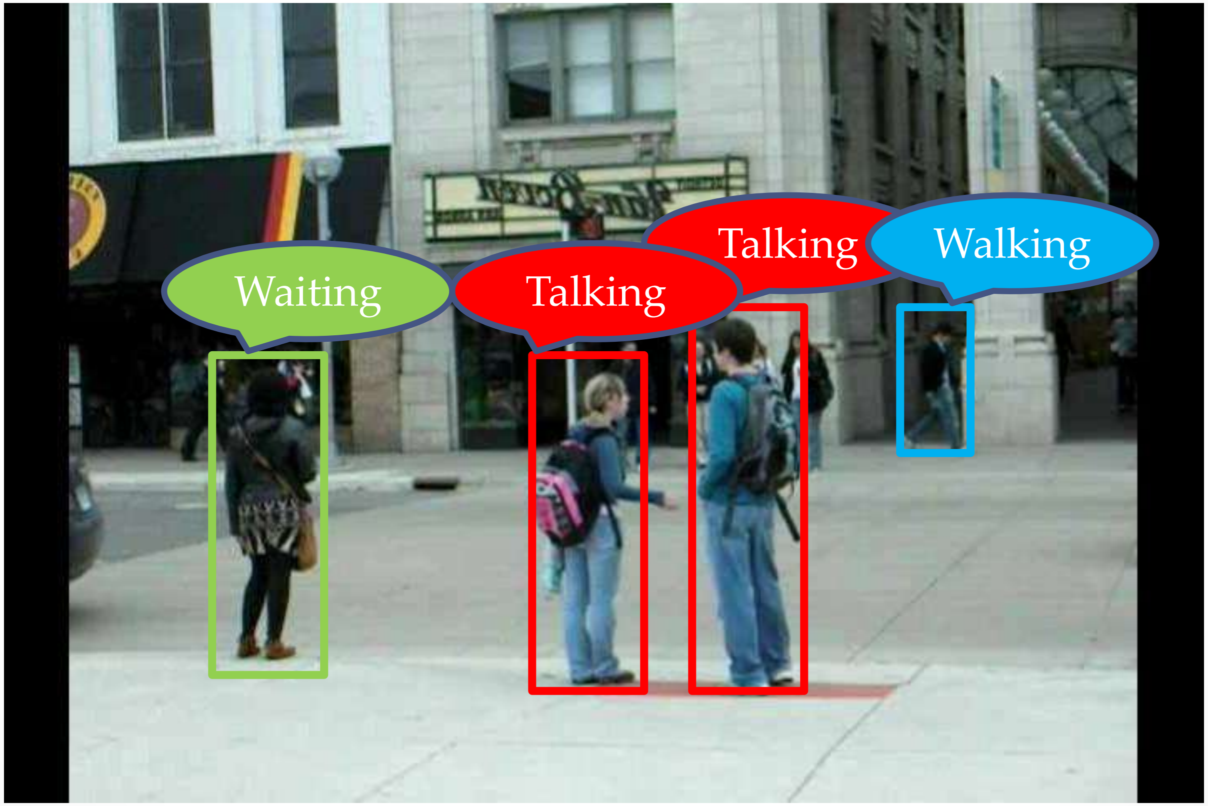
Collective Activity Detection using Hinge-loss Markov Random Fields

Ben London, Sameh Khamis, Stephen H. Bach,
Bert Huang, Lise Getoor, Larry Davis

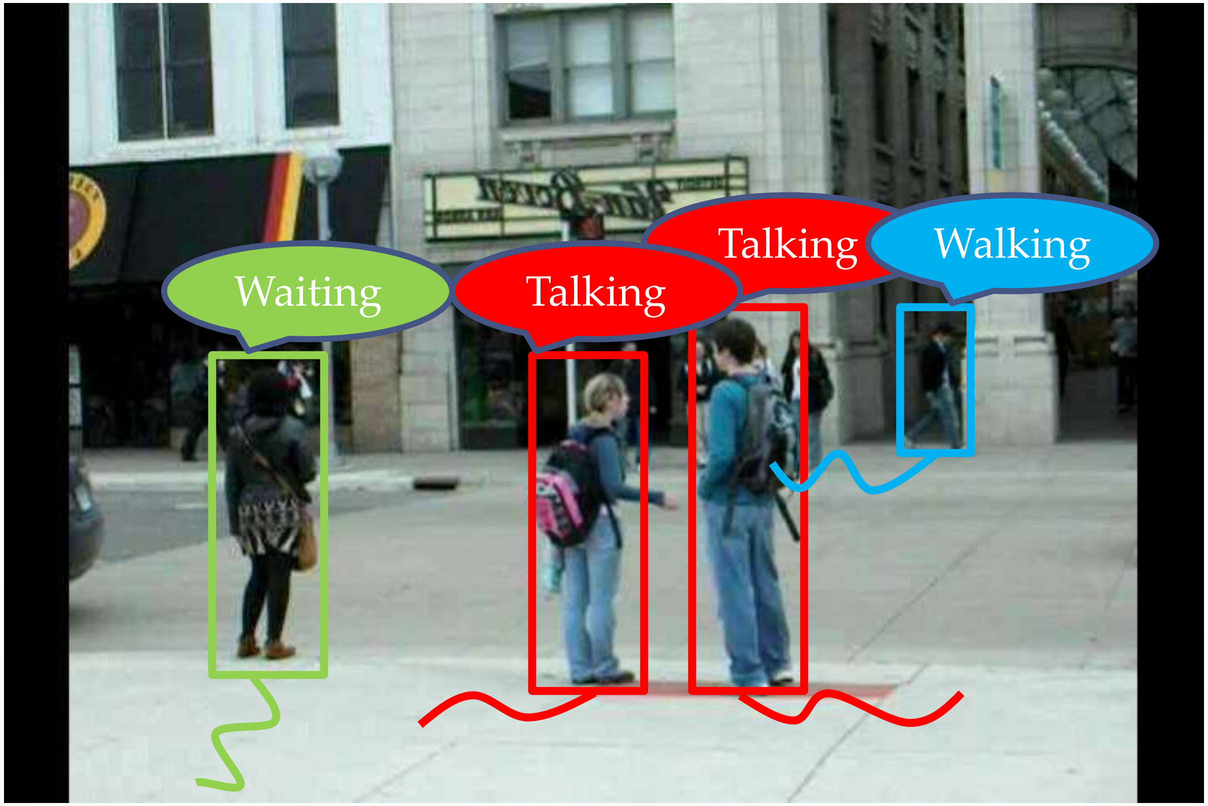
Motivation

Motivation



- Classify the individual actions

Motivation



- Classify the individual actions
- Track the multiple targets

Intuition

- Action transitions are unlikely

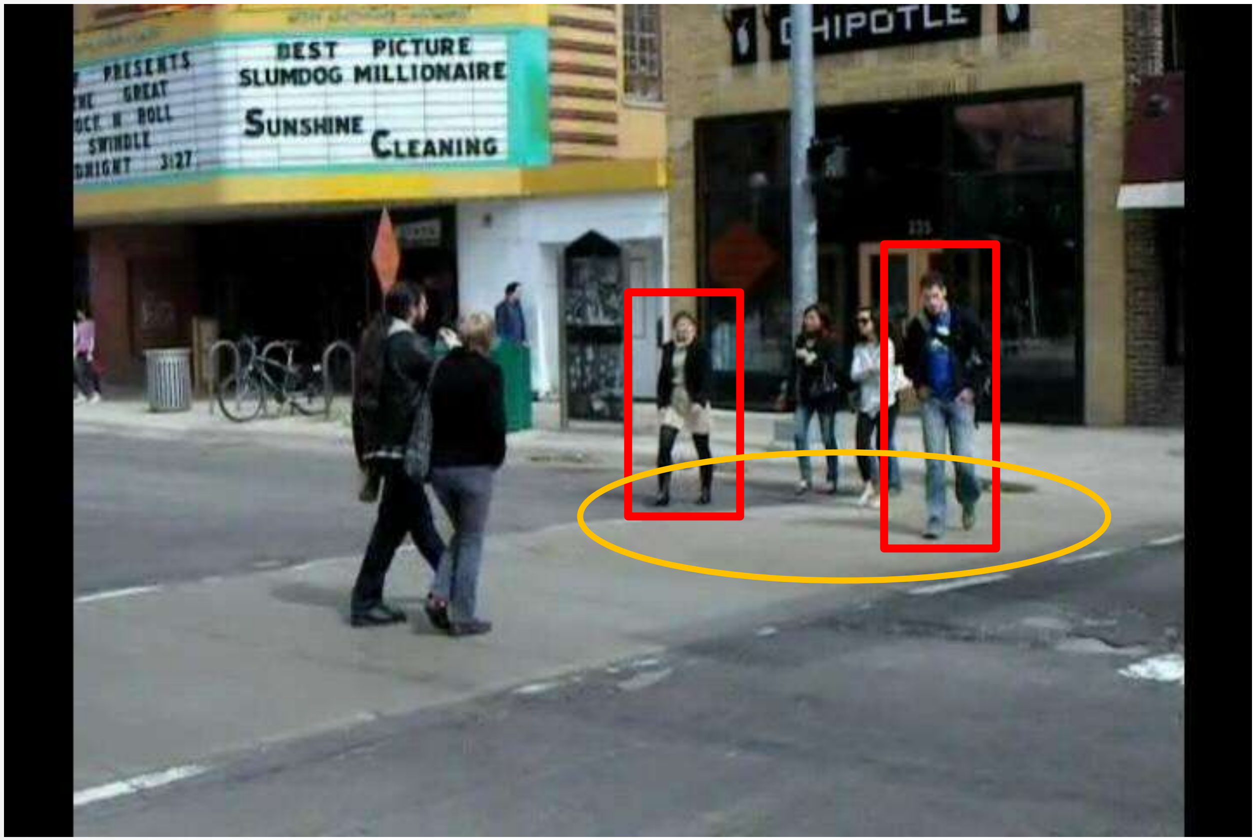
Intuition

- Action transitions are typically not arbitrary

Intuition

- Individual actions are consistent in proximity

Intuition

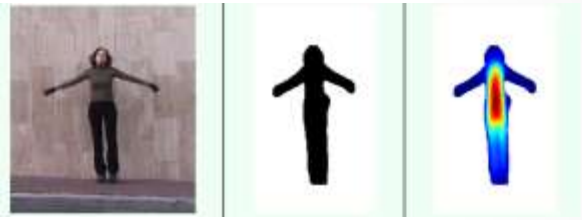


- Individual actions are consistent in proximity

Related Work

- Original action recognition work focused on the isolated person case

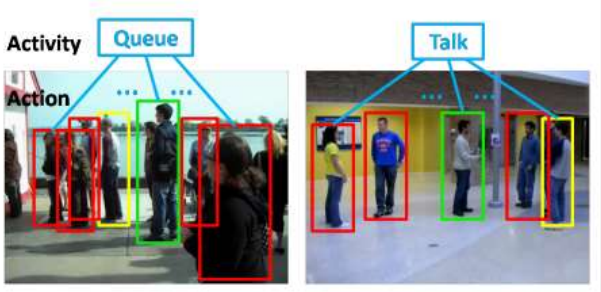
Shuldt et al., ICPR 2004



Blank et al., CVPR 2005

- Following work investigated either pairwise interactions or group activity as the activity of the majority

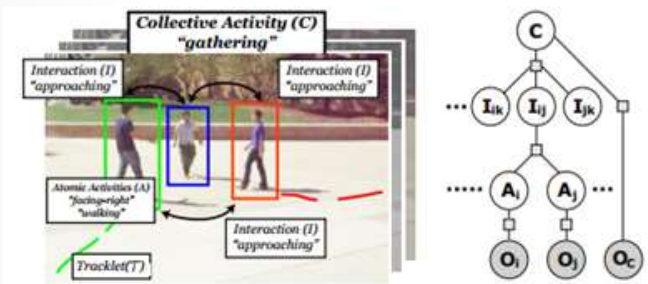
Ryoo and Agarwal, ICCV 2009



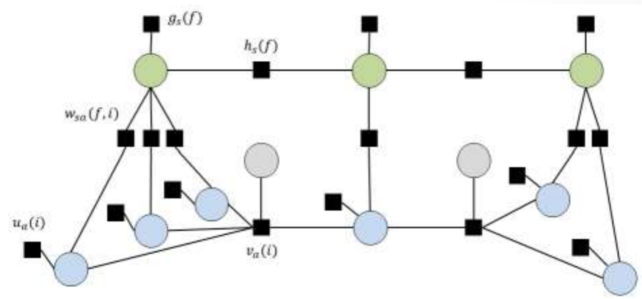
Lan et al., NIPS 2010

Related Work

- More recent work looked at coupling activity recognition, tracking, and scene labeling



Choi and Savarese, ECCV 2012



Khamis et al., ECCV 2012

- While others modeled activities at multiple levels: individual, group, and inter-group

Amer et al., ECCV 2012

Our Approach

An Introduction to Hinge-loss MRFs and PSL

Our Approach

- Problem needs **scalable** solution that handles complex dependencies and tracking constraints

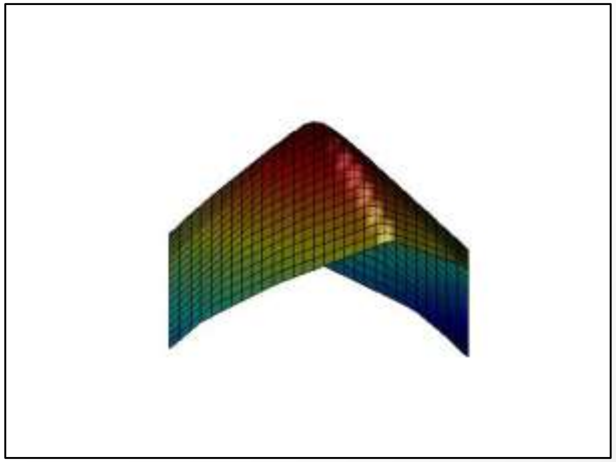
- *Hinge-loss Markov Random fields* (HL-MRFs) are a new class of models that meet these goals
  - Log-concave densities over continuous variables
  - Support fast inference of global solutions
  - New paper on structured prediction at UAI 2013

- *Probabilistic soft logic* (PSL) allows easy encoding of intuitions
  - Converts logical rules to HL-MRFs

Hinge-loss Markov Random Fields

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z} \exp\left[-\sum_{j=1}^{m} w_j \max\{\ell_j(\mathbf{Y}, \mathbf{X}), 0\}^{p_j} \right]$$

- Continuous variables in [0,1]
- Potentials are hinge-loss functions
- Subject to arbitrary linear constraints
- Log-concave!



Inferring Most Probable Explanations

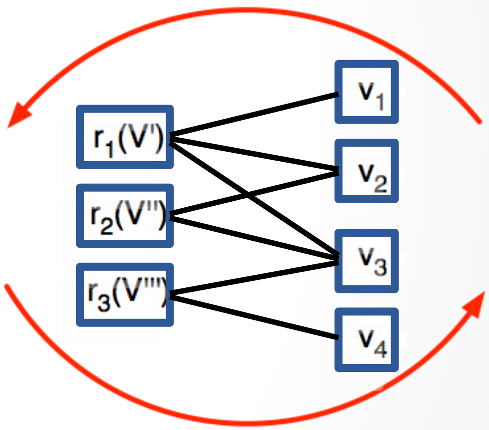
- Objective:

$$\arg\max_{\mathbf{Y}} p(\mathbf{Y}|\mathbf{X}) = \arg\min_{\mathbf{Y}} \sum_{j=1}^{m} w_j \max\{\ell_j(\mathbf{Y}, \mathbf{X}), 0\}^{p_j}$$

- Convex optimization
- Decomposition-based inference algorithm using the ADMM framework

Alternating Direction Method of Multipliers

- Inference with ADMM is fast, scalable, and straightforward
- Optimize subproblems (ground rules) independently, in parallel
- Auxiliary variables enforce consensus across subproblems



Weight Learning

- Various methods to learn from training data:
  - approximate maximum likelihood
  - maximum pseudolikelihood
  - large-margin estimation
  - [Broecheler et al., UAI 2010; Bach et al., UAI 2013]

- State-of-the-art learning performance on
  - Collective classification
  - Social-trust prediction
  - Preference prediction
  - Image reconstruction

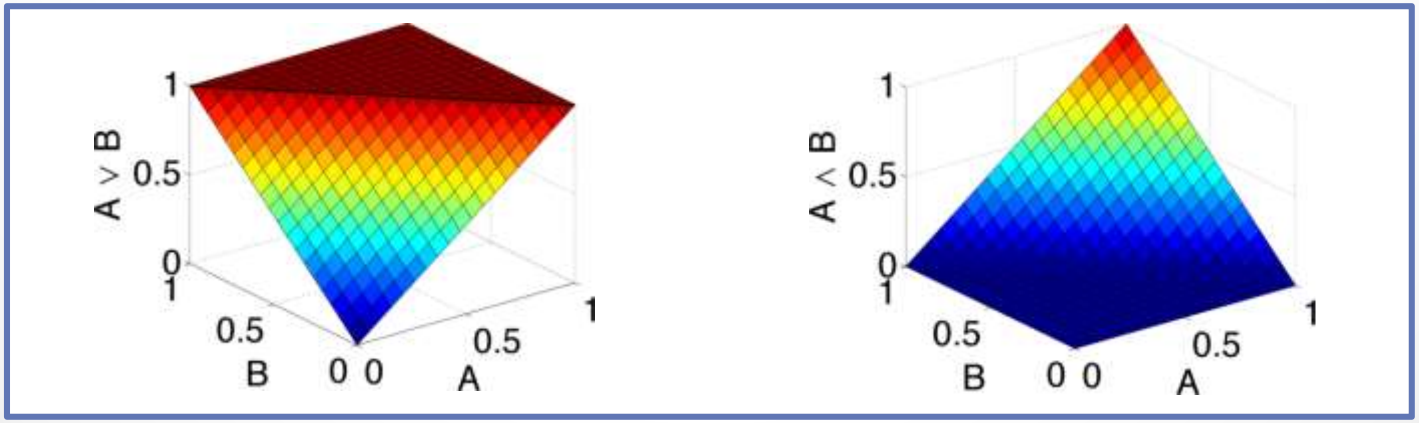
- Here we use approximate maximum likelihood

Probabilistic Soft Logic

- HL-MRFs are easy to define
- Hinge-losses can generalize logical operators

1.8: Doing(X, walking) $\leftarrow$ SamePerson(X, Y)$\wedge$ Doing(Y, walking)

- Lukasiewicz T-norm
  - A $\vee$ B = min{1, A + B}
  - A $\wedge$ B = max{0, A + B – 1}



Grounding to HL-MRFs

- Ground out first-order rules
  - Variables: soft-truth values of atoms
  - Hinge-loss potentials: weighted **distances to satisfaction** of ground rules

- $w : A \to B$

  $w : \neg A \vee B$

  $w \times (1 - \min\{1 - A + B, 1\})$

  $w \times \max\{A - B, 0\}$

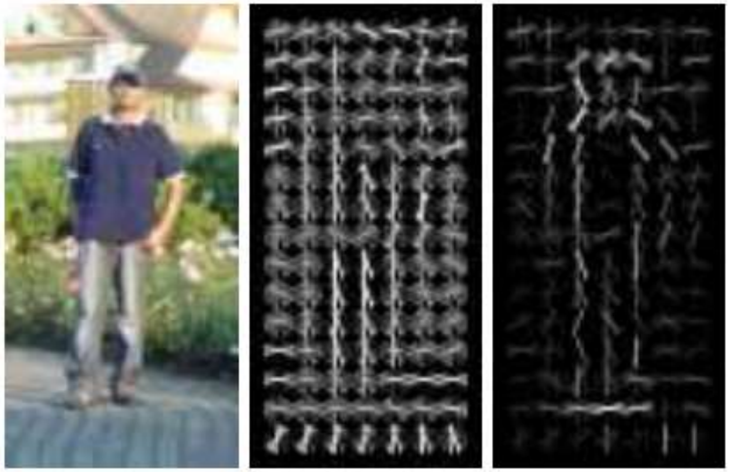
- The effect is assignments that satisfy weighted rules more are more probable

A PSL Model for Collective Activity Detection

A Collective Activity Detection Model in PSL

Features: Low-Level

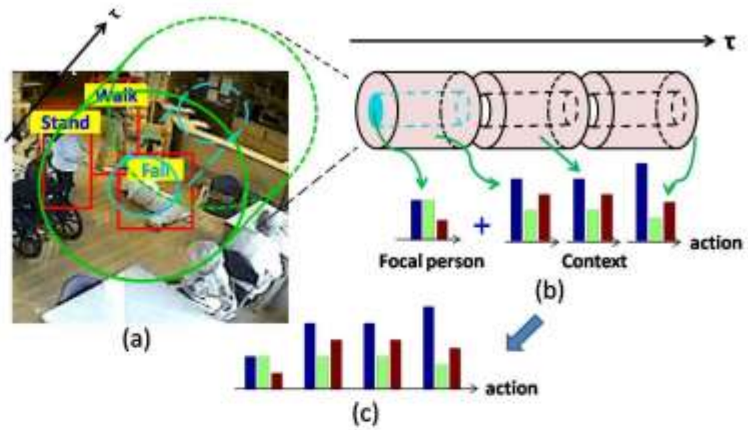
- Histogram of Oriented Gradients (HOG) [Dalal & Triggs, CVPR 2005]



- Describe image patches by a distribution of gradient magnitudes binned by angle
- We train SVMs to predict on HOG features

Features: Low-Level

- Action Context Descriptor (ACD) [Lan et al, NIPS 2010]



- Model context by aggregating SVM outputs on HOG features across multiple spatiotemporal neighborhoods
- E.g, actions like talking cannot be represented by the HOG features of one person

Local Information

- Use low-level detectors

$$w_{local,a}: \text{Doing}(X, a) \leftarrow \text{Detector}(X, a)$$

- E.g.,

$$w_{local,\textit{walking}}: \text{Doing}(X, \textit{walking}) \leftarrow \text{Detector}(X, \textit{walking})$$

$$w_{local,\textit{talking}}: \text{Doing}(X, \textit{talking}) \leftarrow \text{Detector}(X, \textit{talking})$$

$$w_{local,\textit{waiting}}: \text{Doing}(X, \textit{waiting}) \leftarrow \text{Detector}(X, \textit{waiting})$$

$\vdots$ (defined for all actions)

Frame Consistency

- Most people in frame do the same action
- Ground truth is aggregate of descriptors

$$w_{frame,a} : \text{Doing}(X, a) \leftarrow \text{Frame}(X, F) \wedge \text{FrameAction}(F, a)$$

Effect of Proximity

- People that are close (in frame) are likely doing the same action

$$w_{prox,a} : \mathrm{Doing}(X, a) \leftarrow \mathrm{Close}(X, Y) \wedge \mathrm{Doing}(Y, a)$$

- Closeness is measured via a radial basis function

Tracking

- Persistence rules
  - People are likely to continue doing the same action

$$w_{persist,a} : \text{Doing}(Y, a) \leftarrow \text{SamePerson}(X, Y) \wedge \text{Doing}(X, a)$$

  - Requires identity maintenance for SamePerson

- Identity maintenance

$$w_{id} : \text{Same}(X, Y) \leftarrow \text{Sequential}(X, Y) \wedge \text{Close}(X, Y)$$

Action Transitions

- Can define rules for transitioning between actions

$$w_{trans,a,b} : \text{Doing}(Y, b) \leftarrow \text{SamePerson}(X, Y) \wedge \text{Doing}(X, a)$$

- Defined over all pairs of actions (a,b)
- Effect is similar to the state transition matrix of an HMM

Priors and Constraints

- Prior beliefs
  - Encode prior beliefs about SamePerson and Doing predicates

w : ~SamePerson(X, Y) w : ~Doing(X, a)

- Constraints
  - Functional constraint on Doing ensures that soft-truth values for each person sum to 1
  - Partial-functional constraint on SamePerson ensures that soft-truth values for each person sum to at most 1

Experiments

Dataset

- University of Michigan, "Collective Activity"
- Annotated activities, poses, trajectories
  - We don't use poses, trajectories
  - We only use activity annotations for training
- 2 common splits:
  - 5-label: [*crossing, walking, waiting, talking, queueing*]
    - 44 sequences
  - 6-label: [*crossing, waiting, talking, queueing, dancing, jogging*]
    - 63 sequences

http://www.eecs.umich.edu/vision/activity-dataset.html

PSL Model

w_{id} : Same(X, Y) ← Sequential(X, Y) ∧ Close(X, Y)

$w_{idprior}$: ~SamePerson(X, Y)

For all actions a:

$w_{local,a}$: Doing(X, a) ← Detector(X, a)

$w_{frame,a}$: Doing(X, a) ← Frame(X, F) ∧ FrameAction(F, a)

$w_{prox,a}$: Doing(X, a) ← Close(X, Y) ∧ Doing(Y, a)

$w_{persist,a}$: Doing(Y, a) ← SamePerson(X, Y) ∧ Doing(X, a)

$w_{prior,a}$: ~Doing(X, a)

Methodology

- Measure benefit of high-level reasoning
  - One model using HOG SVM scores, another using ACD SVM scores
  - Measure lift over low-level detectors
- Leave-one-out cross-validation
  - Train on all but one sequence
  - Test on hold-out
  - Accumulate test statistics over all hold-outs
    - Compensates for varying lengths and label distributions

Results

	5-Action		6-Action	
	Accuracy	F1	Accuracy	F1
HOG SVM	0.474	0.481	0.596	0.582
HL-MRF + HOG	**0.598**	**0.603**	**0.793**	**0.789**
ACD SVM	0.675	0.678	0.835	0.835
HL-MRF + ACD	**0.692**	**0.693**	**0.860**	**0.860**

What about MLNs?

- Also compare against an identical Markov logic network (MLN) model
  - Inference and MLE in MLNs are generally intractable
  - MaxWalkSat for learning
  - MCSAT for test-time inference

Results

	5-Action		6-Action	
	Accuracy	F1	Accuracy	F1
HOG SVM	0.474	0.481	0.596	0.582
MLN + HOG	**0.657**	**0.657**	**0.809**	**0.803**
HL-MRF + HOG	0.598	0.603	0.793	0.789
ACD SVM	0.675	0.678	0.835	0.835
MLN + ACD	0.687	0.685	0.850	0.850
HL-MRF + ACD	**0.692**	**0.693**	**0.860**	**0.860**

Speed

Average running time

	Cora	Citeseer	Epinions	Activity
MLN	110.9 s	184.3 s	212.4 s	344.2 s
HL-MRF	**0.4 s**	**0.7 s**	**1.2 s**	**0.6 s**

[Bach et al., UAI 2013]

- MLN inference is **slow**
  - MCSAT is poly-time, but slow
- HL-MRF inference is **fast**
  - In practice, we find that inference scales linearly with the number of potentials

Improved PSL Model

- Scene consistency
  - Certain sequences tend to have a single majority action
  - Improved performance in [Khamis et al., ECCV 2012]
- In-frame/sequence interactions
  - E.g., Maybe *walking* and *crossing* frequently co-occur together?
- Latent variables
  - E.g., Group actors into same-action clusters, reason about cluster interactions

Conclusion

- HL-MRFs are a powerful class of graphical models
  - Capable of fast MPE inference
  - Faster inference than discrete models (e.g., MLNs)
- PSL facilitates easy construction of HL-MRFs
  - First-order-logic syntax
- Using HL-MRFs/PSL for high-level vision yields significant improvement over low/mid-level detectors

Thank you!

- PSL info at http://psl.cs.umd.edu/

- M. R. Amer, D. Xie, M. Zhao, S. Todorovic, S. C. Zhu: Cost-Sensitive Top-Down/Bottom-Up Inference for Multiscale Activity Recognition. ECCV 2012
- S. Bach, B. Huang, B. London, L. Getoor. Hinge-loss Markov random fields: convex inference for structured prediction. *UAI*, 2013
- M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri. Actions as Space-Time Shapes. ICCV, 2005
- M. Broecheler, L. Mihalkova, L. Getoor. Probabilistic similarity logic. *UAI*, 2010.
- W. Choi, S. Savarese. A Unified Framework for Multi-target Tracking and Collective Activity Recognition. ECCV 2012
- S. Khamis, V. I. Morariu, L. S. Davis. Combining Per-Frame and Per-Track Cues for Multi-Person Action Recognition. ECCV, 2012
- T. Lan, Y. Wang, W. Yang, G. Mori: Beyond Actions: Discriminative Models for Contextual Group Activities. NIPS 2010
- M. S. Ryoo, J. K. Aggarwal. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. ICCV 2009
- C. Schult, I. Laptev, B. Caputo. Recognizing Human Actions: A Local SVM Approach. ICPR, 2004