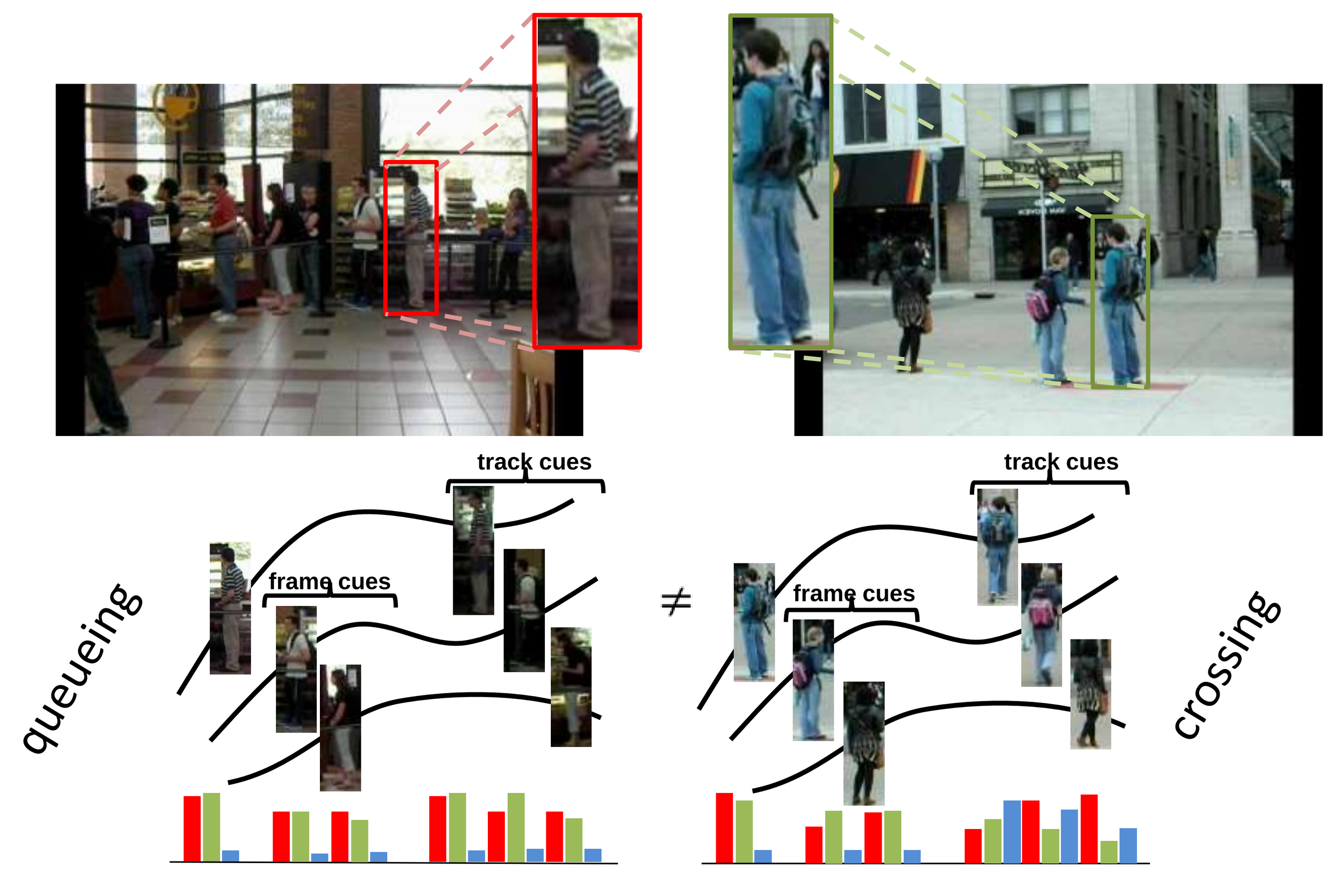
Combining Per-Frame and Per-Track Cues for Multi-Person Action Recognition

Sameh Khamis
Vlad I. Morariu
Larry S. Davis

The Motivation

- Recognizing human activity from pose and motion still subject to error due to appearance aliasing.



- Integrate tracking and scene context into action recognition to overcome this.
- Solve the coupled problem jointly!

The Approach

- Our goal is to formulate the problem as a tractable optimization function.
- The function should minimize
  - The action classification costs.
  - The per-track identity association costs.
  - The per-frame scene harmony costs.
- The action classification cost is based on the Action-Context (AC) descriptor [2] using HOG as the underlying representation.
- The identity association cost penalizes appearance and action transition inconsistencies.
  - Appearances are modeled by a distance matrix learned using LMNN [4] between the downsampled detection boxes as raw features.
  - Action transitions are modeled by a transition matrix learned by counting action pairs on the same track.
- The scene harmony cost is modeled by the joint likelihood of scene types and actions.
  - Scene types are approximated by the cluster centroids of K-means on the per-frame action histograms.
  - Scene prior is also estimated from the output of K-means.

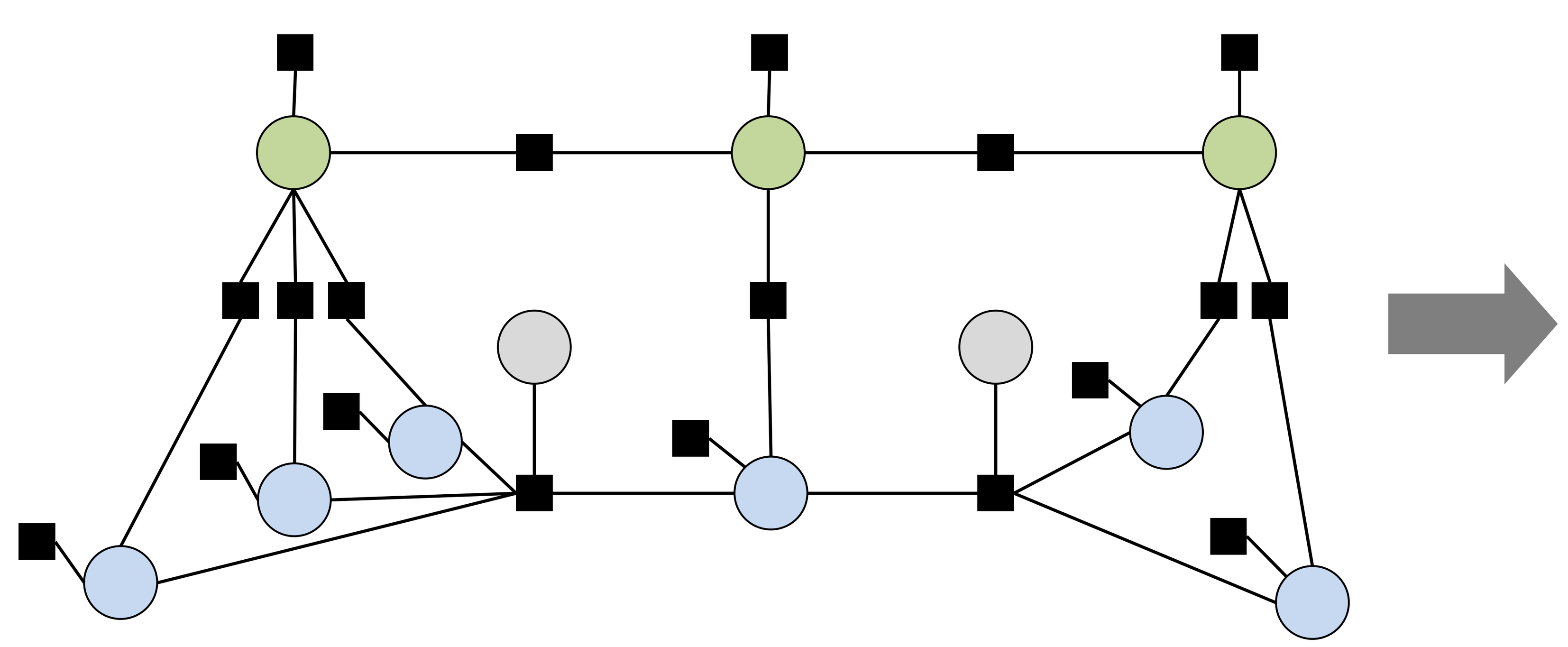
The Model

- Inference can be formulated as a linear program relaxation, but it is more advantageous to leverage the underlying structure of our model.
- As a function of (A)ctions, (S)cenes, and (I)dentities, our problem can be broken into two smaller and easier-to-solve subproblems.

$$\min_{A,S,I} F(A,S,I) = \min_{A,S,I} [F_1(A,S) + F_2(A,I)]$$

- Separate the subproblems by duplicating the "complicating" variables.

$$\min_{A_1,A_2,S,I} \quad F_1(A_1,S) + F_2(A_2,I)$$
$$s.t. \qquad A_1 = A_2$$

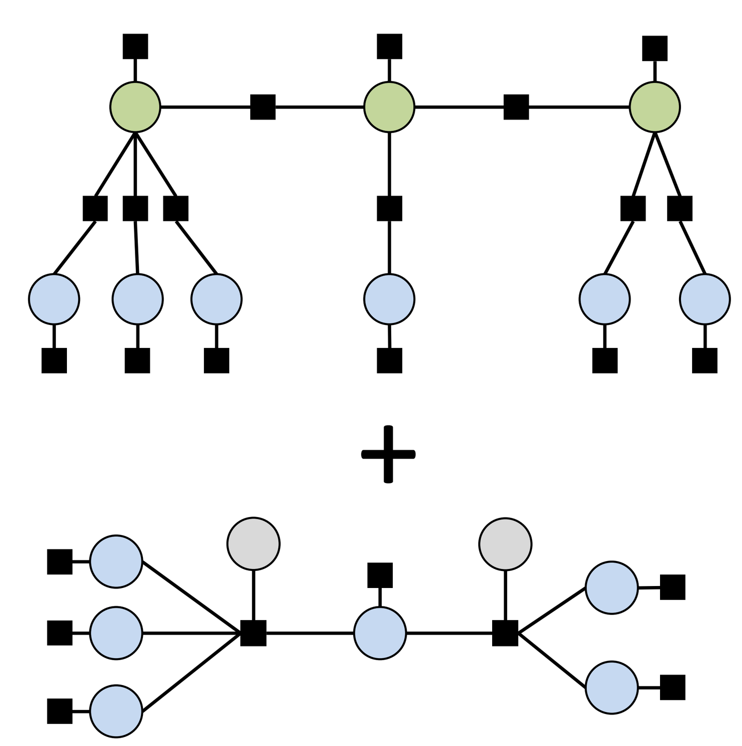


- Form the Lagrangian to reveal the separable but modified subproblems.

$$L(A_1,A_2,S,I,\nu) = F_1(A_1,S) + F_2(A_2,I) + \nu A_1 - \nu A_2$$

- Optimizing the Lagrangian by an iterative primal-dual approach tightens the bound on the optimal solution of the original problem.

$$\max_{\nu} L(A_1,A_2,S,I,\nu) =$$
$$\max_{\nu} \Big[\underbrace{\min_{A_1,S} [F_1(A_1,S) + \nu A_1]}_{\text{Belief Propagation}} + \underbrace{\min_{A_2,I} [F_2(A_2,I) - \nu A_2]}_{\text{Minimum Cost Flow}} \Big]$$



- A tree-structured pairwise graphical model
- Solves action recognition consistent with scene context
- Max-Product Belief Propagation is exact and efficient

\+

- A minimum cost flow problem based on [3, 5]
- Jointly solves action recognition and tracking
- Constraint matrix is totally unimodular, so a globally optimal integral solution exists

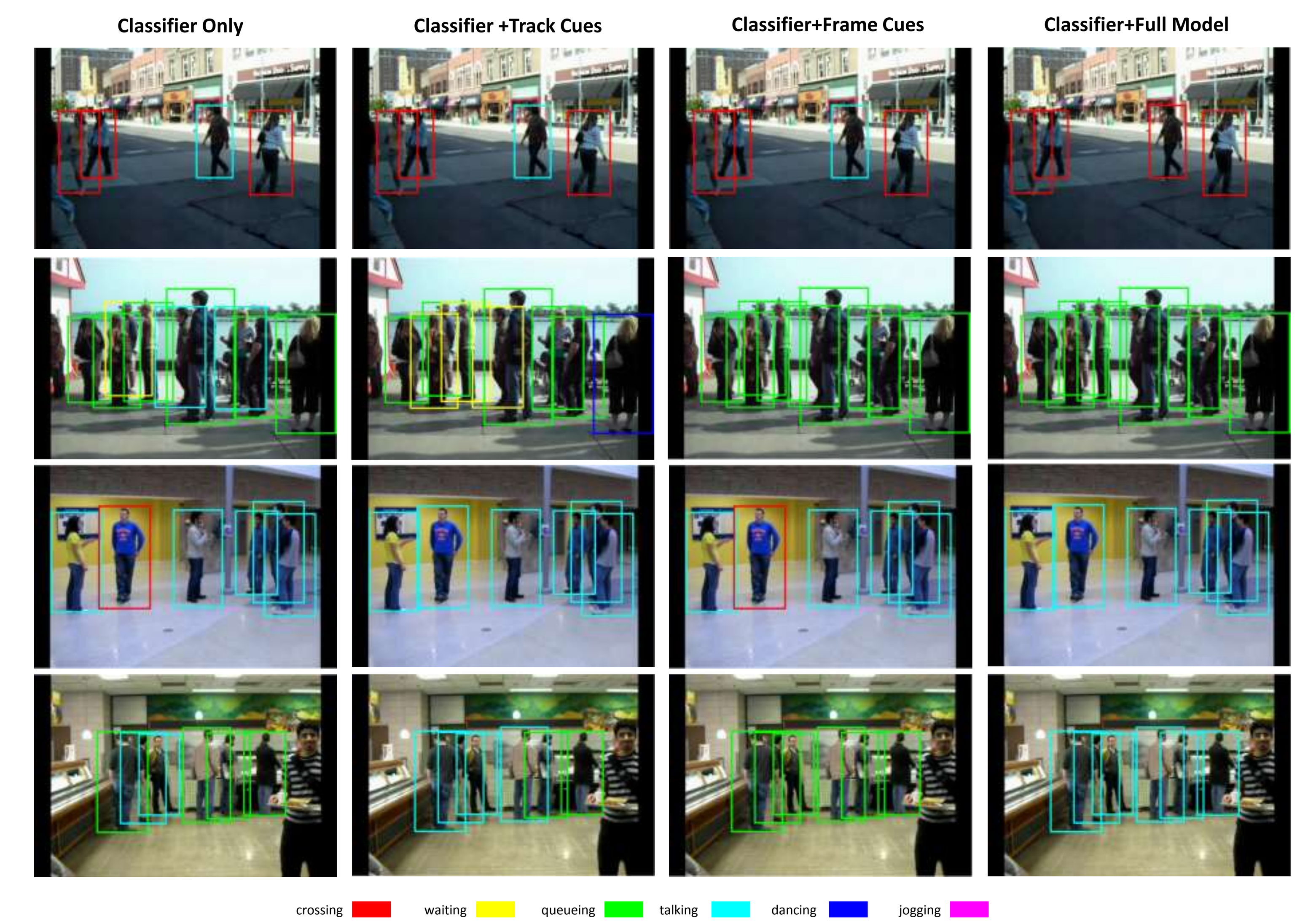
The Results

- We report results on two public multi-person action recognition datasets [1]

Approach / Dataset	5 Activities	6 Activities
Classifier Only	68.8%	81.5%
Classifier + Track Cues	70.9%	83.7%
Classifier + Frame Cues	70.7%	84.8%
Classifier + Full Model	**72.0%**	**85.8%**

	crossing	waiting	queueing	walking	talking
crossing	67.2%	5.3%	0.9%	19.3%	2.2%
waiting	2.9%	56.8%	13.1%	10.4%	0.8%
queueing	4.7%	29.4%	81.1%	3.5%	0.5%
walking	24.6%	5.9%	0.8%	61.5%	3.2%
talking	0.5%	2.7%	4.2%	5.4%	93.3%

	crossing	waiting	queueing	talking	dancing	jogging
crossing	90.8%	6.0%	0.5%	3.8%	0.0%	0.3%
waiting	8.3%	57.8%	14.6%	1.6%	0.0%	0.0%
queueing	0.4%	24.6%	84.7%	3.6%	0.0%	6.6%
talking	0.4%	11.6%	0.1%	90.2%	0.0%	1.8%
dancing	0.0%	0.0%	0.1%	0.2%	100.0%	0.0%
jogging	0.1%	0.0%	0.0%	0.7%	0.0%	91.2%



The References

1. W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In VS, 2009.
2. T. Lan, Y. Wang, G. Mori, and S. N. Robinovitch. Retrieving actions in group contexts. In SGA, 2010.
3. S. Khamis, V. I. Morariu, and L. S. Davis. A flow model for joint action recognition and identity maintenance. In CVPR, 2012.
4. K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In ICML, 2008.
5. L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In CVPR, 2008.