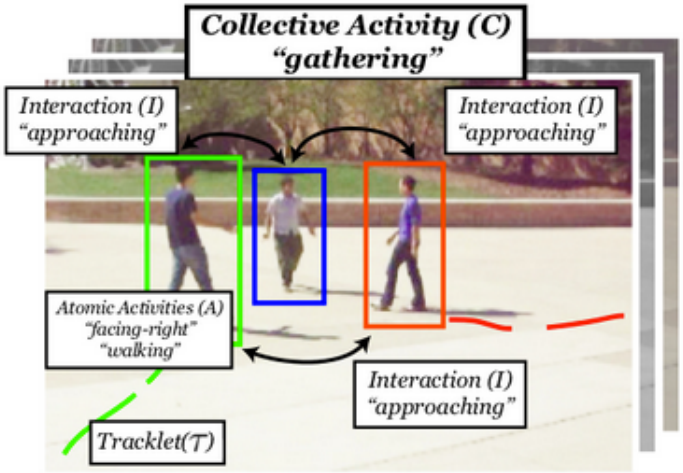
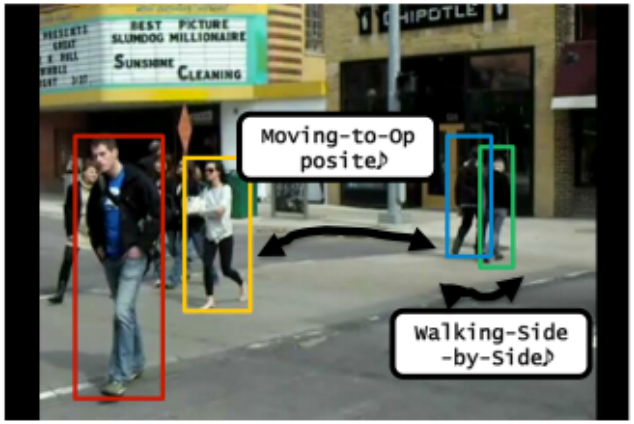
A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

Wong Choi and Silvio Savarese

Presented by: David J. Garcia

November 9, 2012





Organization

- Introduction and Problem Description
- Contributions
- Technical Details
- Experiments
- Conclusions and Extensions

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] http://www-personal.umich.edu/~wgchoi/eccv12/wongun_eccv12.html

Introduction and Problem Description

- Consider the following scenes. What labels would you attach to the "activities" in them?

Introduction and Problem Description

- In this image, perhaps the people are “standing in a line?”

Introduction and Problem Description

- Maybe this one is a talking scene?

Introduction and Problem Description

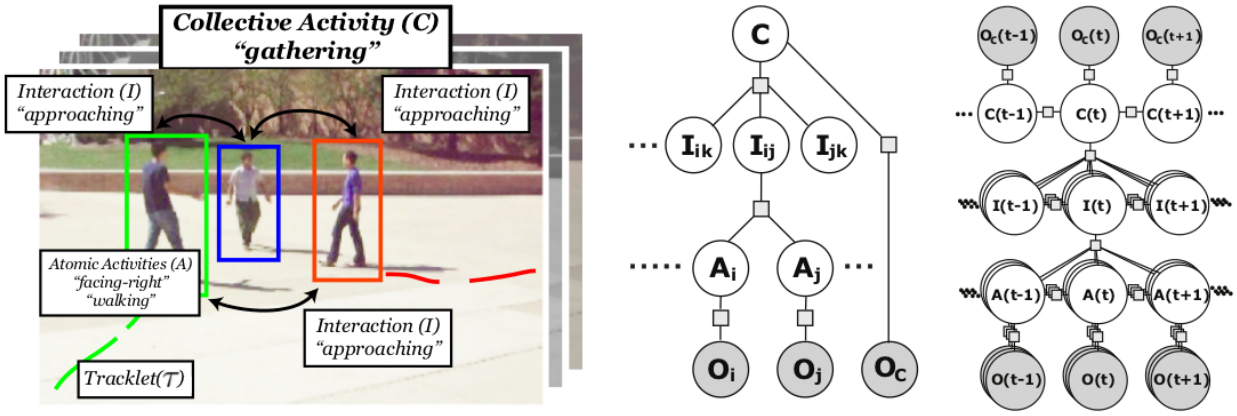
- What about these two? They seem similar in some ways (a group of people focused inward).

- Maybe one group is converging, and maybe the other is simply waiting? It's difficult to tell without more information. Video might be able to add additional information.

Introduction and Problem Description

- To infer an activity in a video scene…track subjects and infer pairwise activities.

- Problems addressed simultaneously with a hierarchical graphical model.



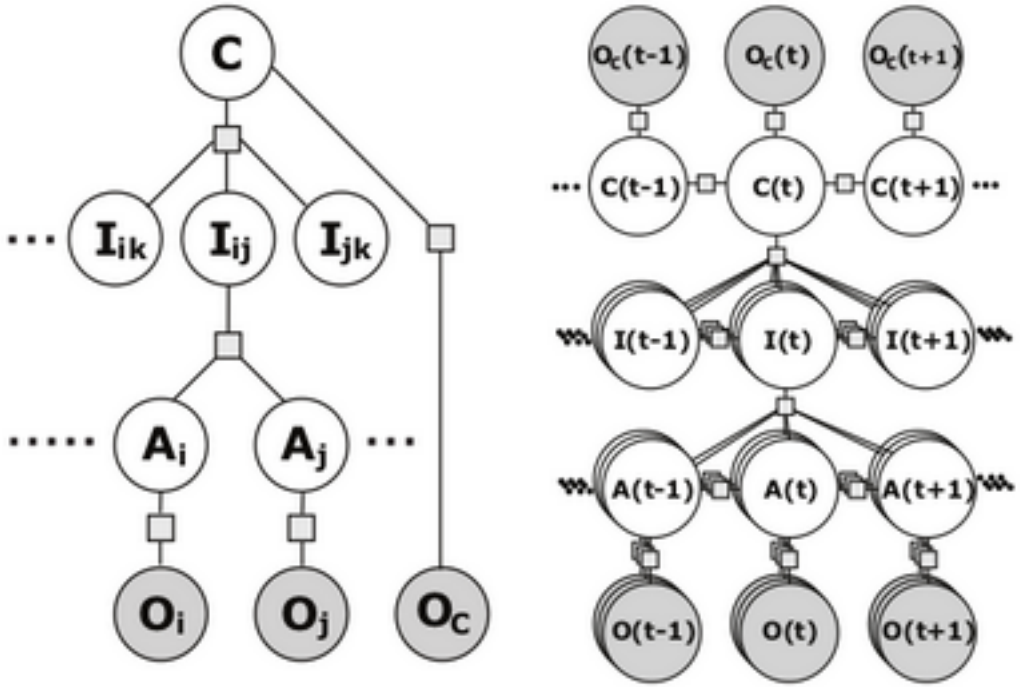
- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] http://www-personal.umich.edu/~wgchoi/eccv12/wongun_eccv12.html

Contributions

- Graphical model that correlates collective activity, pairwise activity and individual activity in a hierarchical fashion.
- Simultaneously solving the tracking problem and the activity inference problem: The tracking problem is solved with help from annotated data.
- Solving this “joint inference problem” with a novel algorithm that combines belief propagation and the branch and bound algorithm.
- The authors also evaluate their method against challenging datasets.

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] http://www-personal.umich.edu/~wgchoi/eccv12/wongun_eccv12.html

Technical Details: Activity Modeling



Collective-Activity: C (Oc is called the crowd context descriptor)

Pairwise-Activity: I (i and j are the individual subjects in the interaction)

Individual-Activity: A (Oi are the "appearance features" of subject i)

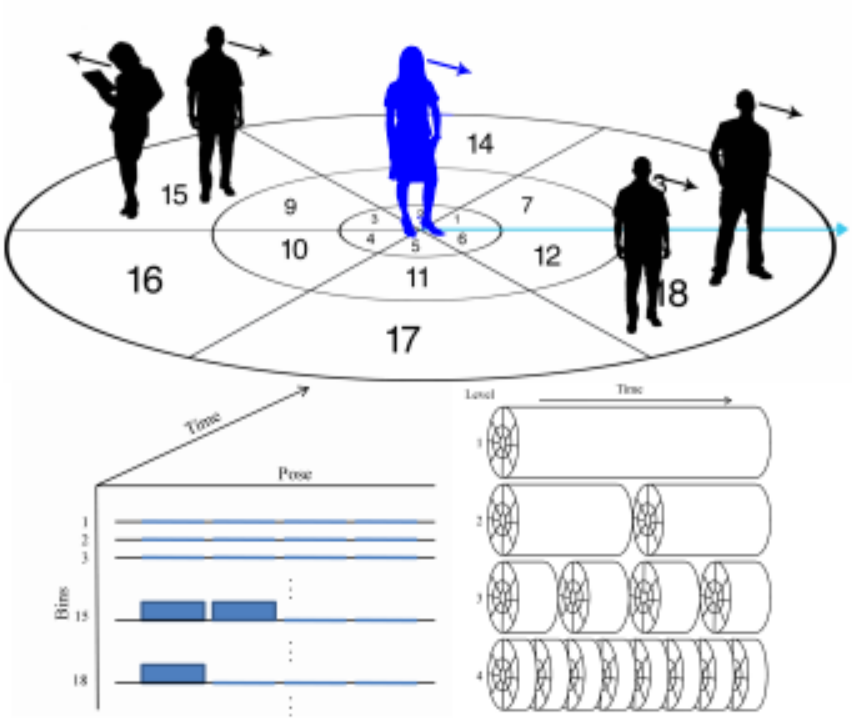
These include features like HoG [3] and BoV [4]

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] http://www-personal.umich.edu/~wgchoi/eccv12/wongun_eccv12.html
- [3] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR. 2005
- [4] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S. Behavior recognition via sparse spatio-temporal features. VS-PETS. 2005

Technical Details: Collective Observations…the STL descriptor

- The spatio-temporal local (STL) descriptor is a binning-style descriptor.
- An “anchor” is chosen (indicated in blue).
- The location (and pose) of every other person is noted and placed into a bin relative to the anchor.
- These histograms are “stacked” along the time dimension..
- For more information, please see [3] and [4]

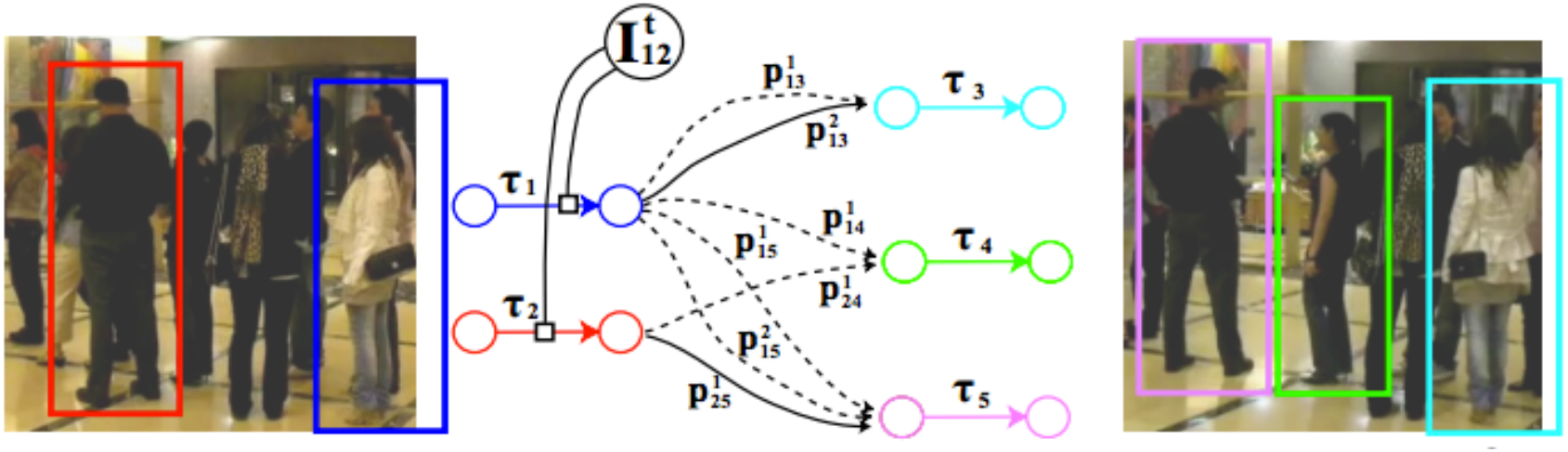
- Side Note: Almost like “letter” recognition?



- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] http://www-personal.umich.edu/~wgchoi/eccv12/wongun_eccv12.html
- [3] W. Choi, S. K. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. VSWS 2009
- [4] W. Choi, S. K. Savarese. Learning context for collective activity recognition. CVPR 2011

Technical Details: Tracking Problem

- Although it's assumed that only one collective activity is present in any one scene(multiple frames), multiple subjects need to be tracked.



- The tracklet association problem can be understood as a matching problem.
- The best association is the one of LEAST resistance…

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

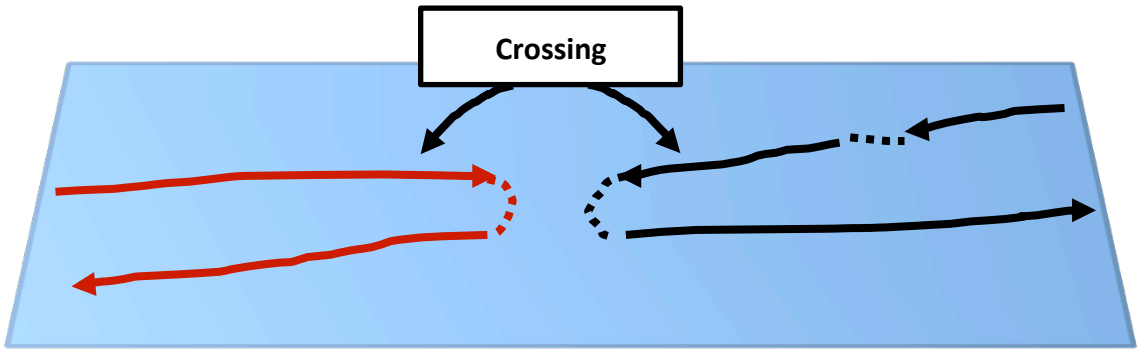
Technical Details: Combining Everything

$$\Psi(C, I, A, O, f) =$$

$$\Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) +$$

Tracking

$$\Psi(C) + \Psi(I) + \Psi(A) + c \uparrow T f, f \in \boldsymbol{S}$$



- This track association is not good if this is a crossing activity.

- [1]W. Choi, S. Savarese. Slides: http://www.umich.edu/~wgchoi/eccv12/choi_eccv12_final_web.pptx
- [2] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

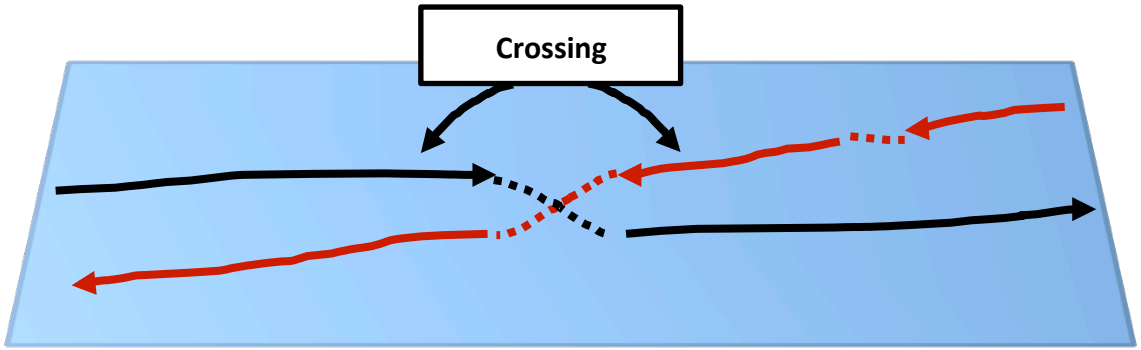
Technical Details: Combining Everything

$$\Psi(C, I, A, O, f) =$$

$$\Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) +$$

Tracking

$$\Psi(C) + \Psi(I) + \Psi(A) + c \uparrow T f, f \in \boldsymbol{S}$$



- This is a much better association.

- [1]W. Choi, S. Savarese. Slides: http://www.umich.edu/~wgchoi/eccv12/choi_eccv12_final_web.pptx
- [2] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

Technical Details: Classification + ? + Tracking = Profit !!!

- The two problems combined as a maximization of the following:

$$\hat{y} = \underset{f,C,I,A}{\operatorname{argmax}} \underbrace{\Psi(C, I, A, O, T(f))}_{Sec.3} - \underbrace{c^T f}_{Sec.4}, \; s.t. \; f \in \mathbb{S}$$

- Section 3 dealt with the classification bit.

- Section 4 defined the tracking problem.

- Note that ‘f’ is incorporated in both terms.

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

Technical Details: Divide and Conquer

- The compact equation on the previous slide is broken up into two problems and solved iteratively.

$$\{\hat{C}, \hat{I}, \hat{A}\} = \underset{C,I,A}{\operatorname{argmax}} \Psi(C, I, A, O, T(\hat{f})) \; AND \; \hat{f} = \underset{f}{\operatorname{argmin}} \, c^T f - \Psi(\hat{I}, \hat{A}, T(f)), \; s.t. \; f \in \mathbb{S}$$

Hold f constant, and solve via Belief Propagation. Obtain 'I' and 'A'

With 'I' and 'A', find f with Branch-and-Bound algorithm[2].

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] Choi, W., Savarese, S.: Supplementary material. In: ECCV. (2012)

Experiments

- Video-words are obtained by applying PCA (principal component analysis) with 200 dimensions and using K-means with 100 code words on the cuboids described in [2].

- Collective activity features are computed using STL on tracklets[3].

- Presumably, tracklets are obtained with [4].

- STL set for 8 meters and 60 frames for reinforcement.

- Experiments are done over two datasets.

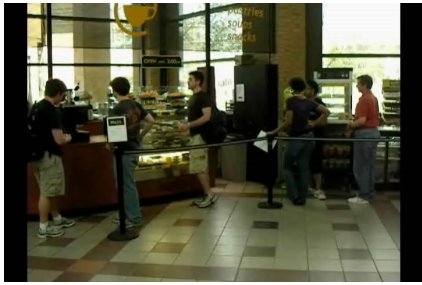
- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S. Behavior recognition via sparse spatio-temporal features. VS-PETS. 2005
- [3] W. Choi, S. K. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. VSWS 2009
- [4] Choi, W., Savarese, S.: Multiple target tracking in world coordinate with single, minimally calibrated camera. In: ECCV. (2010)

Experiments: Dataset One

- First dataset is composed of 44 video clips with annotations for 5 collective activities (crossing, waiting, queuing, walking, and talking) and 8 poses (right, right-front,…, right back). 8 types of interactions are annotated:
- AP (approaching)
- LV (leaving)
- PB (passing-by)
- FE (facing-each-other)
- WS (walking-side-by-side)
- SR (standing-in-a-row)
- SS (standing-side-by-side)
- NA (no interactions)

Crossing

Waiting



Queuing

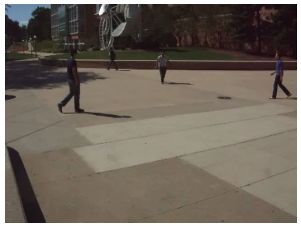
Walking

Talking

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2]]W. Choi, S. Savarese. Slides: http://www.umich.edu/~wgchoi/eccv12/choi_eccv12_final_web.pptx

Experiments: Dataset Two

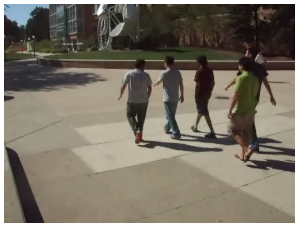
- Second dataset is composed of 32 video clips with annotations for 6 collective activities (gathering, talking, dismissal, walking, together, chasing, queueing) and 8 poses (right, right-front,…, right back). 9 types of interactions are annotated:
- AP (approaching)
- WO (walking-in-opposite-direction)
- WR (walking-one-after-the-other)
- RS (running-side-by-side)
- RR (running-one-after-the-other)
- FE (facing-each-other)
- WS (walking-side-by-side)
- SR (standing-in-a-row)
- NA (no interactions)



Gathering

Talking

Dismissal



Walking-together

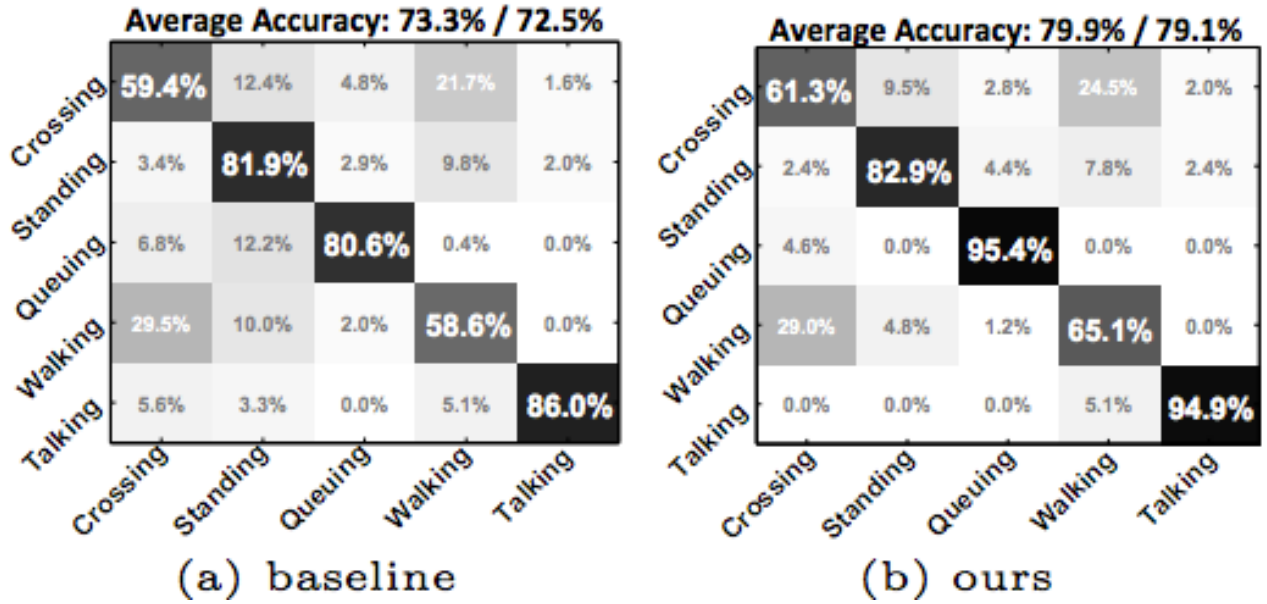
Chasing

Queuing

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2]]W. Choi, S. Savarese. Slides: http://www.umich.edu/~wgchoi/eccv12/choi_eccv12_final_web.pptx

Experiments: Dataset One for Classification

- Accuracy measure show significant improvement for the method

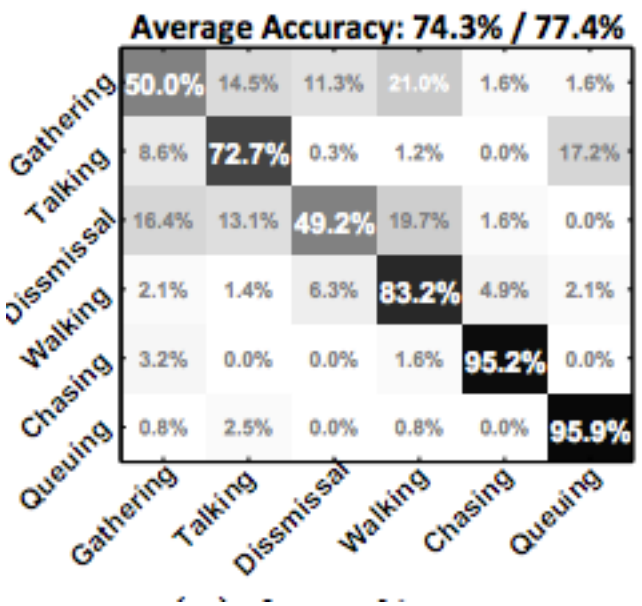


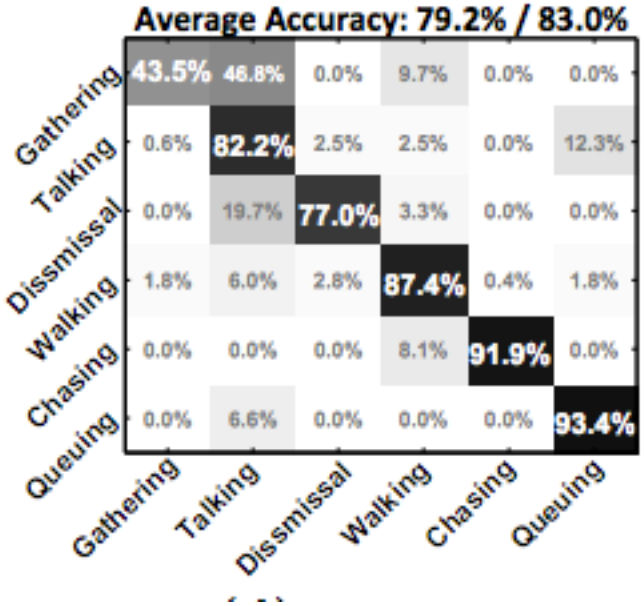
- The numbers on top are mean-accuracy per class, and overall accuracy.
- The basic SVM isn't terrible and much simpler to do.

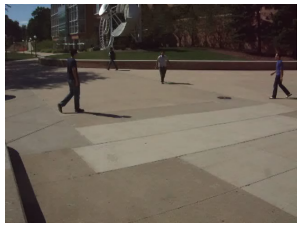
- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

Experiments: Dataset Two for Classification

- Dataset two also shows improvement over the baseline







Gathering

Dismissal

- The numbers on top are mean-accuracy per class, and overall accuracy.
- What do the “Gathering” and “Dissmissal” activities have in common?

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition

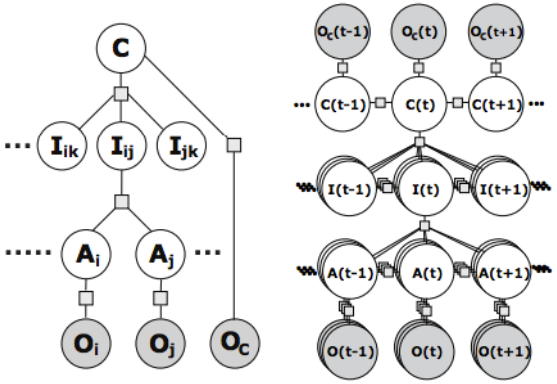
Experiments: Graphical Model Experiments

- Is the graphical model correct? The authors test different versions:

	Dataset [1]				New Dataset			
Method	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)
without O_C	38.7	37.1	40.5	37.3	59.2	57.4	49.4	41.1
no edges between C and I	67.7	68.2	42.8	37.7	67.8	54.6	42.4	32.8
no temporal chain	66.9	66.3	42.6	33.7	71.1	68.9	41.9	46.1
no temporal chain between C	74.1	75.0	54.2	48.6	77.0	76.1	**55.9**	**48.6**
full model ($\triangle t_C = 20, \triangle t_I = 25$)	**79.0**	**79.6**	**56.2**	**50.8**	**83.0**	**79.2**	53.3	43.7
baseline	72.5	73.3	-	-	77.4	74.3	-	-

- The authors cut some of these links and see what happens to classification accuracy.

- The Collective observations are clearly the most important.



- Baseline: SVM over Collective activity observations (i.e. STL[2]).

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] W. Choi, S. K. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. VSWS 2009

Experiments: Temporal Support's effect on Classification

- What is the sensitivity of $\Psi(C, I)$ and $\Psi(I, A, T)$ to the temporal support window? The authors isolate each window and test different sizes.

	Dataset [1]				New Dataset			
Method	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)	Ovral (C)	Mean (C)	Ovral (I)	Mean (I)
$\triangle t_C = 30, \triangle t_I = 25$	79.1	79.9	56.1	50.8	80.8	77.0	**54.3**	**46.3**
$\triangle t_C = 20, \triangle t_I = 25$	79.0	79.6	**56.2**	**50.8**	**83.0**	**79.2**	53.3	43.7
$\triangle t_C = 10, \triangle t_I = 25$	77.4	78.2	56.1	50.7	81.5	77.6	52.9	41.8
$\triangle t_C = 30, \triangle t_I = 15$	76.1	76.7	52.8	40.7	80.7	71.8	48.6	34.8
$\triangle t_C = 30, \triangle t_I = 5$	**79.4**	**80.2**	45.5	36.6	77.0	67.3	37.7	25.7

- Authors claim that classification becomes more robust. Perhaps for collective activity more so than for pairwise-inference? How conclusive are these results?
- These values directly affect the "cuboid" based features[2]. Perhaps some windows are better for particular activities.

- [1] W. Choi, S. Savarese. A Unified Framework for Multi-Target Tracking and Collective Activity Recognition
- [2] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS. (2005)

Conclusions/Extensions

- Graphical model links that seemed most important were Collective-Observations link and the temporal link between collective activities.

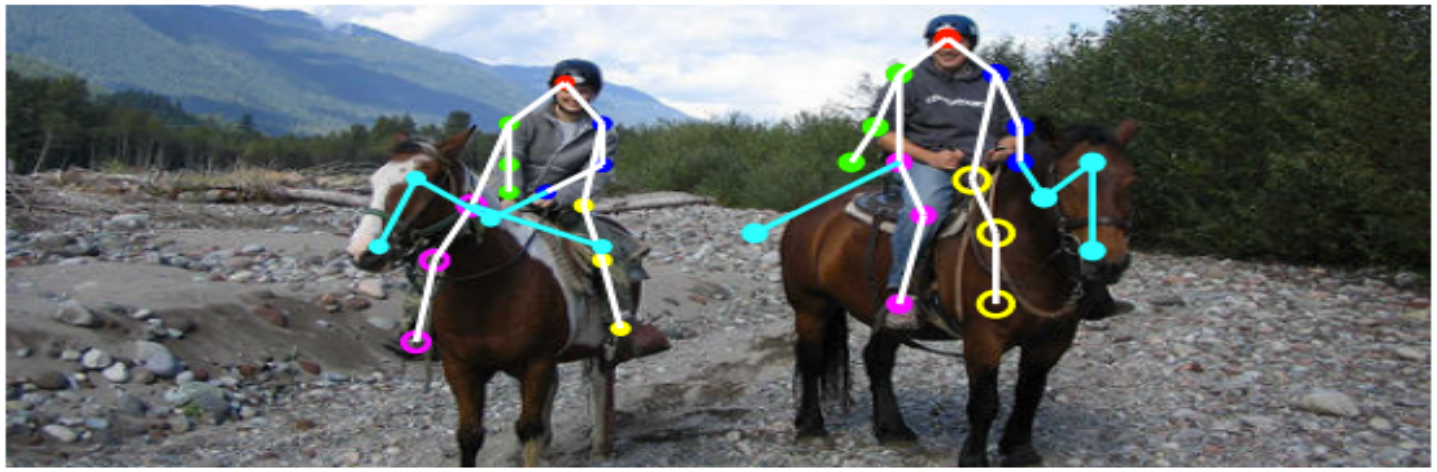
- Is the still camera assumption reasonable (maybe for some applications)? This stems from the 'cuboid' [1].

- Authors suggest a multi-activity detector as a possible extension.

- Finer grained temporal considerations (like different temporal windows for different activities)

- [1] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS. (2005)

Detecting Actions, Poses, and Objects with Relational Phraselets



- Person-Object “composites” are combined “into local patches or ‘phraselets.’”
- Phraselets are then used for learning after a clustering step.
- There are separate “mixtures for visible and occluded parts.”

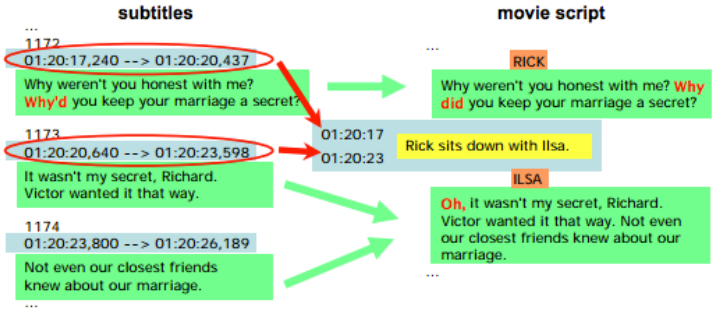
(a) Visible elbow phraselets

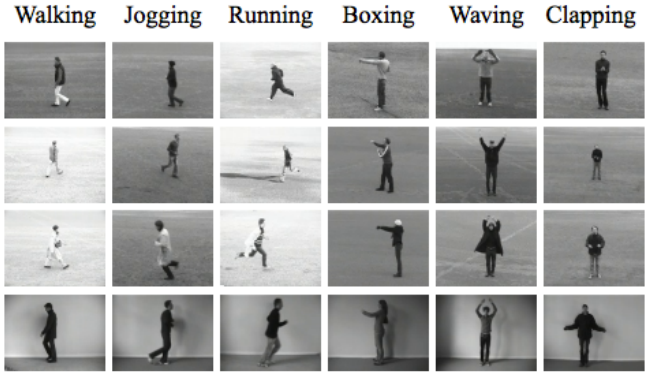
(b) Occluded elbow phraselets

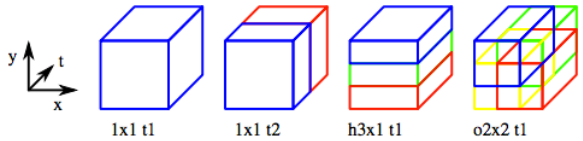
- [1] C. Desai. D. Ramanan. Detecting Actions, Poses, and Objects with Relational Phraselets

Learning realistic human actions from movies

- Movies used to train SVM to detect actions.
- Scripts are used to aid in action annotation.
- Spatial-Temporal features are extracted for the sequences/subsequences.
- SVM is trained on clustered BoF data-points.







- KTH dataset evaluation[2].

Task	HoG BoF	HoF BoF	Best channel	Best combination
KTH multi-class	81.6%	89.7%	91.1% (hof h3x1 t3)	91.8% (hof 1 t2, hog 1 t3)
Action AnswerPhone	13.4%	24.6%	26.7% (hof h3x1 t3)	32.1% (hof o2x2 t1, hof h3x1 t3)
Action GetOutCar	21.9%	14.9%	22.5% (hof o2x2 1)	41.5% (hof o2x2 t1, hog h3x1 t1)
Action HandShake	18.6%	12.1%	23.7% (hog h3x1 1)	32.3% (hog h3x1 t1, hog o2x2 t3)
Action HugPerson	29.1%	17.4%	34.9% (hog h3x1 t2)	40.6% (hog 1 t2, hog o2x2 t2, hog h3x1 t2)
Action Kiss	52.0%	36.5%	52.0% (hog 1 1)	53.3% (hog 1 t1, hof 1 t1, hof o2x2 t1)
Action SitDown	29.1%	20.7%	37.8% (hog 1 t2)	38.6% (hog 1 t2, hog 1 t3)
Action SitUp	6.5%	5.7%	15.2% (hog h3x1 t2)	18.2% (hog o2x2 t1, hog o2x2 t2, hog h3x1 t2)
Action StandUp	45.4%	40.0%	45.4% (hog 1 1)	50.5% (hog 1 t1, hof 1 t2)

- [1] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld. Learning realistic human actions from movies
- [2] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In ICPR, 2004.