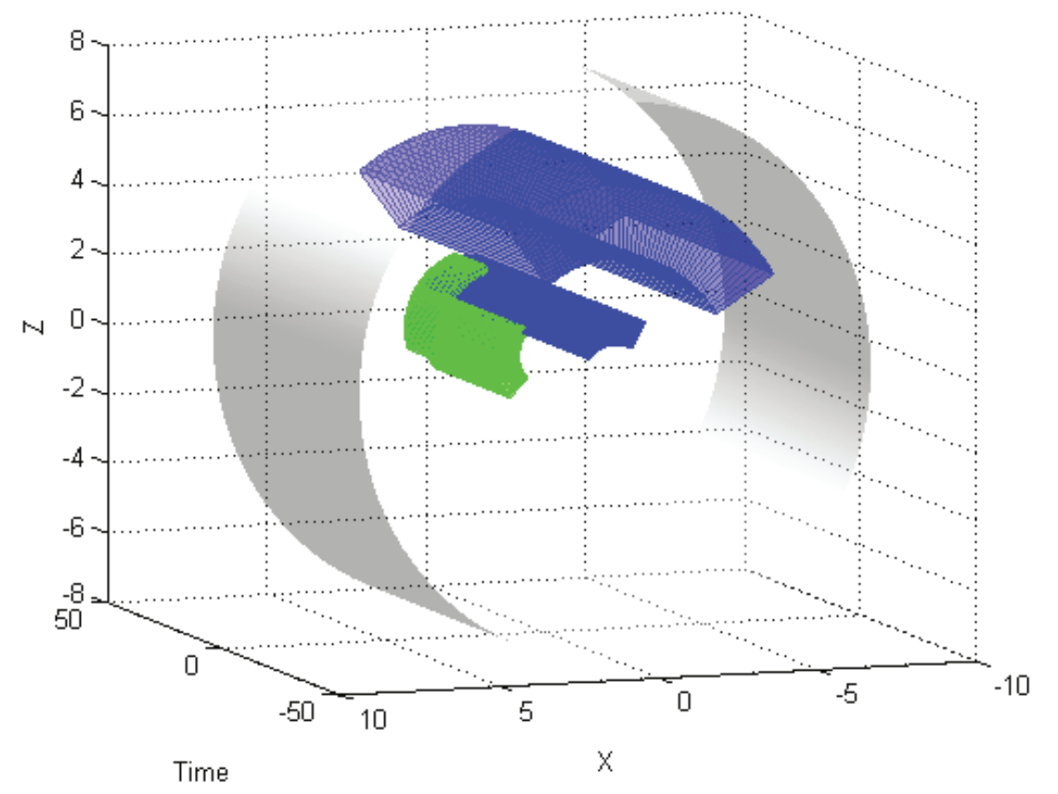
- Inputs for algorithm
  - 3D trajectories + velocity of people
  - N-directional pose classification



- RSTV learning
  - Divide training set into N random subset (bagging)
  - Train a tree with a subset of training data

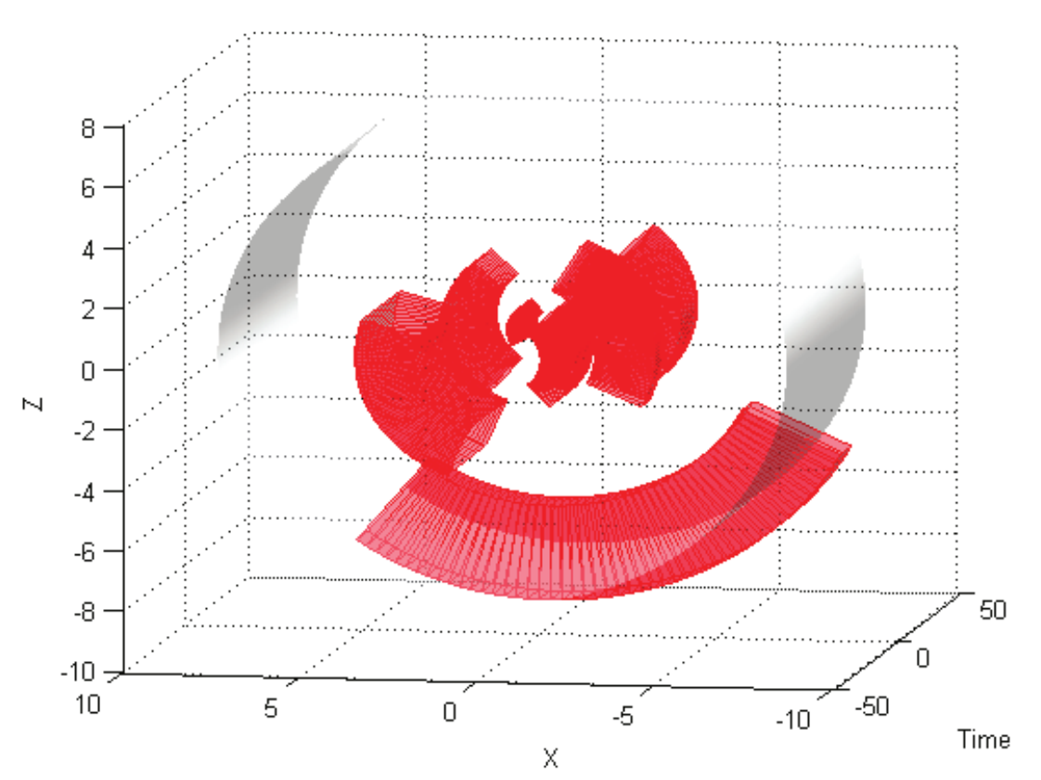


Randomly choose a set of features
- Spatial location (r, θ)
- Temporal support (t, Δt)
- Pose direction (p)
- Threshold (t)



Features in each path

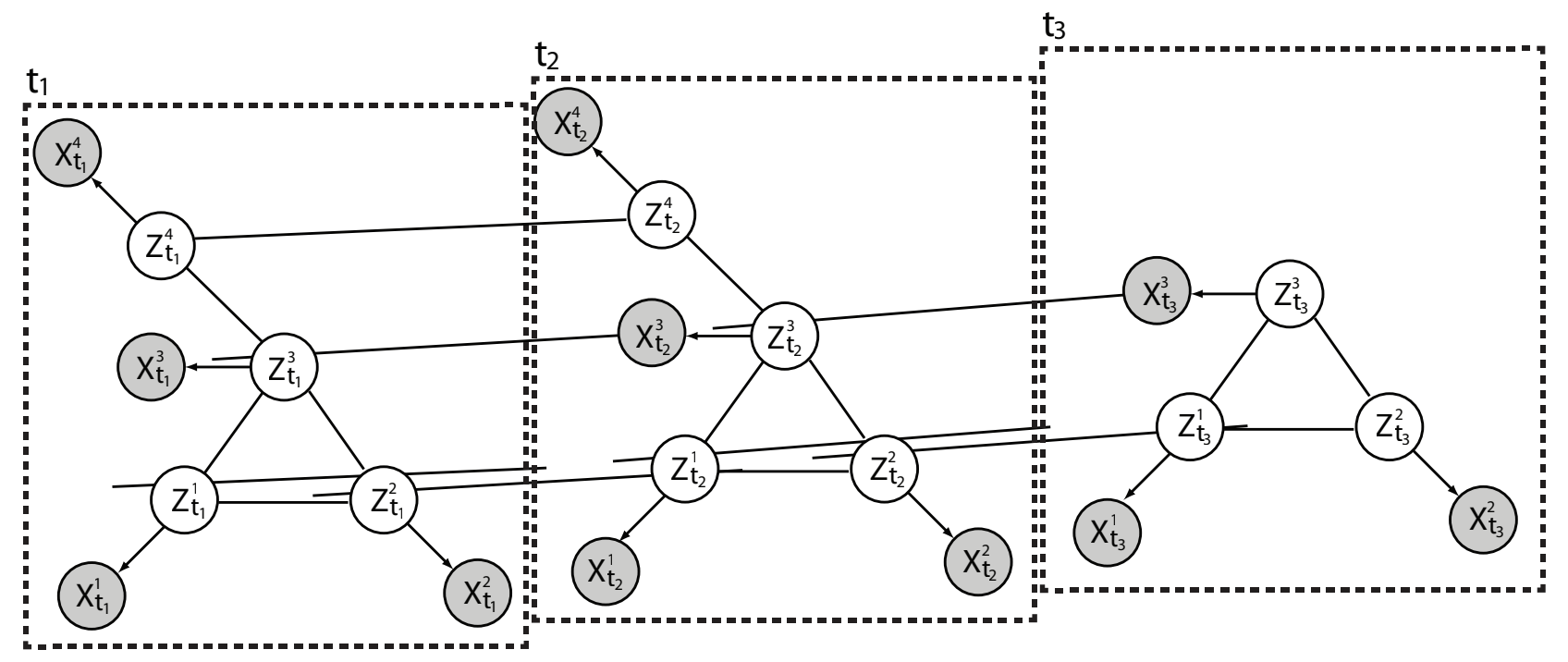
Each color represents a pose
- blue : facing down
- green : facing right
- red : facing up



- Classification
  - For each tree, find the leaf node for each example.
  - Compute sum of posterior probability

$$\frac{1}{N}\sum_{i=1}^{N} P_i(Z|X)$$

- MRF regularization

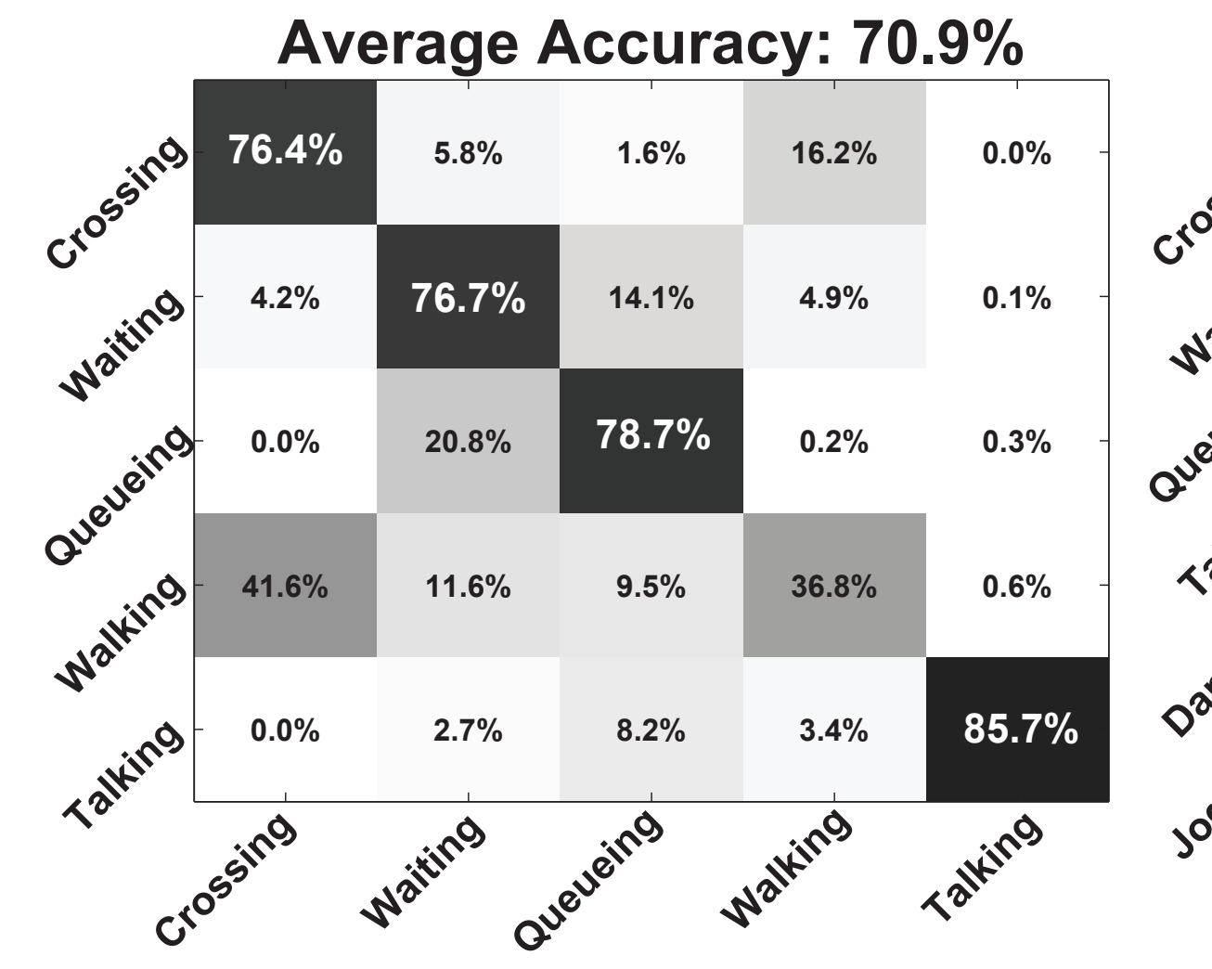


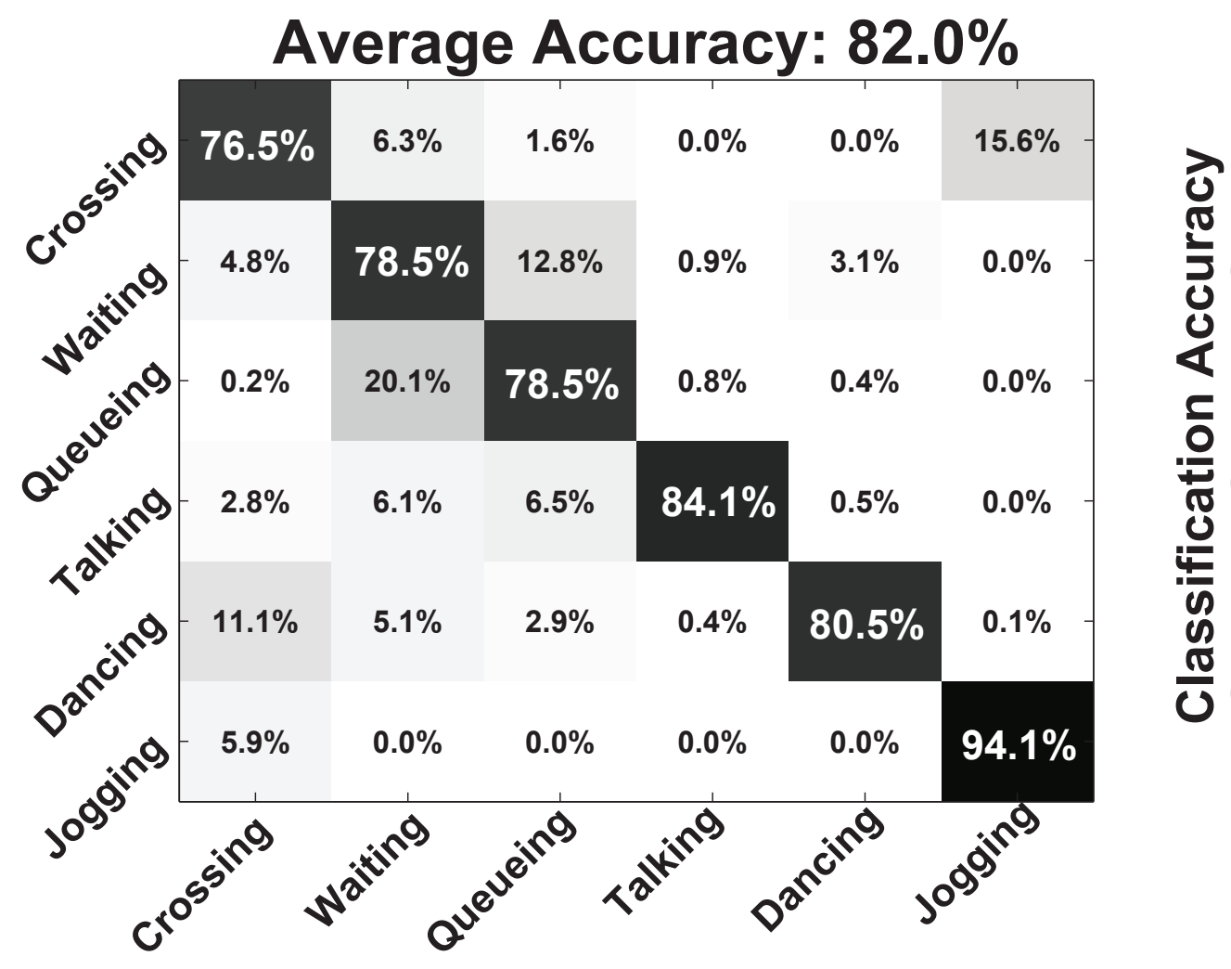
- Spatio-temporal context
- Inference by Gibbs sampling

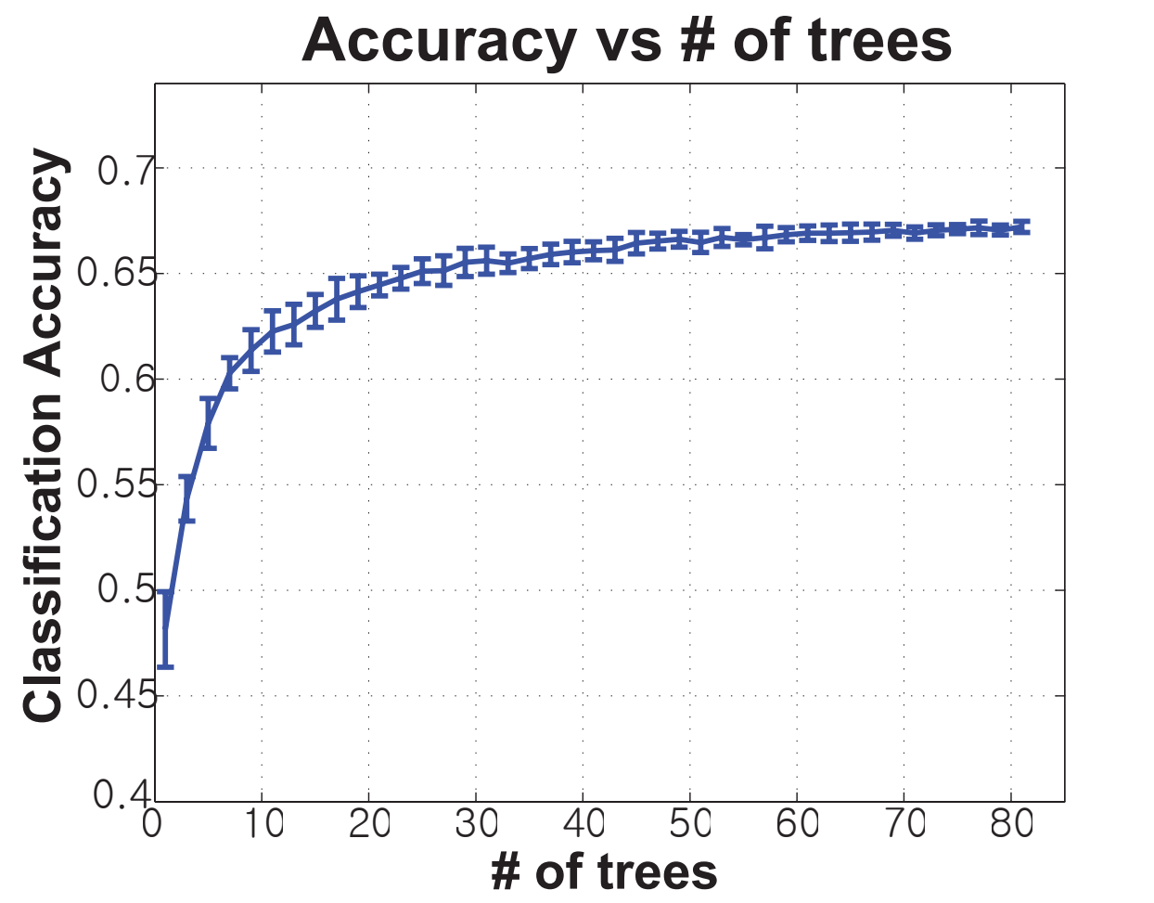
$$P(Z|X,p) \propto \prod_t \prod_i P(Z_t^i|X_t^i) \prod_t \prod_{(i,j)\in E_s} \Phi_S(Z_t^i, Z_t^j; p_t^i, p_t^j) \prod_i \prod_t \Phi_T(Z_{t-1}^i, Z_t^i)$$

Results

- Classification Result

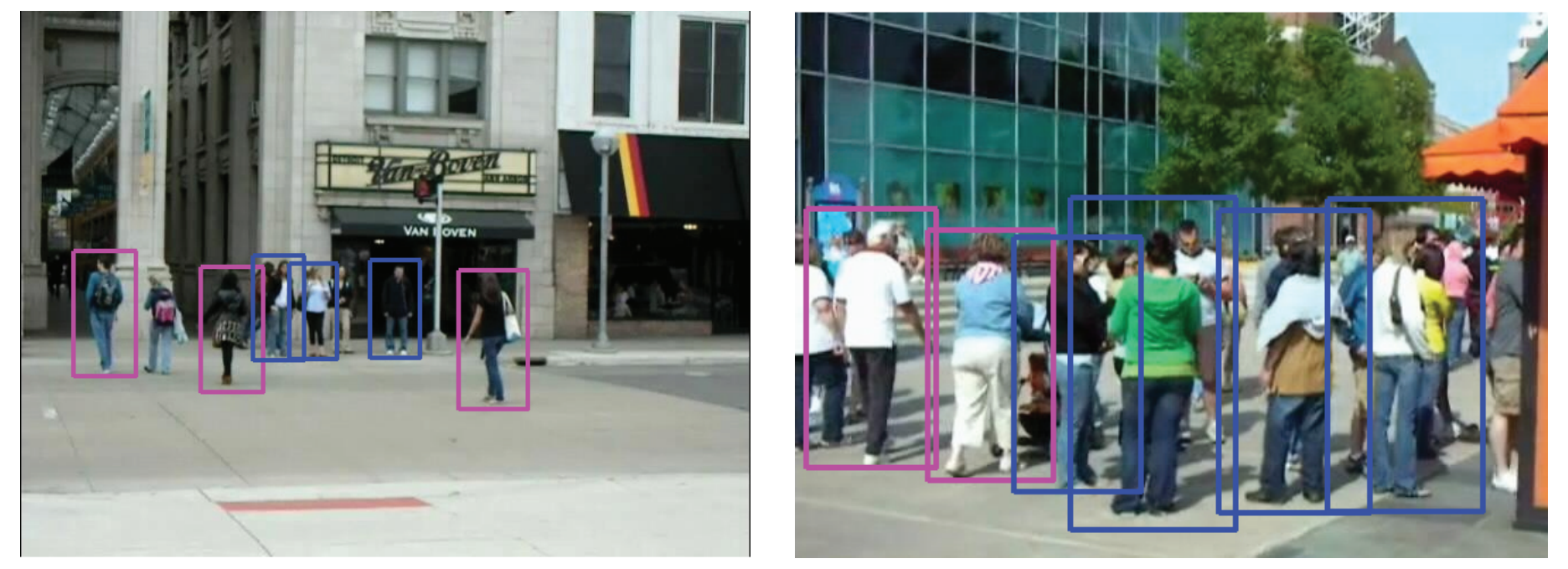




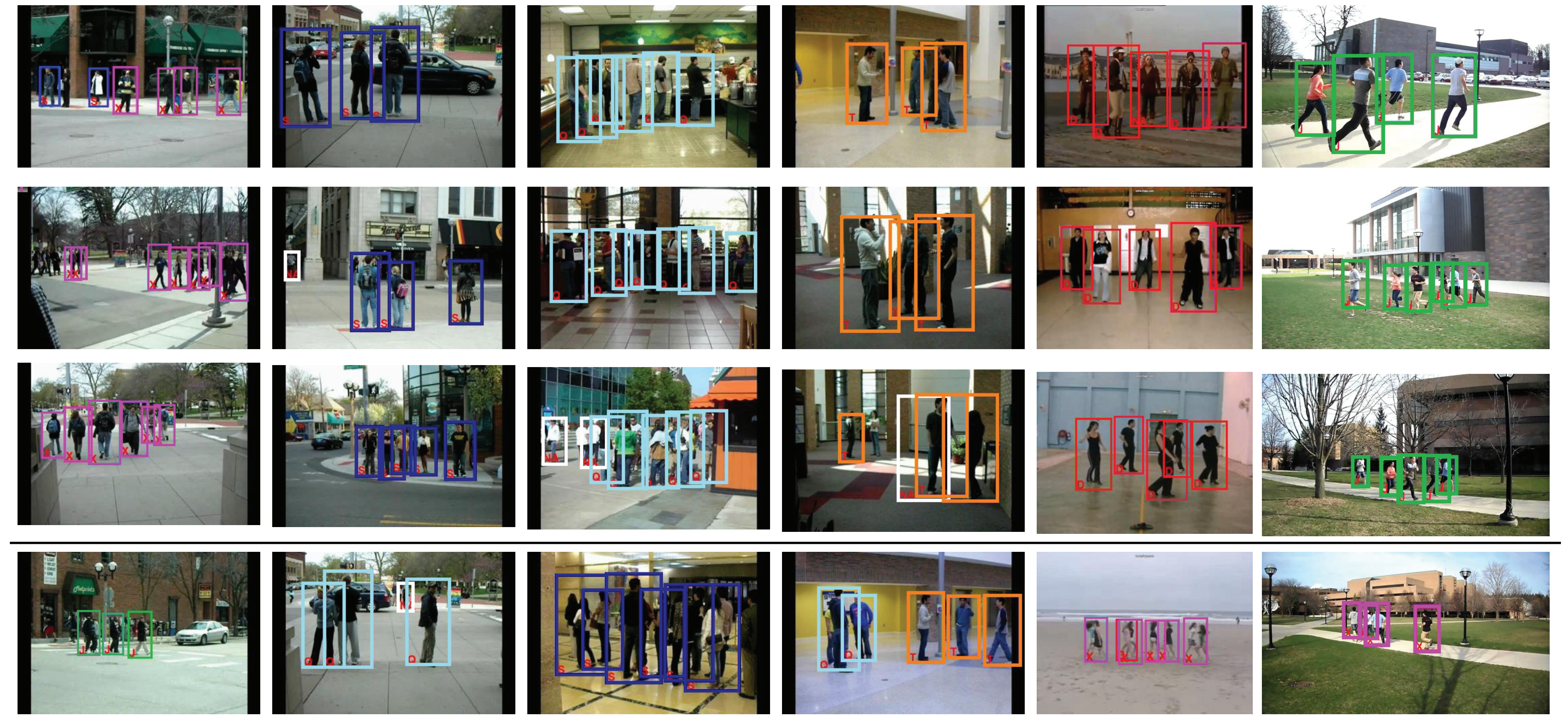


Dataset	5 Activities	6 Activities
AC [2]	68.2%	-
STV [1]	64.3%	-
STV+MC [1]	65.9%	-
STV + RF	64.4%	-
RSTV	67.2%	71.7%
RSTV + MRF	**70.9%**	**82.0%**

- Activity Segmentation



- Example Results



Conclusion

- RSTV enables more accurate classification results than state-of-the-art methods
- Capable of handling multiple activities in the scene.
- Enable segementation of individuals into different collective activities.

[1] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In Visual Surveillance Workshop, ICCV, 2009.
[2] T. Lan, Y. Wang, G. Mori, and S. Robinovitch. Retrieving actions in group contexts. In International Workshop on Sign Gesture Activity, 2010.