Learning Structured Models for Recognizing Human Actions

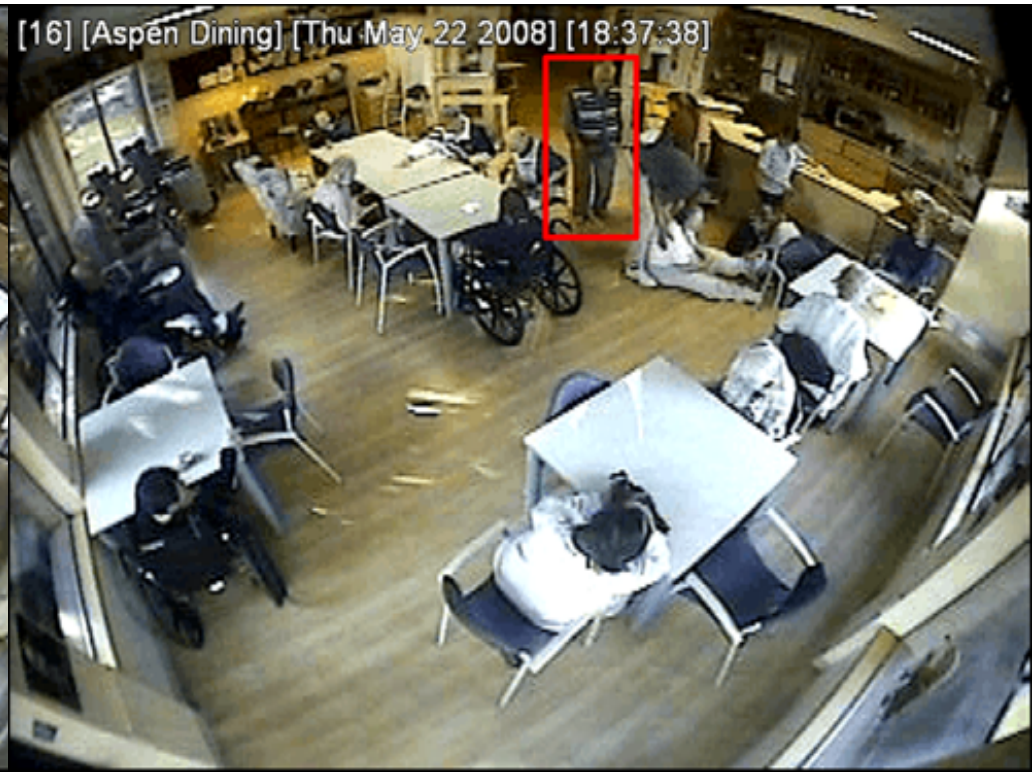
Greg Mori
School of Computing Science
Simon Fraser University

CVPR Workshop on Gesture Recognition
June 20, 2011

SFU Vision and Media Lab

SFU

Action Recognition



Lan, Wang, Yang, Mori NIPS 2010
Lan, Wang, Robinovitch, Mori SGA 2010

Automatically detect falls, near-falls

Applications - HCI

Fathi & Mori CVPR 2008
Bayazit, Couture-Beil, Mori MVA 2009

SFU Vision and Media Lab

SFU

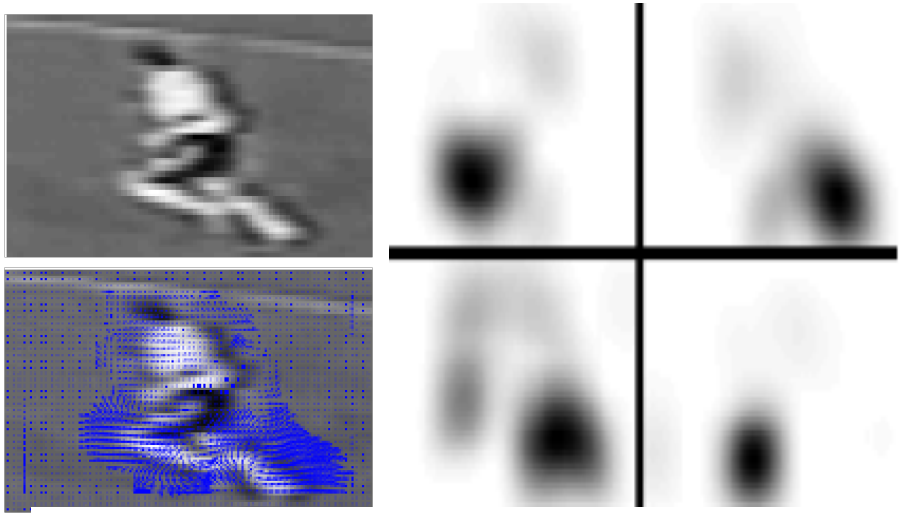
Applications – Human Robot Interaction

Milligan, Mori, Vaughan ACM/IEEE Human Robot Interaction HRI 2011

Structured Models

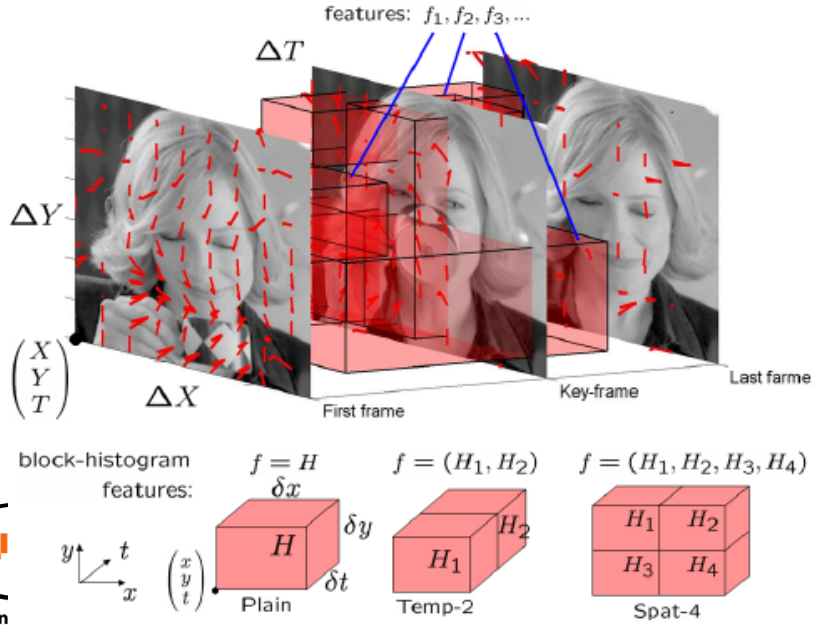
- Models that account for spatial and temporal structure of actions
  - Flexible
    - E.g. local feature models
  - Capture the Gestalt
    - E.g. template representations
- This talk: representations and learning for structured models of human actions

Example – Action Recognition



Large-scale features

[e.g. Efros, Berg, Mori, Malik, ICCV03]



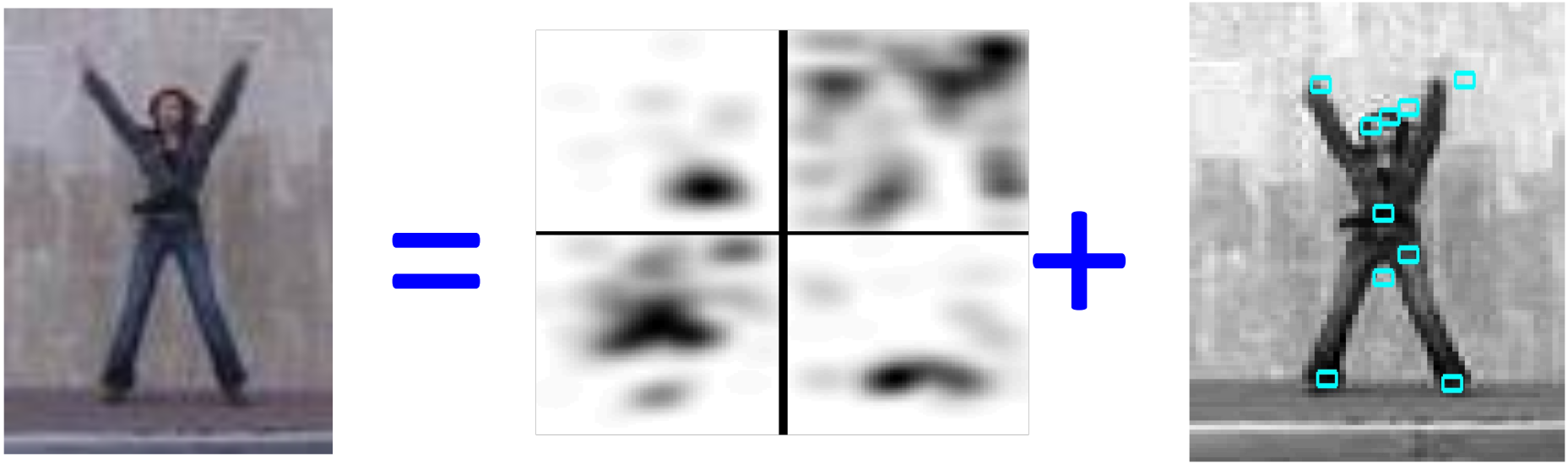
Local patches

[e.g. Laptev & Perez, ICCV07]

SFU Vision

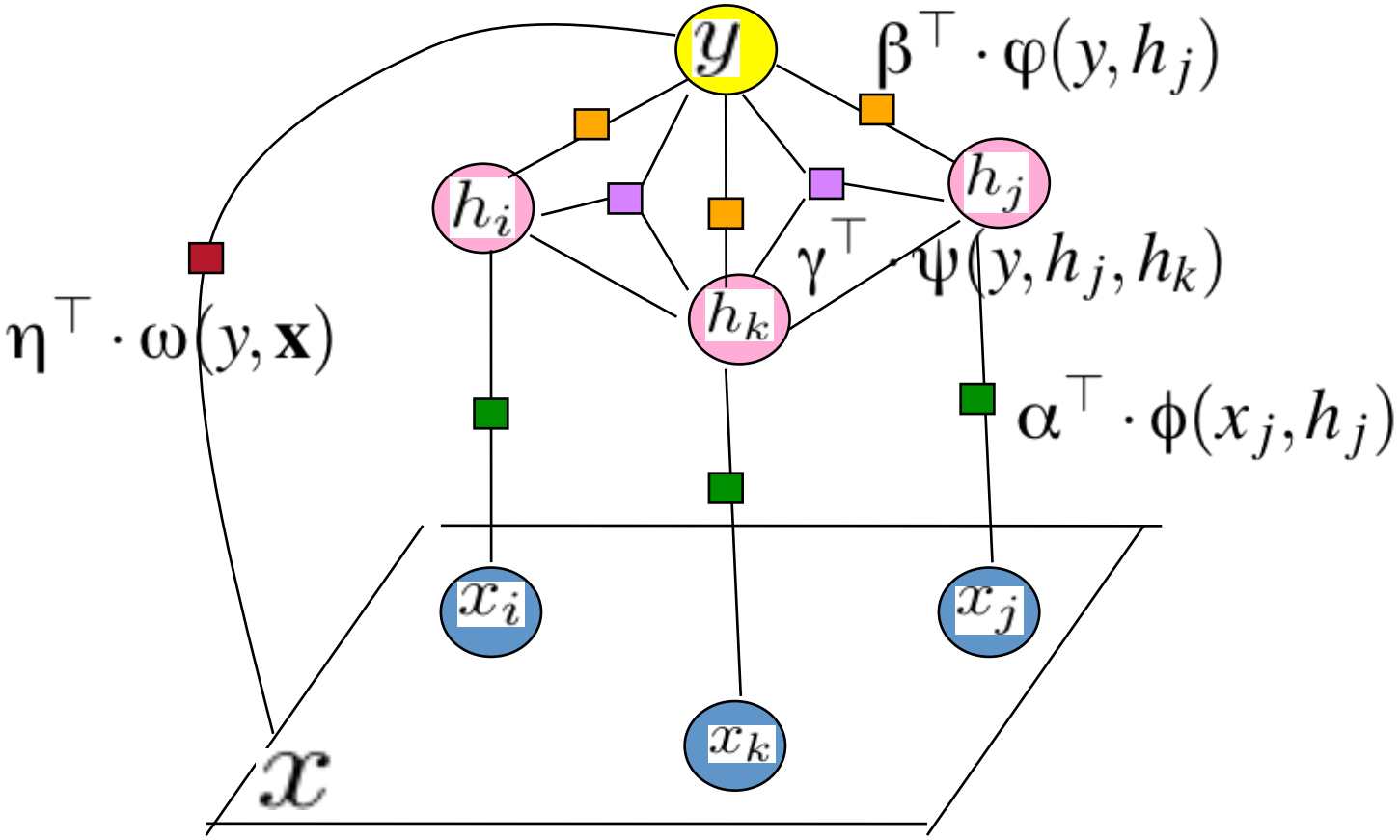
SFU

Large vs. Small Scale Features



Challenge: How to combine in a principled manner?

Hidden Conditional Random Field

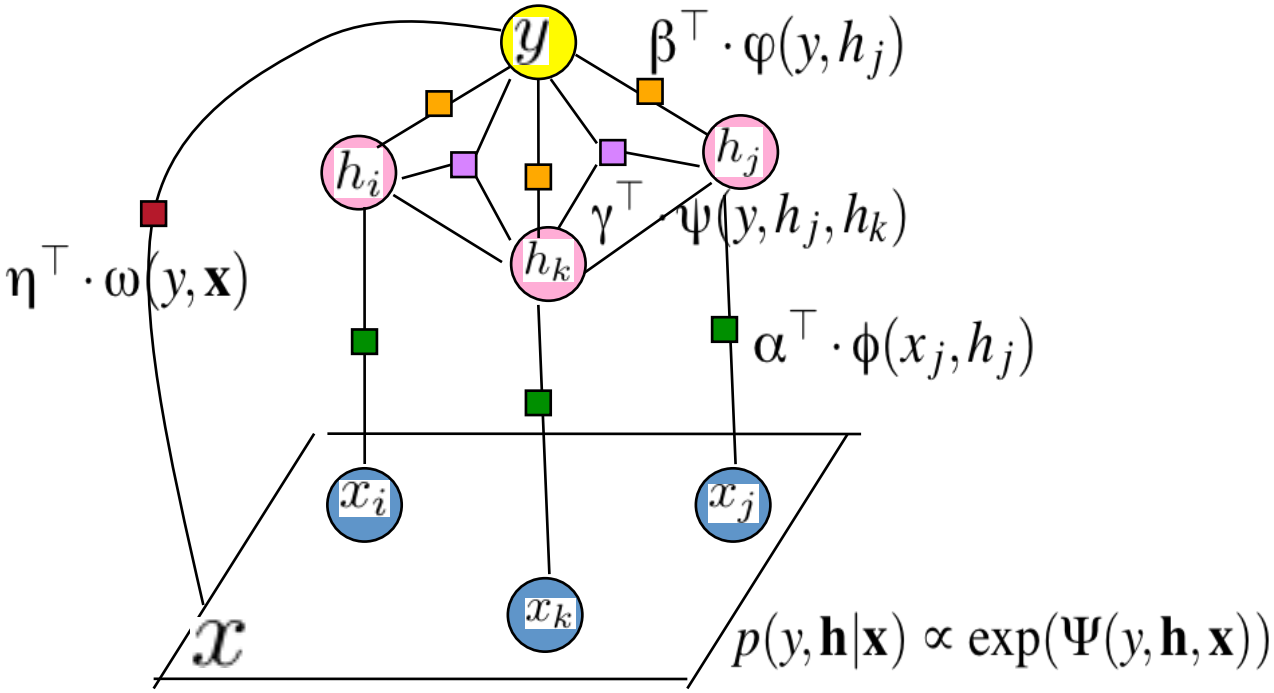


$$p(y, \mathbf{h}|\mathbf{x}) \propto \exp(\Psi(y, \mathbf{h}, \mathbf{x}))$$

Quattoni et al. PAMI 2007
Wang & Mori PAMI 2011

Inferred Part Labels

Learning hCRF Parameters



- Conditional likelihood
  - Integrate out latent part labels h
- Max-margin
  - Examine best setting for latent part labels h
  - Latent-SVM (Felzenszwalb et al. CVPR08), MI-SVM (Andrews et al. NIPS03)

Conditional Likelihood vs. Max-Margin

Weizmann dataset

Method	|H| = 6	|H| = 10	|H| = 20
hCRF-CL	91.7	97.2	94.4
hCRF-MM	97.2	100	97.2

KTH dataset

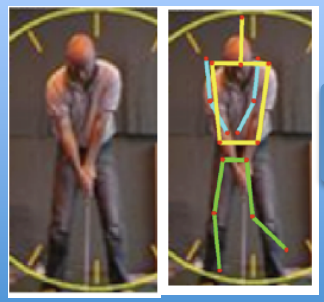
Method	|H| = 6	|H| = 10	|H| = 20
hCRF-CL	78.5	87.6	75.1
hCRF-MM	84.8	92.5	89.7

CL $$\log \sum_{\mathbf{h}} p(Y = y^t, \mathbf{h}|\mathbf{x}^t) \text{ vs. } \log \sum_{\mathbf{h}} p(Y \neq y^t, \mathbf{h}|\mathbf{x}^t)$$

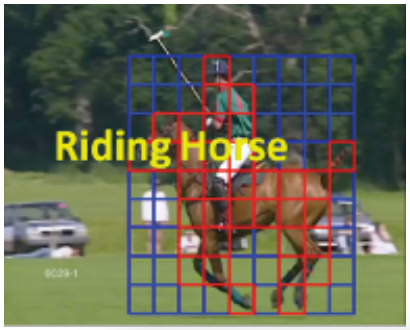
MM $$\max_{\mathbf{h}} p(Y = y^t, \mathbf{h}|\mathbf{x}^t) > \max_{\mathbf{h}} p(Y \neq y^t, \mathbf{h}|\mathbf{x}^t)$$

Outline

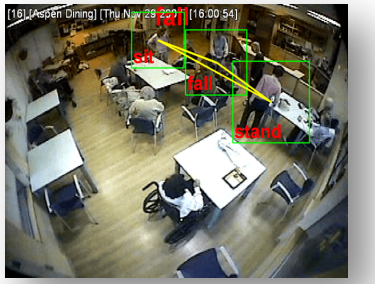
- Latent pose estimation
  - Yang et al. CVPR 2010



- Action localization and recognition
  - Lan et al. ICCV 2011



- Group activity recognition with context
  - Lan et al. NIPS 2010



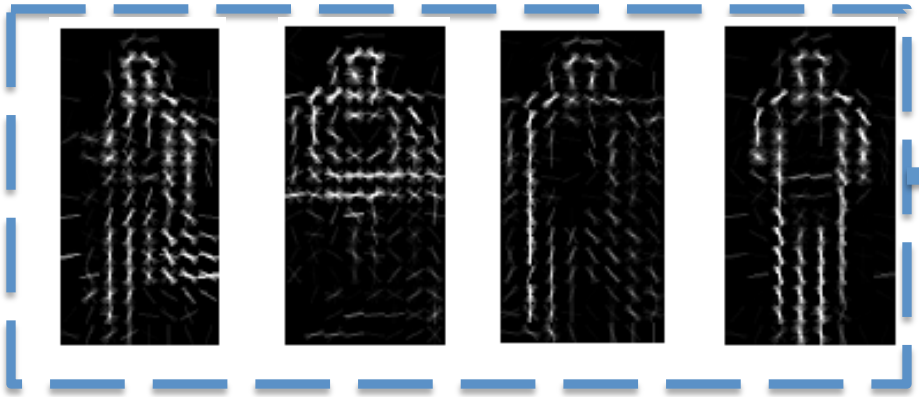
Goal

- Action recognition from still images
  - News/sports image retrieval and analysis
  - An important cue for video-based action recognition

Previous work

- Global template-based representation

  e.g. Wang et al. CVPR06, Ikizler-Cinbis et al. ICCV09



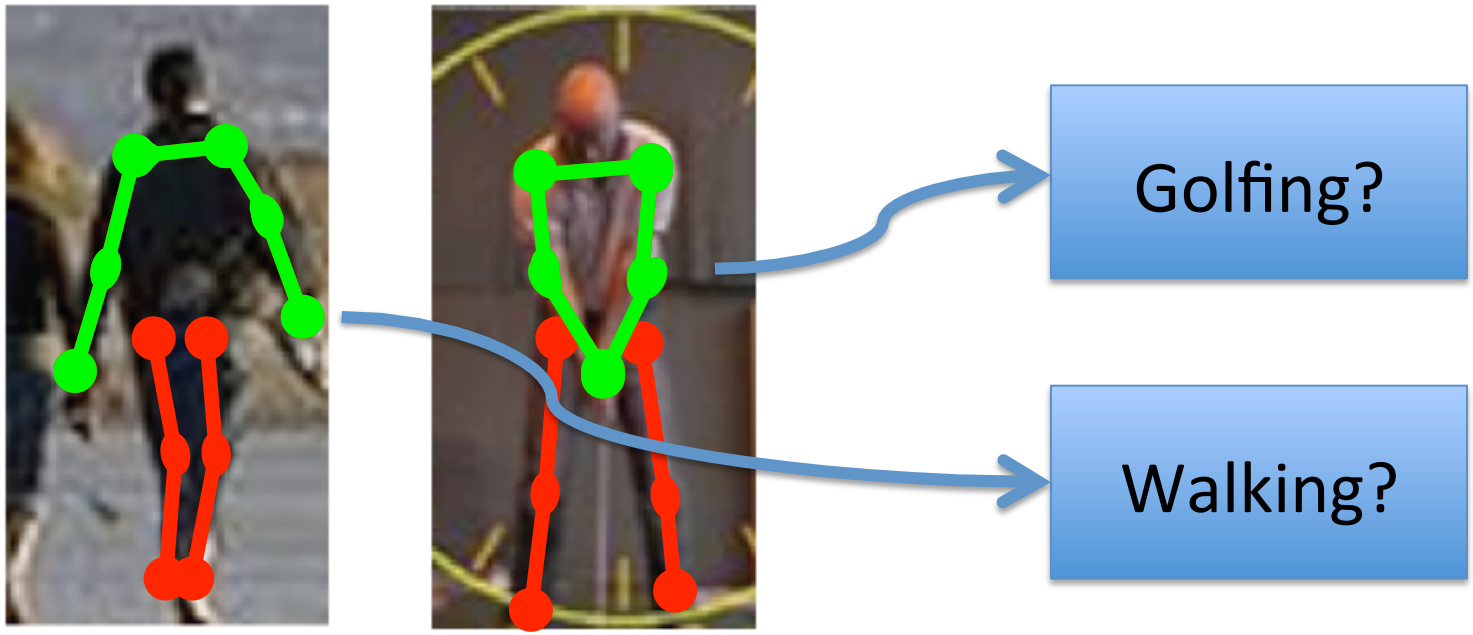
Action Label

- Pose estimation + action recognition

  e.g. Ramanan and Forsyth NIPS03, Ferrari *et al.* CVPR09

Action Label

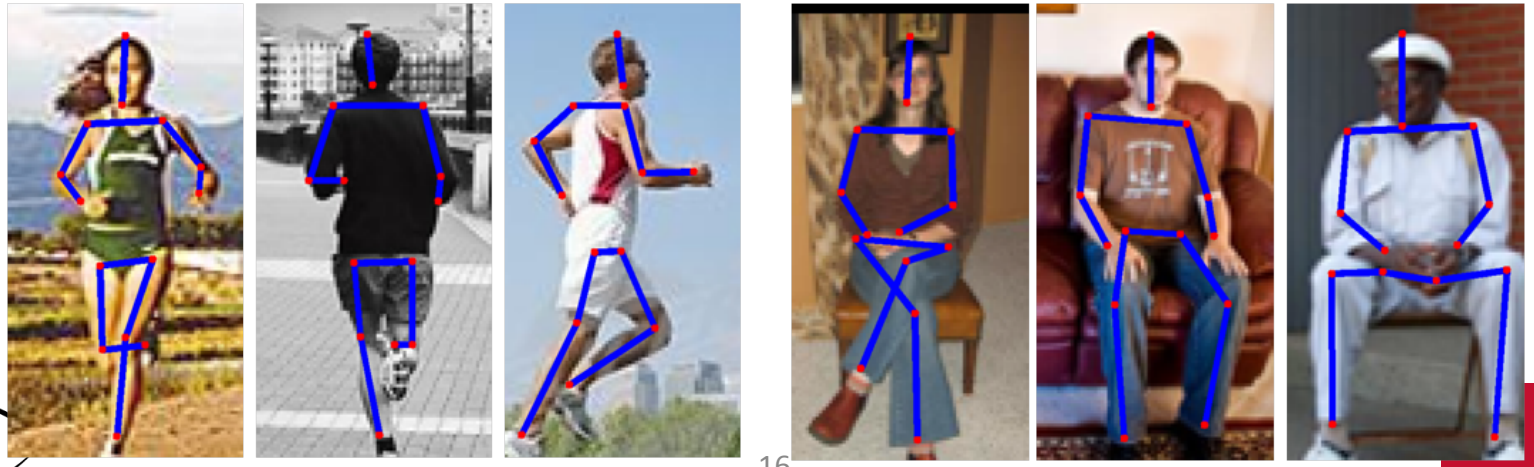
Discriminative Pose



- Not all elements of pose are equally important
- Develop integrated learning framework to estimate pose for action recognition

Pose Representation

- We use a coarse non-parametric pose representation
  - An action-specific variant of the *poselet* [Bourdev & Malik ICCV09]
- A *poselet* is a set of patches not only with similar pose configuration, but also from the same action class.



Poselets

- Poselets obtained by clustering ground-truth joint positions of body parts for each action

Model Formulation

- Develop a scoring function $H(I, Y; \Theta)$
  - Should have high score for correct action label Y
  - Low score for other action labels
  - Model parameters Θ

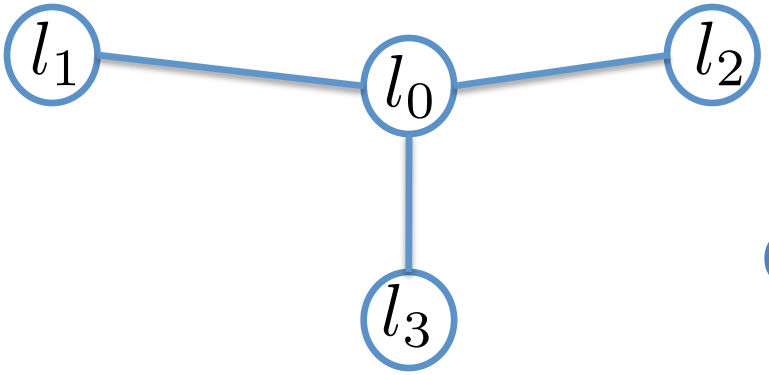
I

Model Formulation

Action Label

Y

Pose



Choose best pose L

Image

I

$$H(I, Y; \Theta) = \max_{L} \Theta^T \Psi(I, L, Y)$$

Model Formulation

Action Label

Running

Pose

Image

I

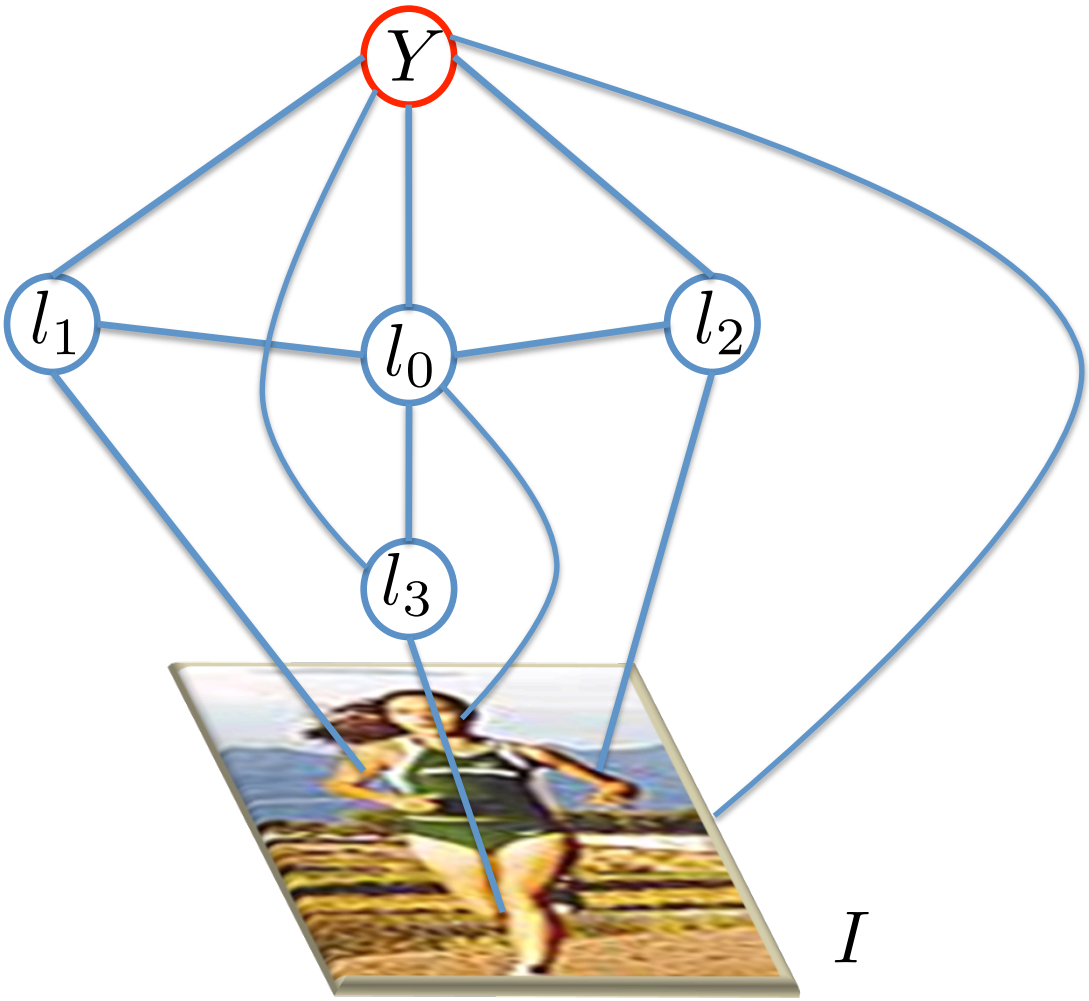
Large score for $H(I, Y = Running; \Theta)$

Full Model

Action Label

Pose

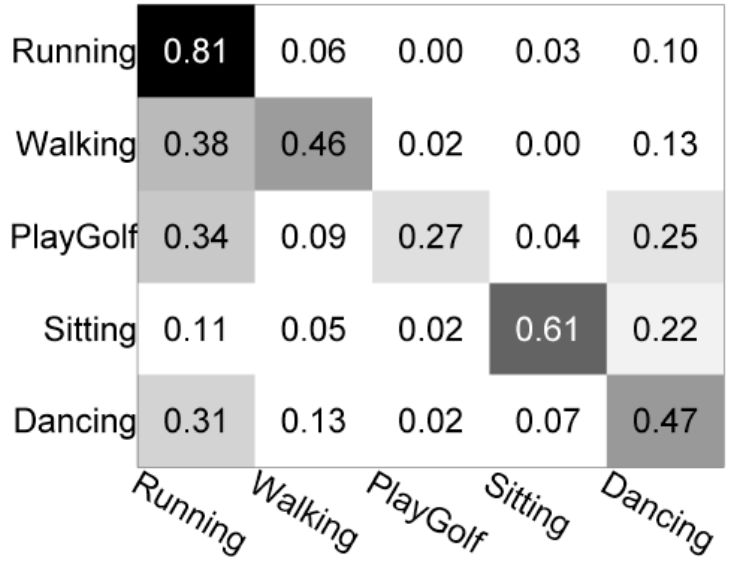
Image

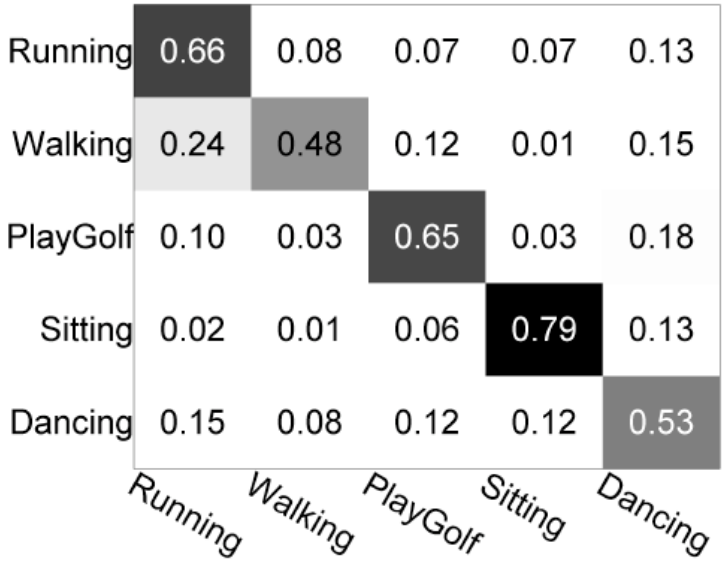


Model parameters learned using max-margin

Experiments

- Still image action dataset
  - Five action categories
  - 2458 images total
  - Train using 1/3 of images from each category

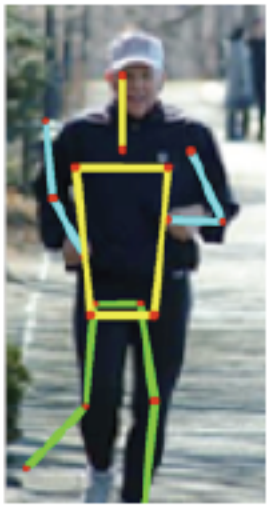


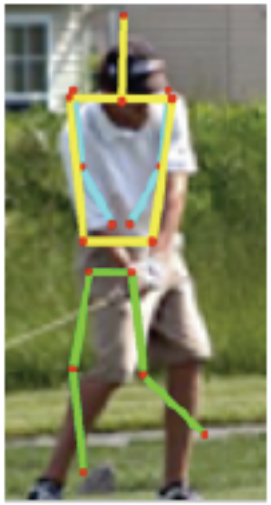


Baseline – HOG/SVM:
52% per class accuracy

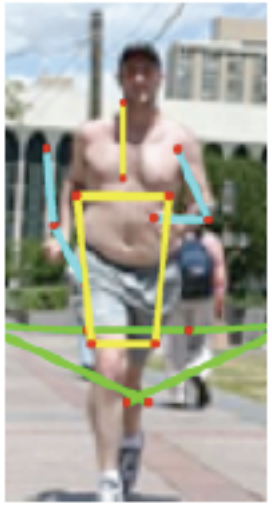
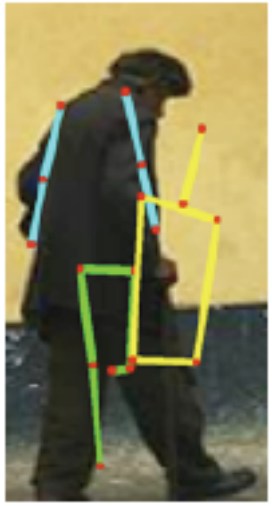
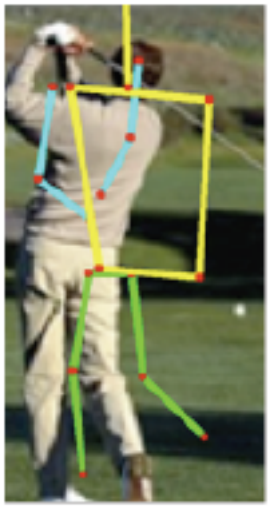
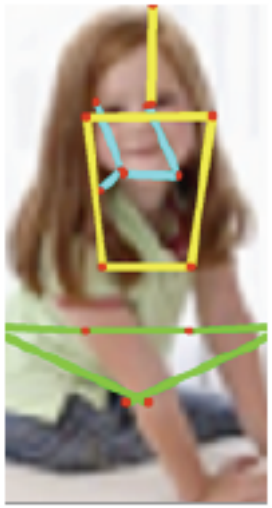
Ours – Latent Pose:
62% per class accuracy

Visualization of latent pose





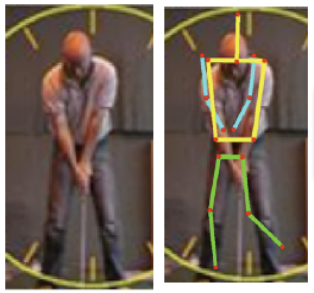

Successful classification examples

Unsuccessful classification examples

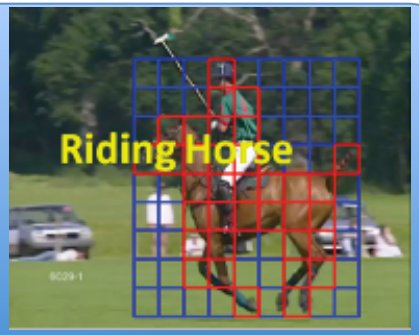
Outline

- Latent pose estimation
  - Yang et al. CVPR 2010

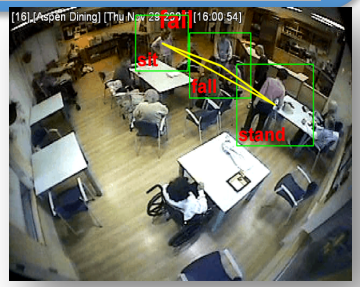


Golfing

- Action localization and recognition
  - Lan et al. ICCV 2011



- Group activity recognition with context
  - Lan et al. NIPS 2010



Action Recognition from Videos

- Statistical Approach (bag-of-words)
  - Laptev et al CVPR 08
  - Neibles & Fei-Fei IJCV 08
  - Ryoo & Aggarwal ICCV 09

  […]

Spatial arrangement of features?
Explicit modeling of human figure?

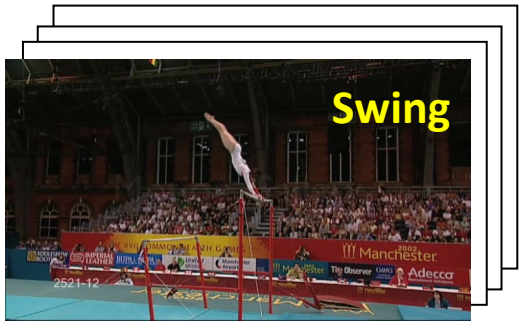
- Structural Approach (figure-centric)
  - Efros et al ICCV 03
  - Shechtman & Irani CVPR 05

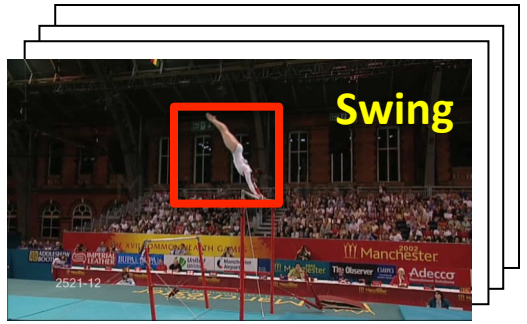
  […]

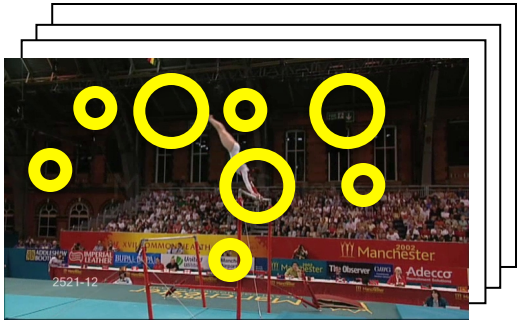
Reliable human detectors?

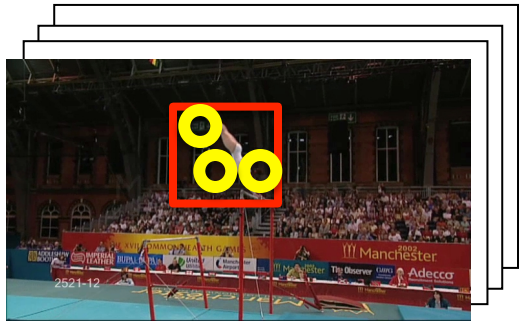
Our method – joint action recognition and localization

Task

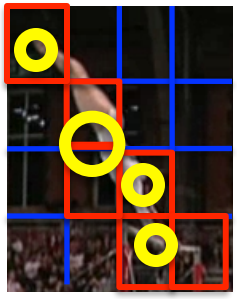








Representation



Approach

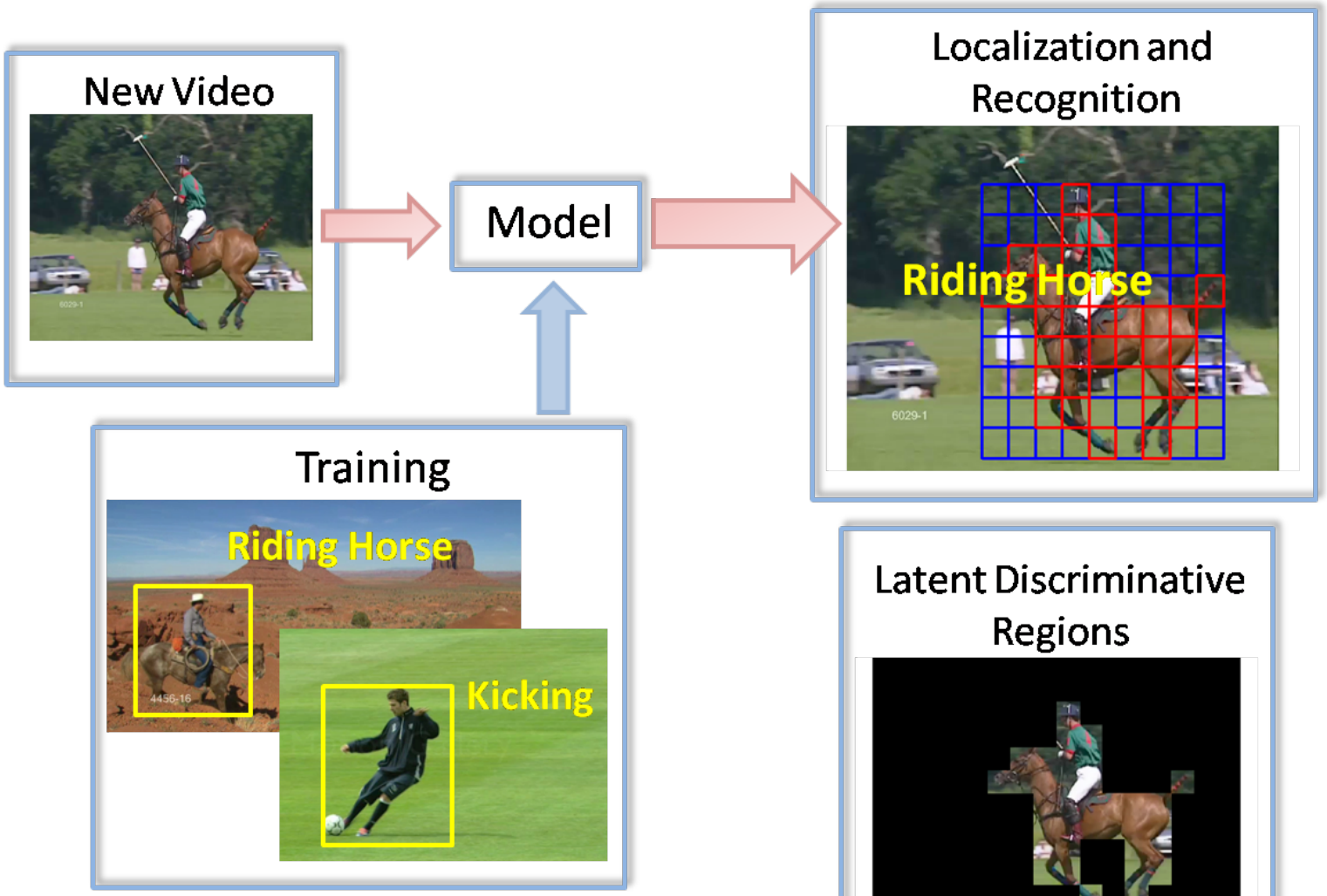

New Video
Model
Localization and Recognition
Riding Horse
Training
Riding Horse
Kicking
Latent Discriminative Regions

Figure-Centric Video Sequence Model

$$\theta^\top \Phi(\mathbf{z}, L, y, \mathbf{I}) = \sum_{i \in \mathcal{V}} \alpha^\top \phi(l_i, \mathbf{z}_i, y, I_i)$$

$$+ \sum_{i,i+1 \in \mathcal{E}} \beta^\top \psi(l_i, l_{i+1}, \mathbf{z}_i, \mathbf{z}_{i+1}, I_i, I_{i+1}) + \gamma^\top \varphi(y, \mathbf{I})$$

Unary Potential -- action model for a frame I_i

l_i : configuration of a bounding box

$\mathbf{z}_i$: {0,1} whether a cell should be selected or not

y: action label

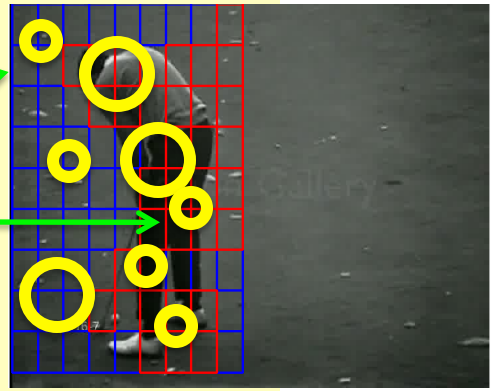


Figure-Centric Video Sequence Model

$$\theta^\top \Phi(\mathbf{z}, L, y, \mathbf{I}) = \sum_{i \in \mathcal{V}} \alpha^\top \phi(l_i, \mathbf{z}_i, y, I_i)$$

$$+ \sum_{i,i+1 \in \mathcal{E}} \beta^\top \psi(l_i, l_{i+1}, \mathbf{z}_i, \mathbf{z}_{i+1}, I_i, I_{i+1}) + \gamma^\top \varphi(y, \mathbf{I})$$

Pairwise Potential -- a tracking constraint between neighboring frames

- similarity of bounding boxes
- similarity of discriminative regions
- similarity of patch appearances

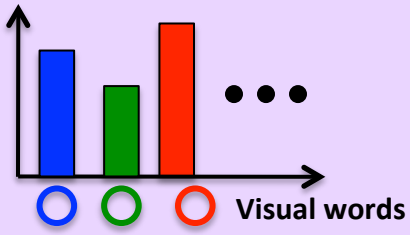
Figure-Centric Video Sequence Model

$$\theta^\top \Phi(\mathbf{z}, L, y, \mathbf{I}) = \sum_{i \in \mathcal{V}} \alpha^\top \phi(l_i, \mathbf{z}_i, y, I_i)$$

$$+ \sum_{i, i+1 \in \mathcal{E}} \beta^\top \psi(l_i, l_{i+1}, \mathbf{z}_i, \mathbf{z}_{i+1}, I_i, I_{i+1}) + \gamma^\top \varphi(y, \mathbf{I})$$

Global Action Potential – action model for a video **I**

Bag-of-words representation for a video:



Max-Margin Learning

Training data: $\{\mathbf{I}^n, L^n, y^n\}$

$$\min_{\theta,\xi\geq 0} \frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\xi^n$$

$$\text{s.t.} \quad f_\theta(y^n, L^n, \mathbf{I}^n) - f_\theta(y, L, \mathbf{I}^n) \geq \Delta(y, y^n, L, L^n) - \xi^n, \forall n, \forall y, \forall L$$

A joint loss on both action recognition and localization

$$\Delta(y, y^n, L, L^n) = \mu\Delta_{0/1}(y, y^n) + (1-\mu)\Delta(L, L^n)$$

Experiment: Dataset

- UCF-Sports dataset [Rodriguez et al. 2008]
  - 150 videos from 10 action categories: diving, golf swinging, kicking, lifting, swinging … (diverse actions, real sports broadcasts)
  - Strong scene correlations among videos, some videos are captured in exactly the same location.
    - ✗ LOO
    - We split the dataset to reduce the chances of videos in the test set sharing the same scene with videos in the training set.

Experiment: Action Recognition

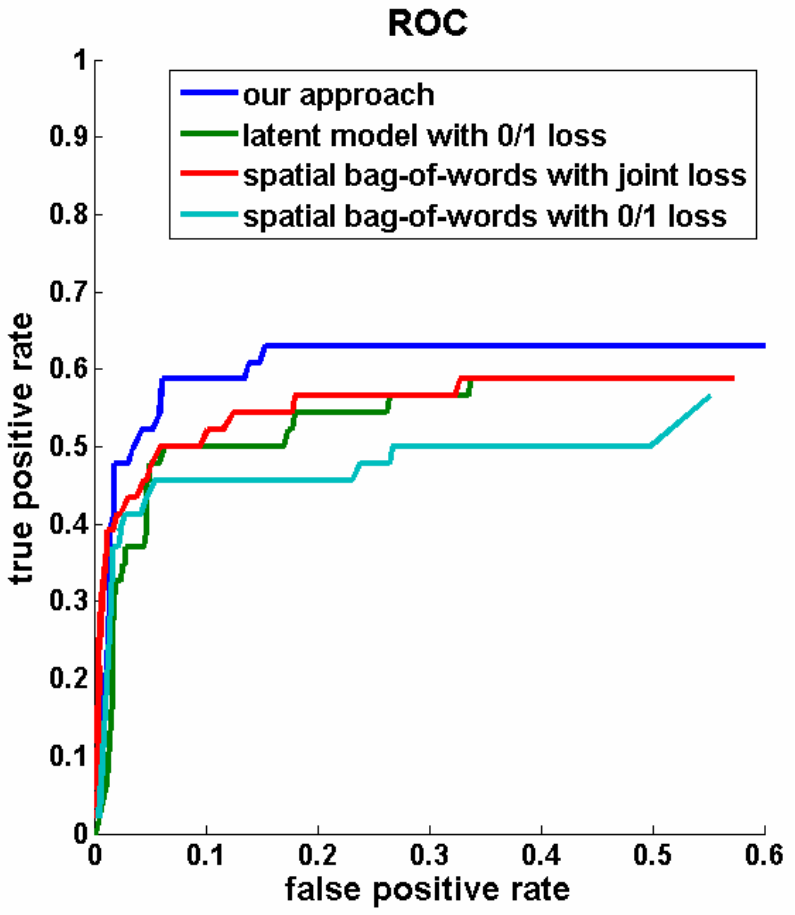
Training / Test Split

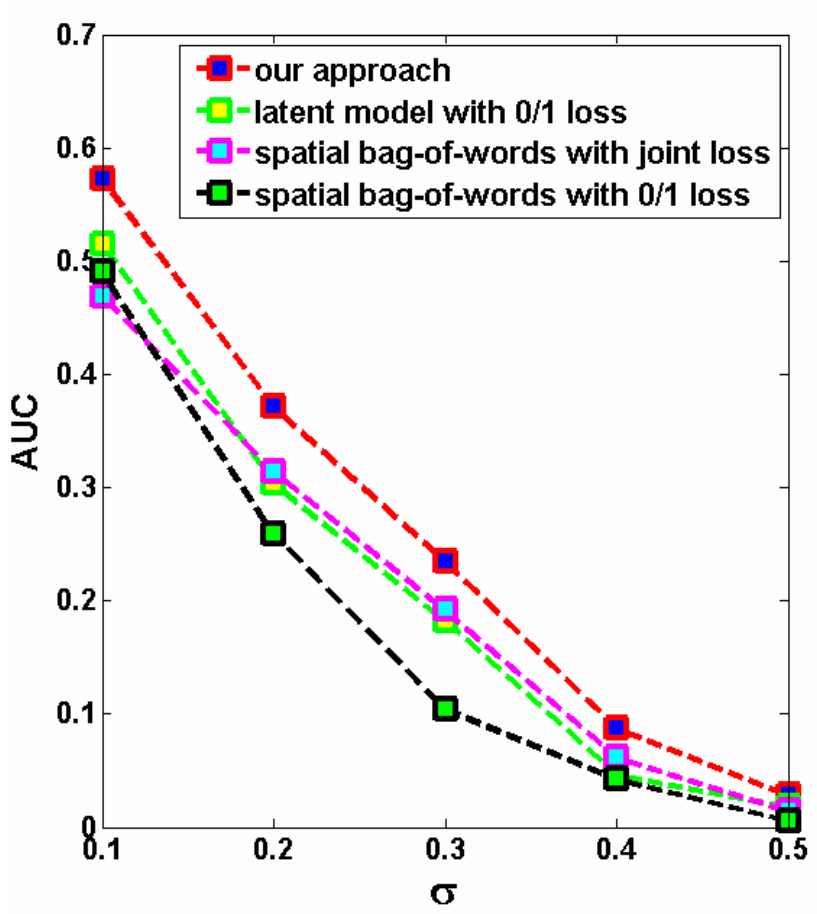
Method	Accuracy
global bag-of-words	63.1
local bag-of-words	65.6
spatial bag-of-words with $\Delta_{0/1}$	63.1
spatial bag-of-words with Δ_{joint}	68.5
latent model with $\Delta_{0/1}$	63.7
latent model with Δ_{joint}	73.1

LOO

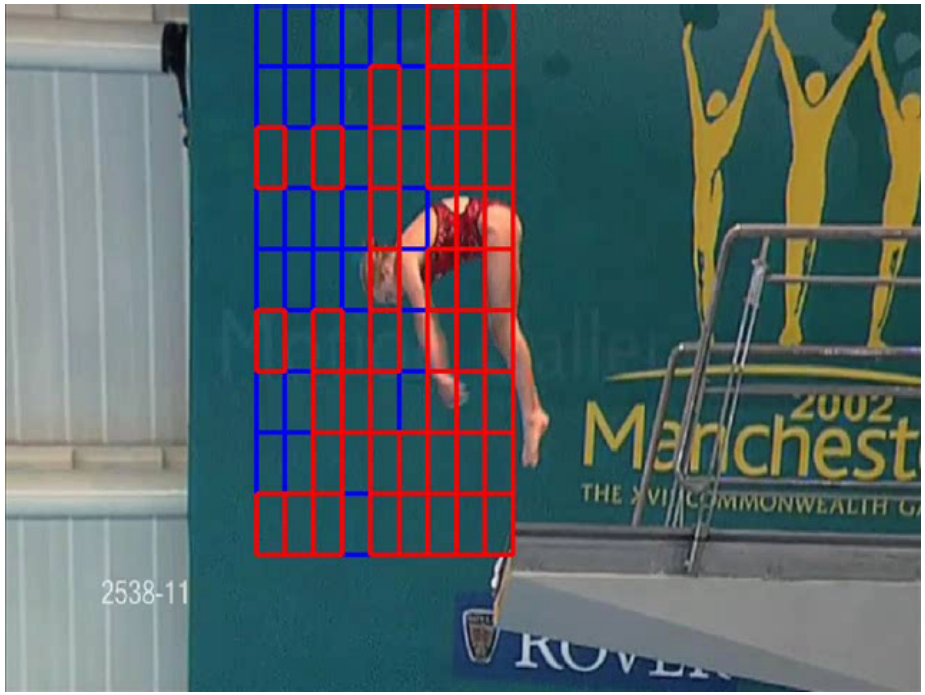
Method	Accuracy
Kovashka et al. 2010	87.3
Wang et al. 2009	85.6
Yeffet & Wolf 2009	79.3
Rodriguez et al. 2008	69.2
global bag-of-words	81.9
Ours	83.7

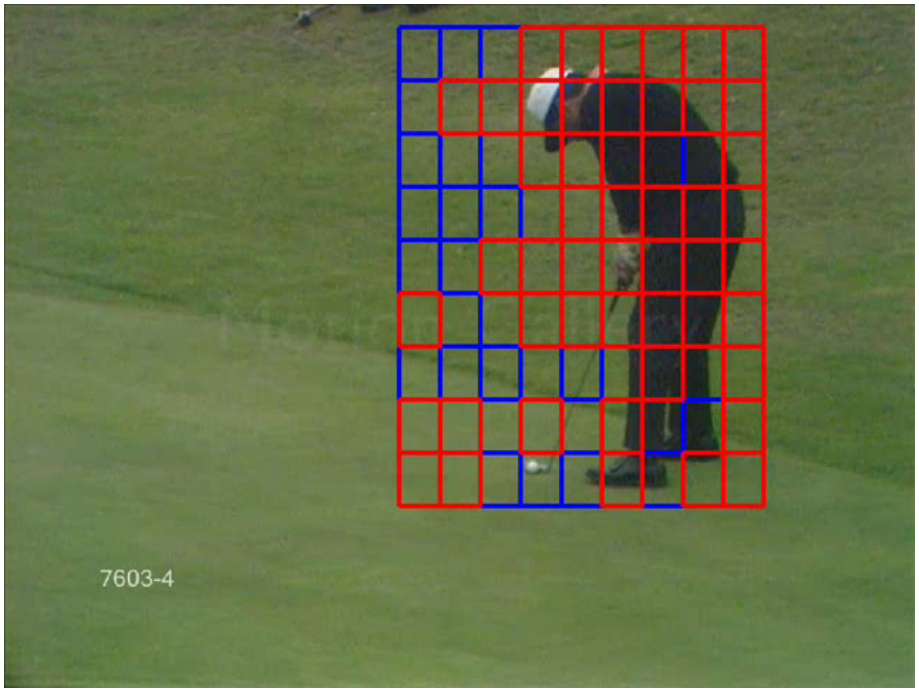
Experiment: Action Localization

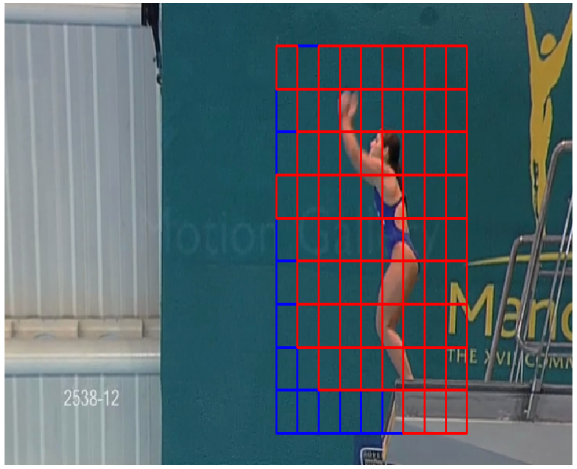


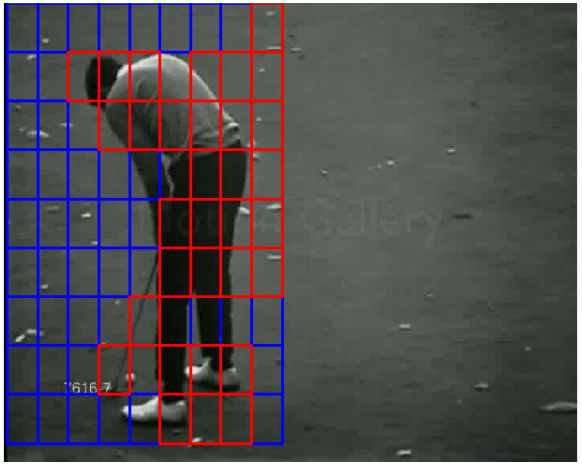


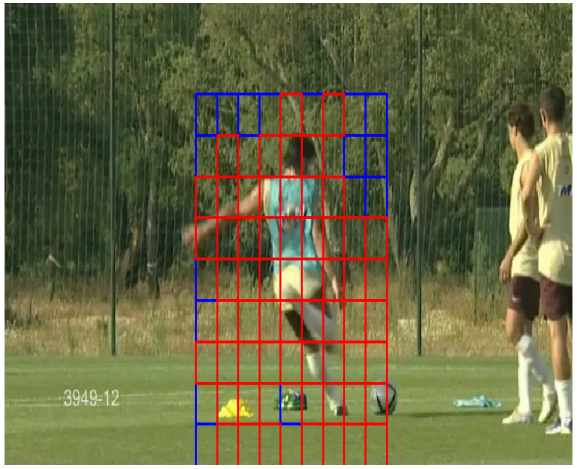
A video is considered as correctly localized
if its intersection-over-union score is larger than σ


2538-11

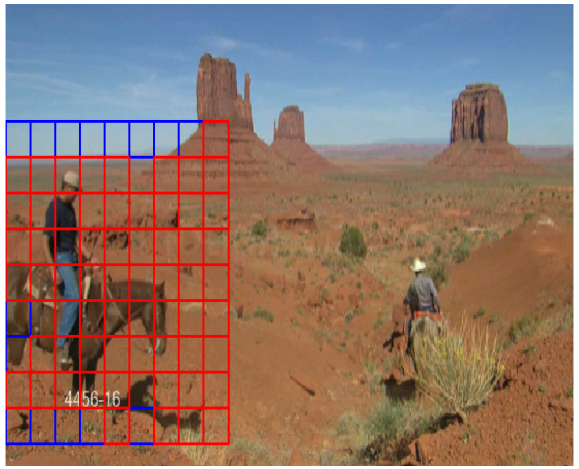

7603-4

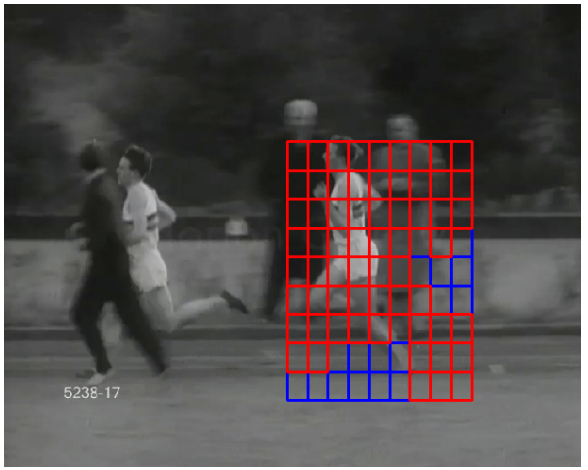

2538-12

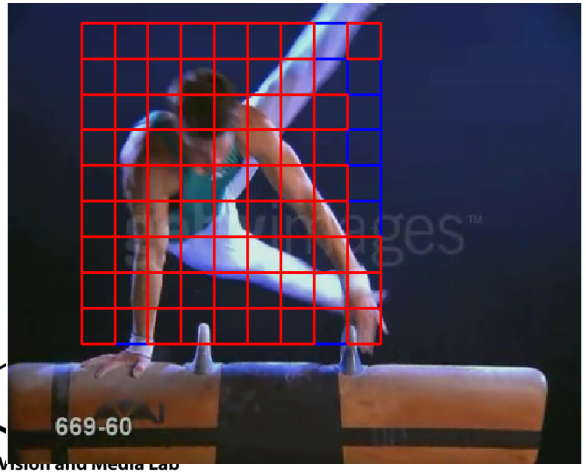


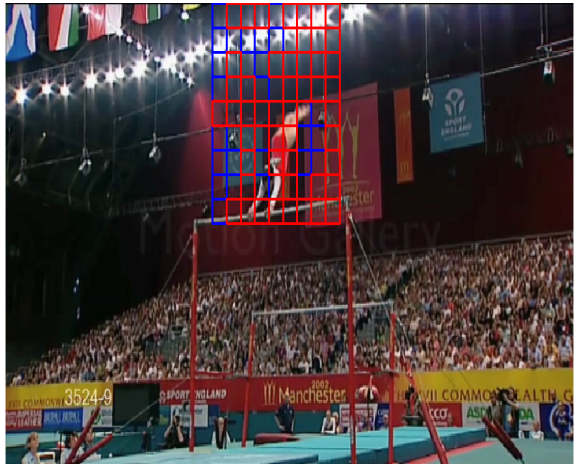

3949-12

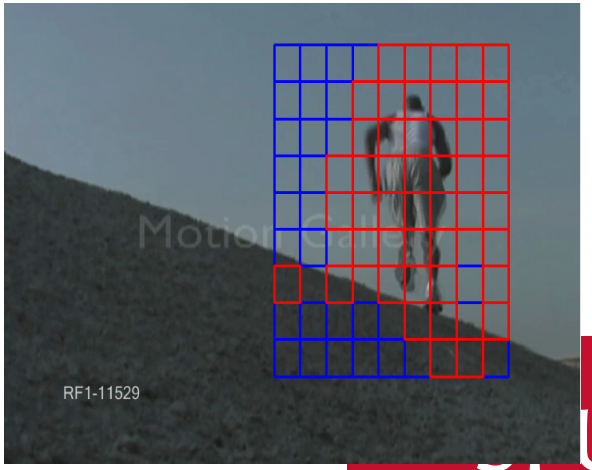

THE XVII COMMONWEALTH GAMES
2002
Manchester
3528-8


4456-16


5238-17

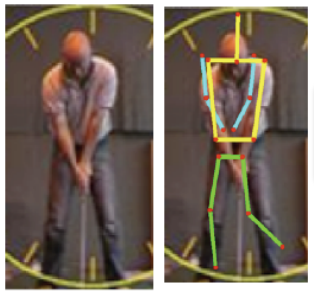

669-60


3524-9
2002
Manchester

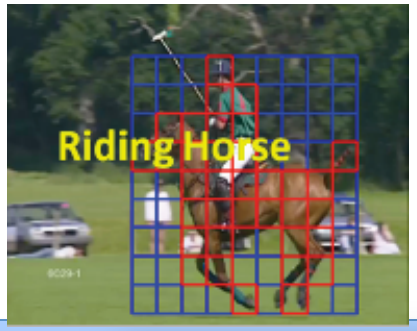

RF1-11529

Outline

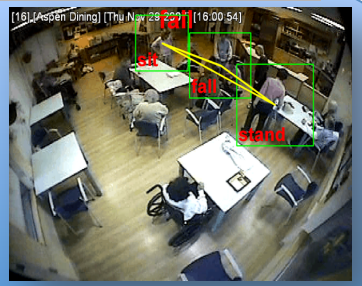
- Latent pose estimation
  - Yang et al. CVPR 2010



- Action localization and recognition
  - Lan et al. ICCV 2011



- Group activity recognition with context
  - Lan et al. NIPS 2010

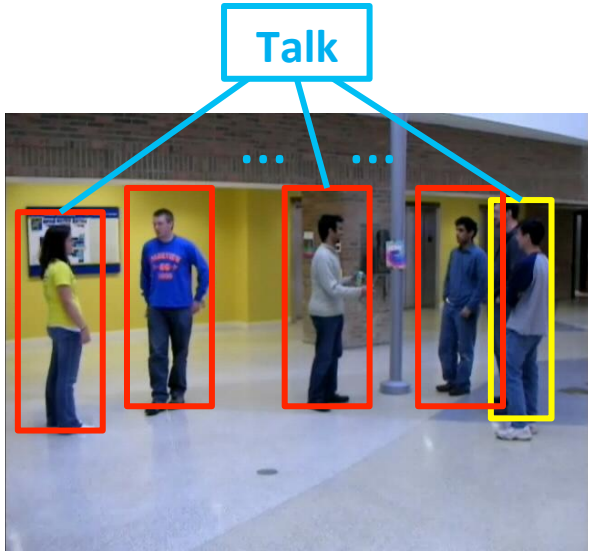


Group Activity Recognition

Context

Group Activity Recognition

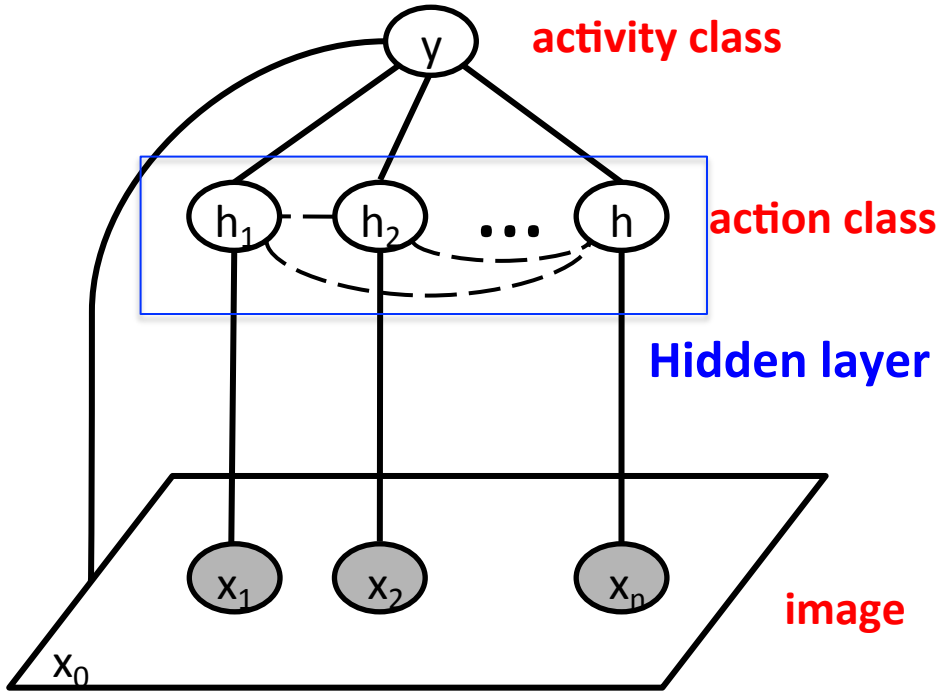
- Two types of Context

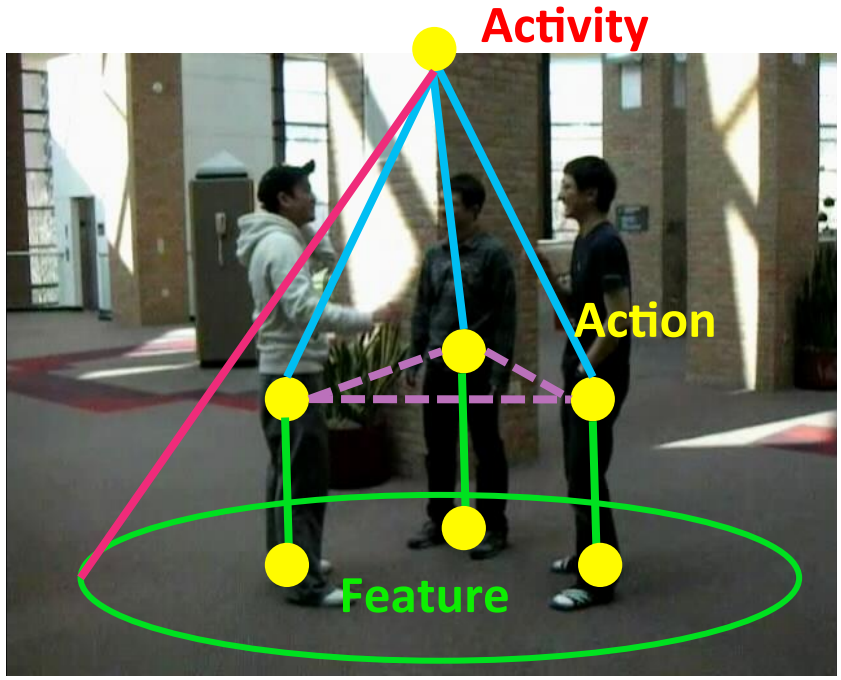


group-person interaction

person-person interaction

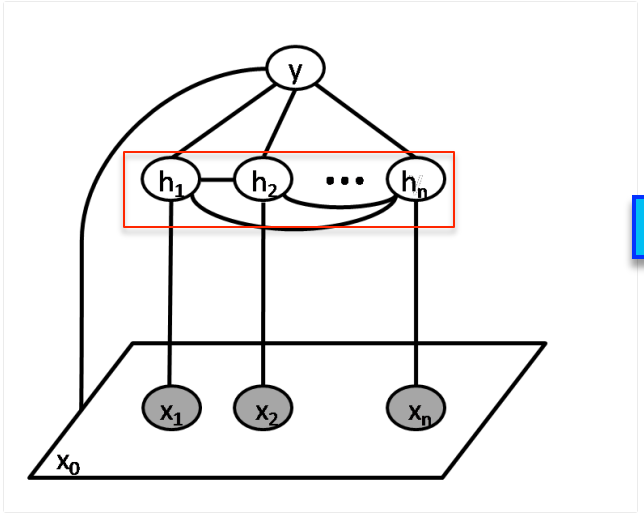
Latent Structured Model

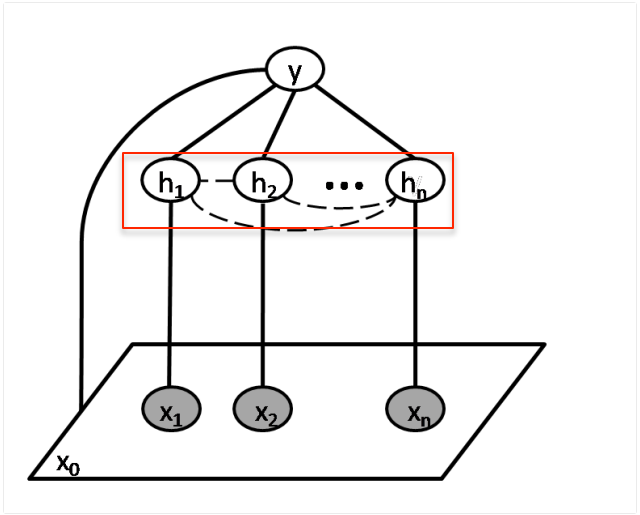




Difference from Previous Work

- Latent Structured Models





Previous work
a pre-defined structure for the hidden layer, e.g. tree (HCRF) (Quattoni et al. PAMI 07, Felzenszwalb et al. CVPR 08)

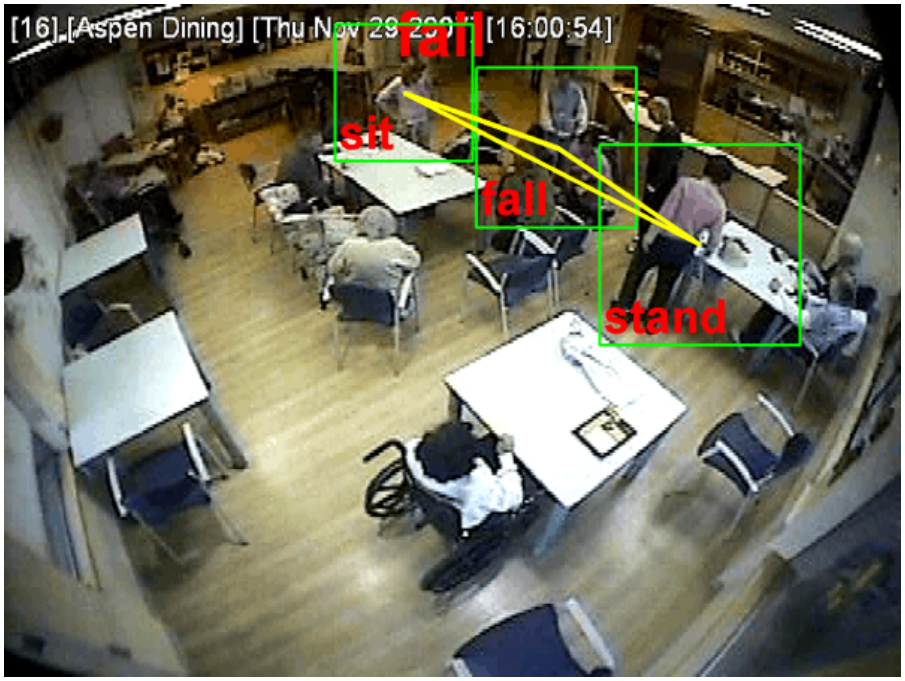
Our work
latent structure for the hidden layer, automatically infer it during learning and inference.

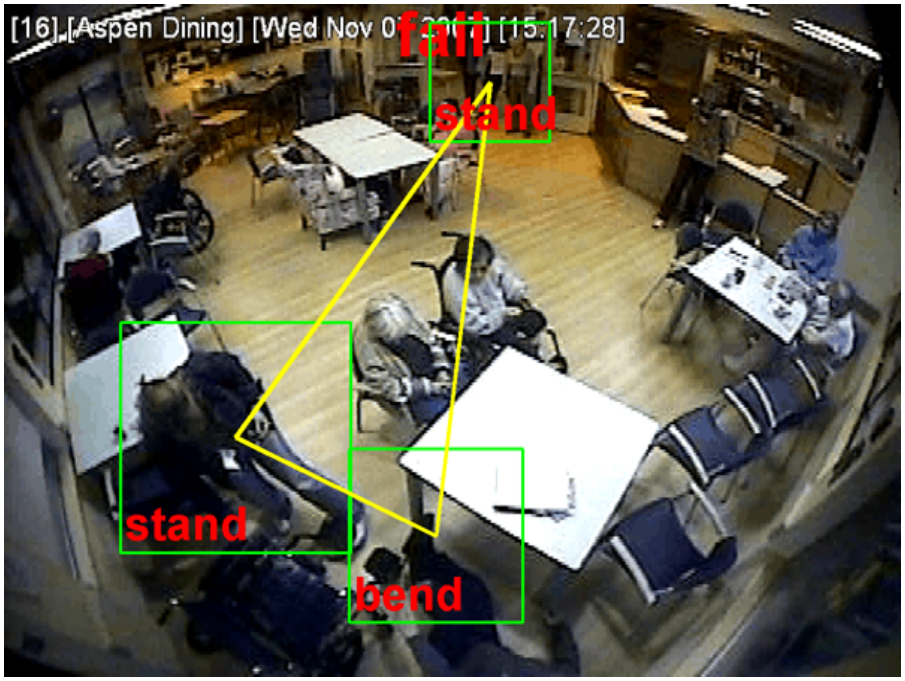
Nursing Home Dataset

Results – Nursing Home Dataset

Method	Mean per-class
root + SVM	52.4
no connection	56.1
minimum spanning tree	62.3
complete graph within r = 100	61.3
complete graph within r = 200	61.1
complete graph within r = 300	64.2
structure-level approach	**67.4**
feature-level approach	**60.3**

Results – Correct Examples



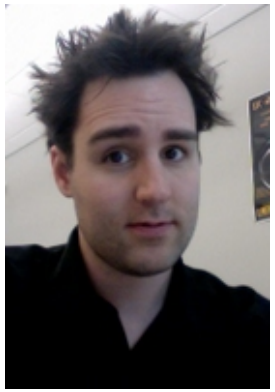


Results – Incorrect Examples

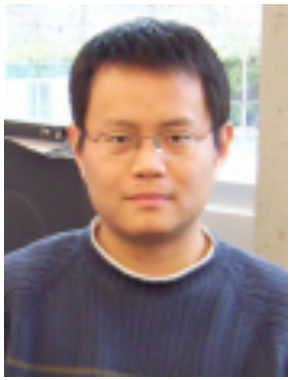
Conclusion

- Structured models
  - Whole versus parts
    - Learning criterion: conditional likelihood vs. max-margin learning
  - Semantically meaningful parts
    - Latent human pose estimation for action recognition
  - Action localization
    - Video model for person location and action label
  - Scene structure
    - Context among people in a scene

Acknowledgements

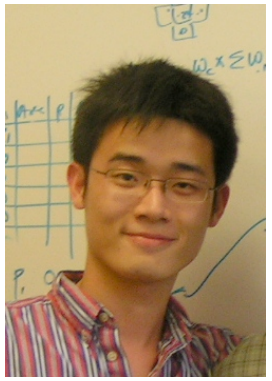


Brian Milligan



Yang Wang

Tian Lan



Weilong Yang

Alex Couture-Beil

Mark Bayazit

Thank you!

Alireza Fathi