Statistical and Structural Recognition of Human Actions

Structural Methods

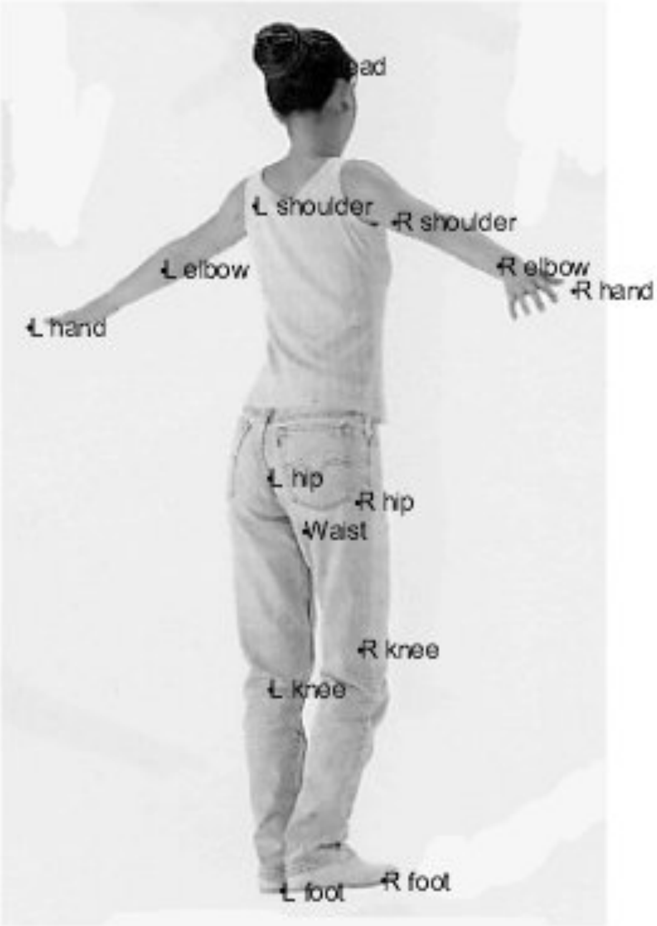
POSE ESTIMATION AND ACTION RECOGNITION

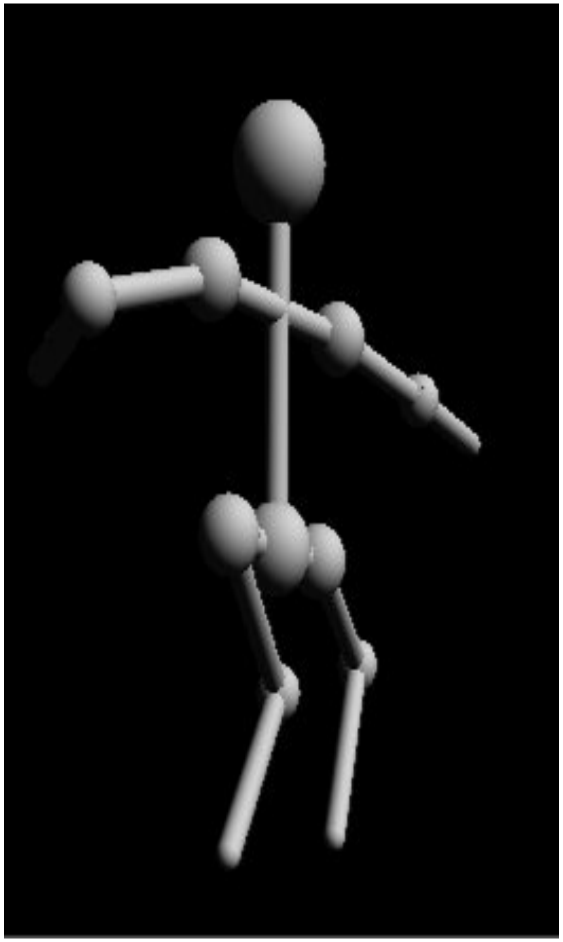
Pose Estimation for Action Recognition

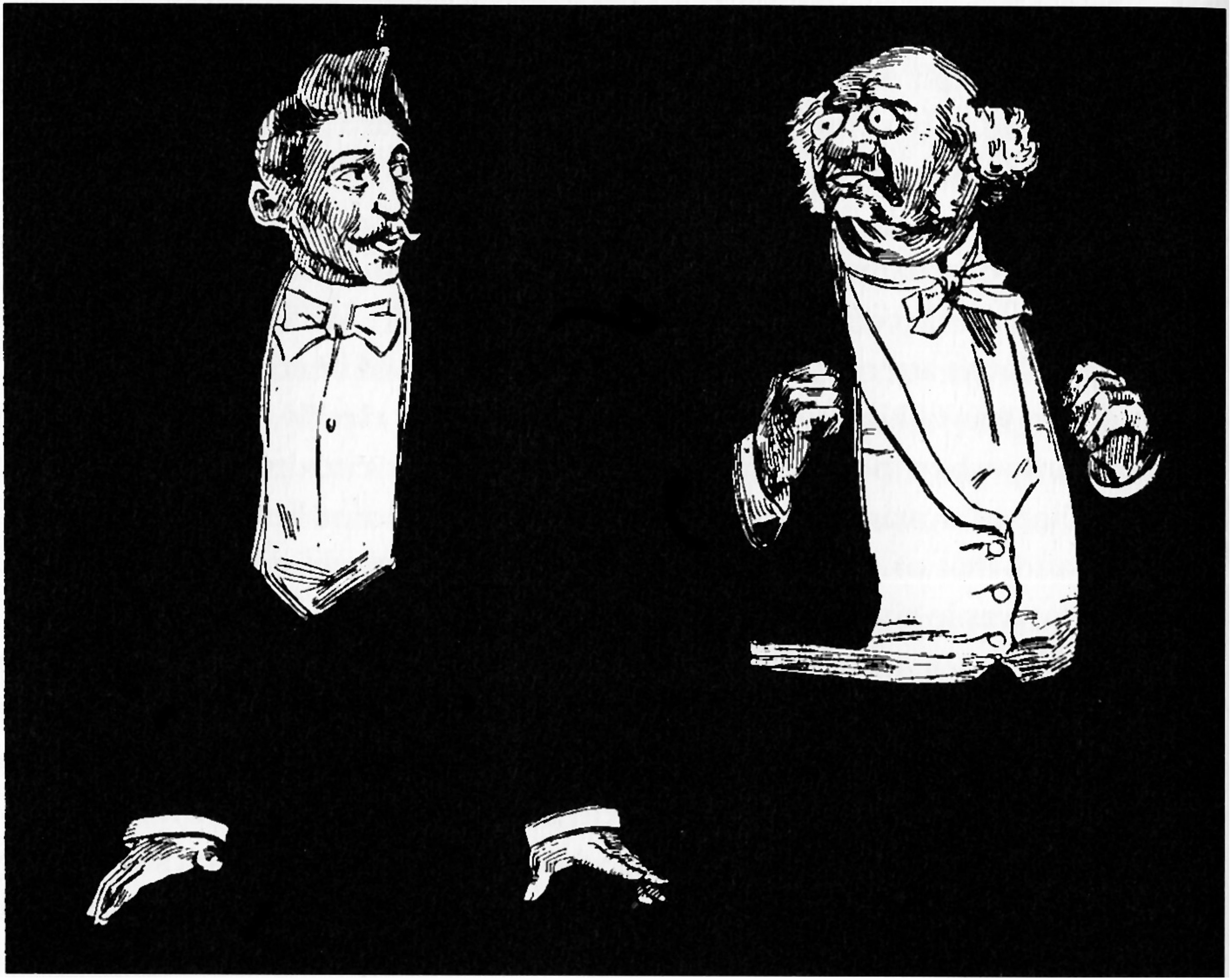
G. Johansson, **Moving Light Displays,** 1973

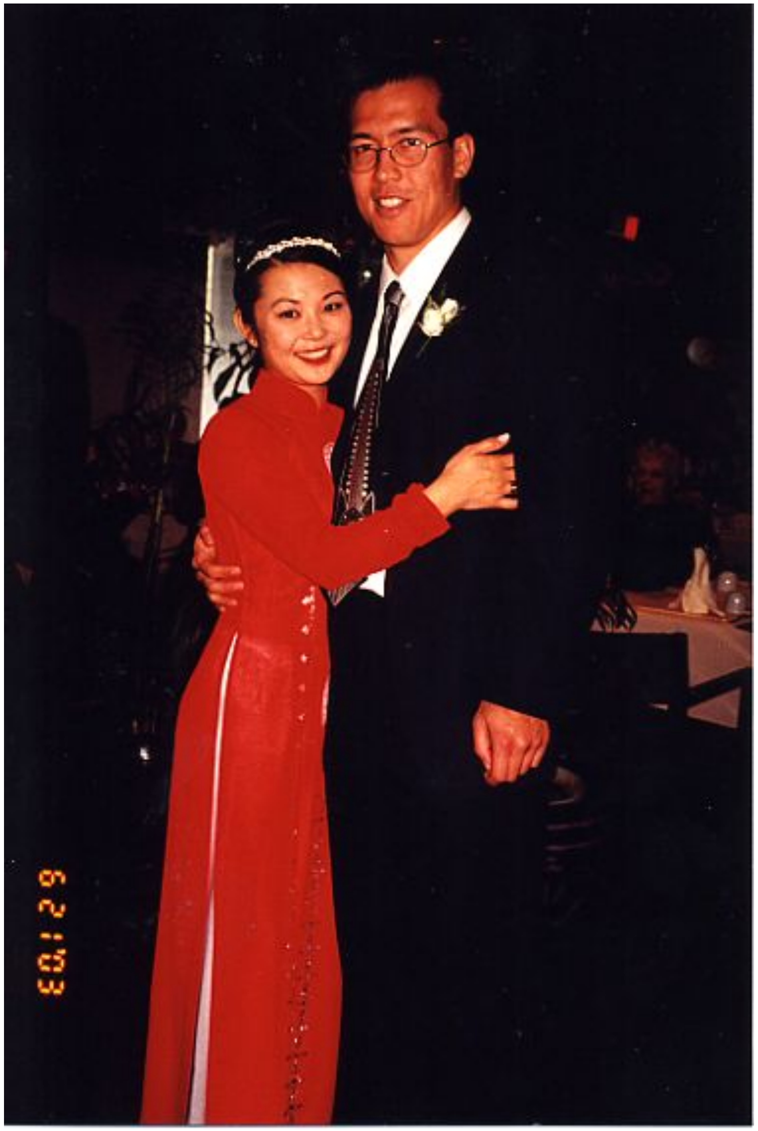
- Pose seems sufficient for certain action categories
- Remove effects of clothing, lighting variation from representation

Pose Estimation – Problem Definition







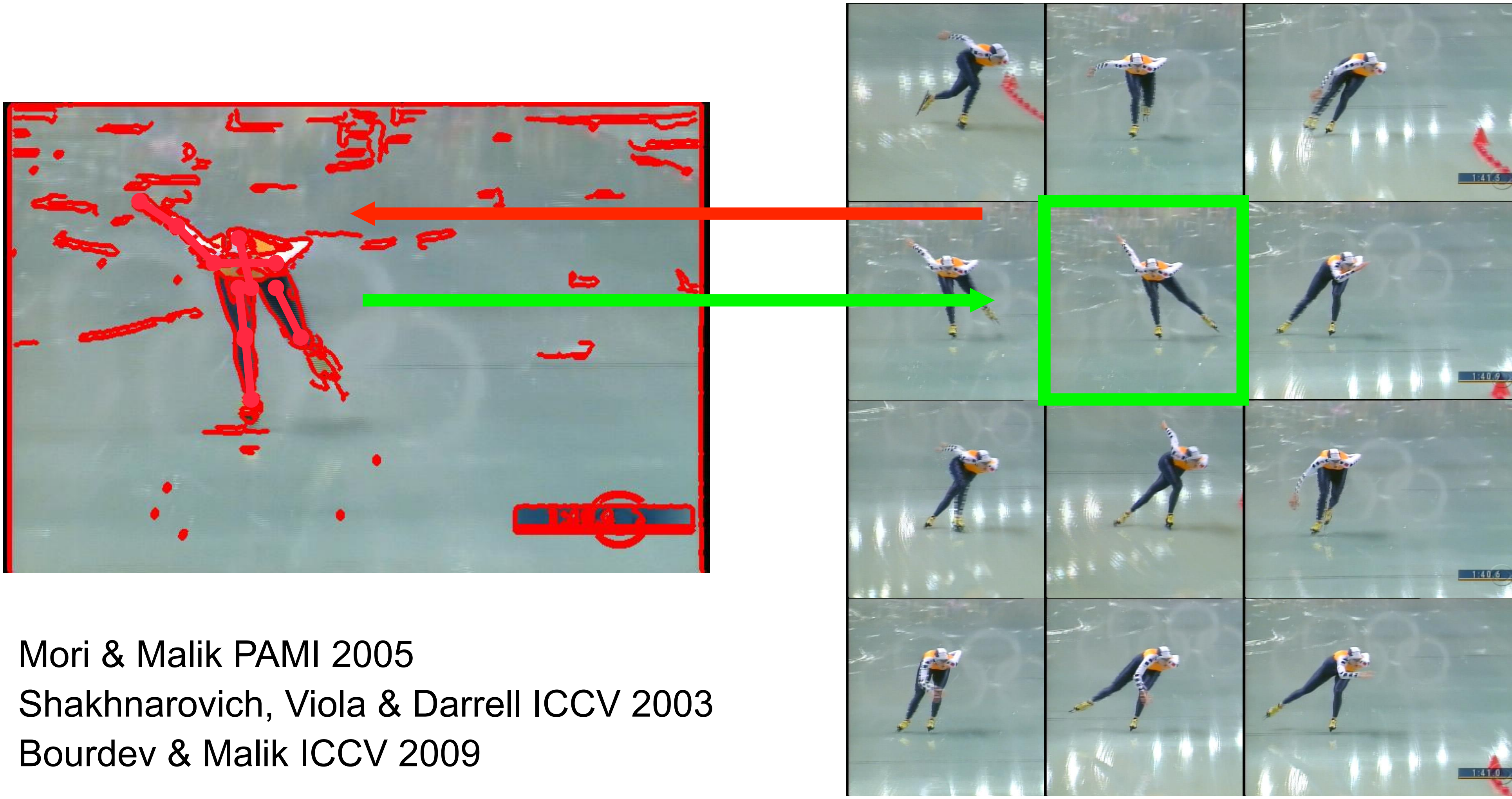

9 2 '03

Problem

Models vs. Exemplars

- Two broad classes of approaches
  - Match templates (exemplar-based)
  - Fit a human body model

Exemplar Matching For Pose Estimation



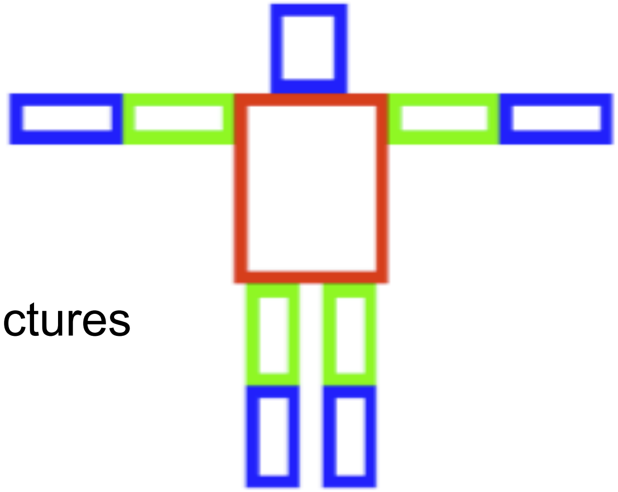
Mori & Malik PAMI 2005

Shakhnarovich, Viola & Darrell ICCV 2003

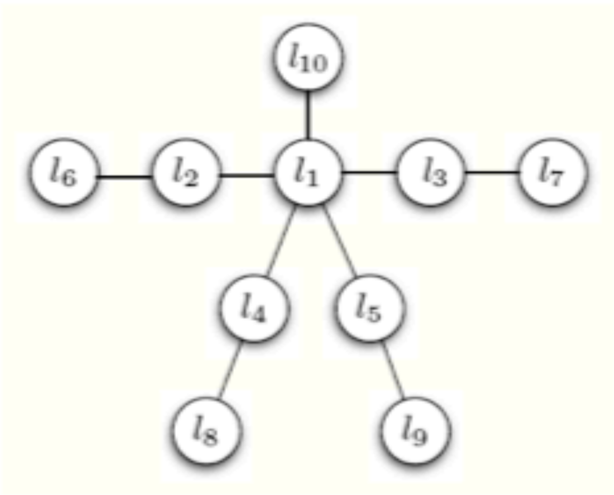
Bourdev & Malik ICCV 2009

Database of Exemplars

Human Body Models for Pose Estimation



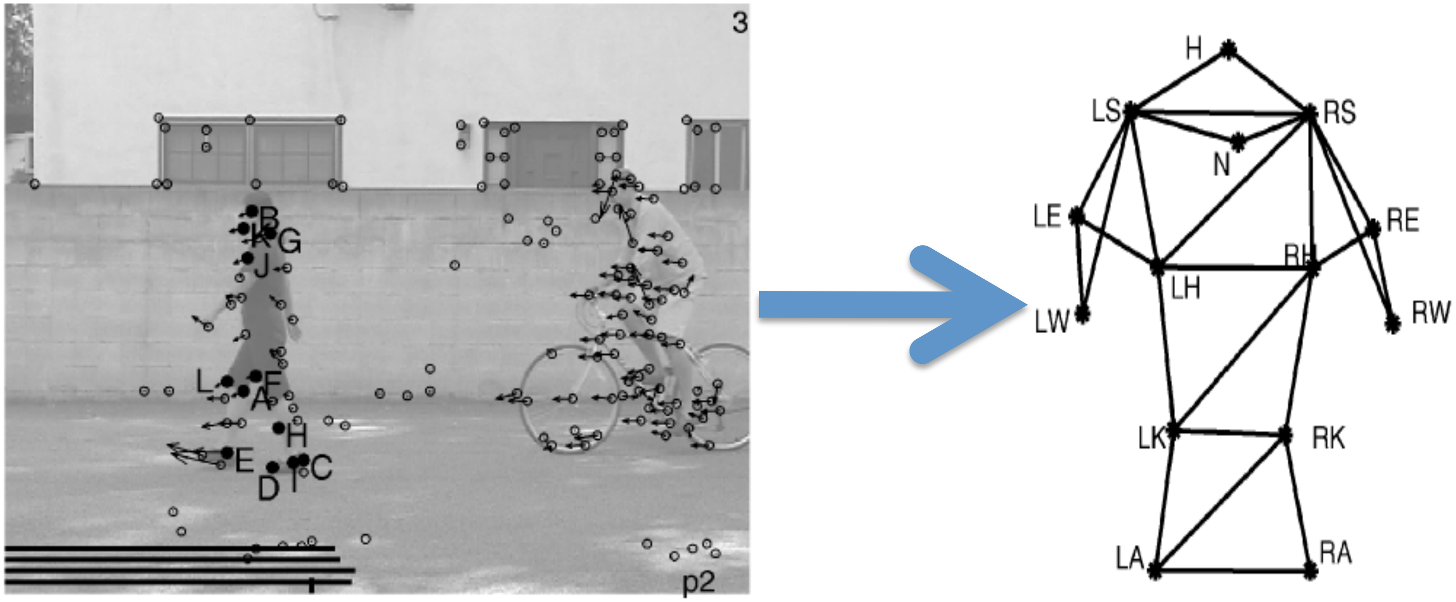
Pictorial Structures model



$$Pr(L|I,\Theta) \propto \exp\left(\underbrace{\sum_{(i,j)\in E} \psi(l_i - l_j)}_{\text{spatial prior}} + \underbrace{\sum_{i=1}^{K} \phi(l_i)}_{\text{part appearance}} \right)$$

Felzenszwalb & Huttenlocher CVPR 2000
Ramanan NIPS 2006
Ferrari, Marin & Zisserman CVPR 2008

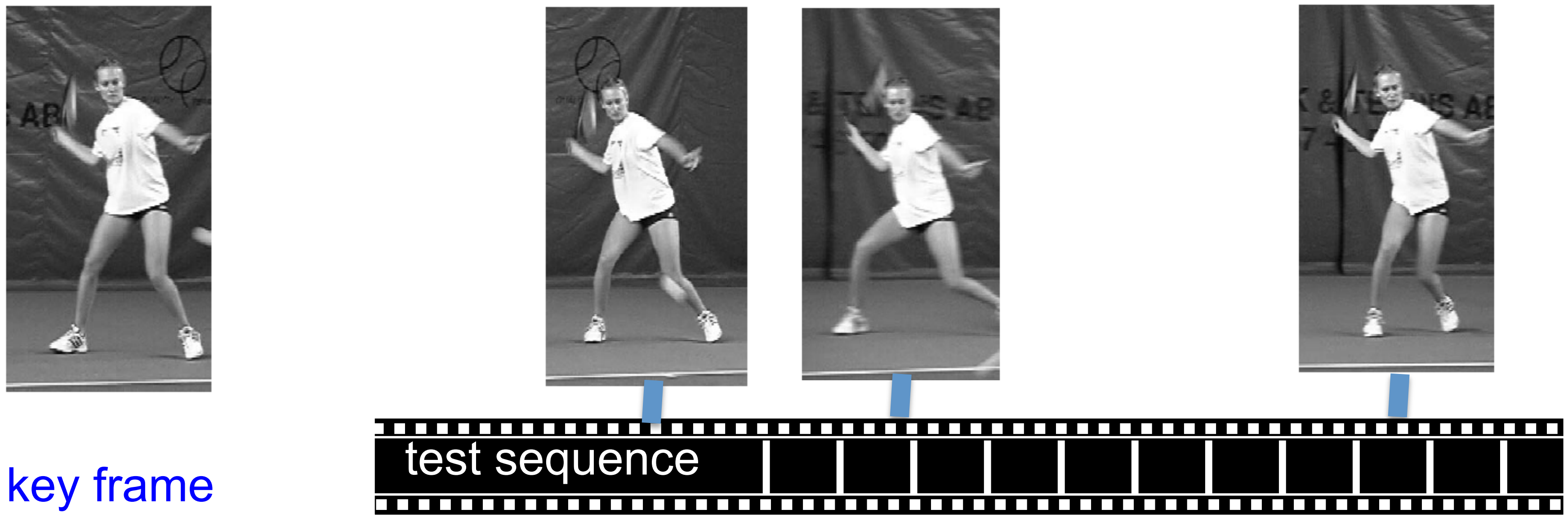
Action from Pose I: Model Likelihood



- Detect corners in images/video
- Assess likelihood under action-specific pose model
- Discriminate between walking directions, bicycle riding

Song, Goncalves & Perona NIPS 2001, PAMI 2003

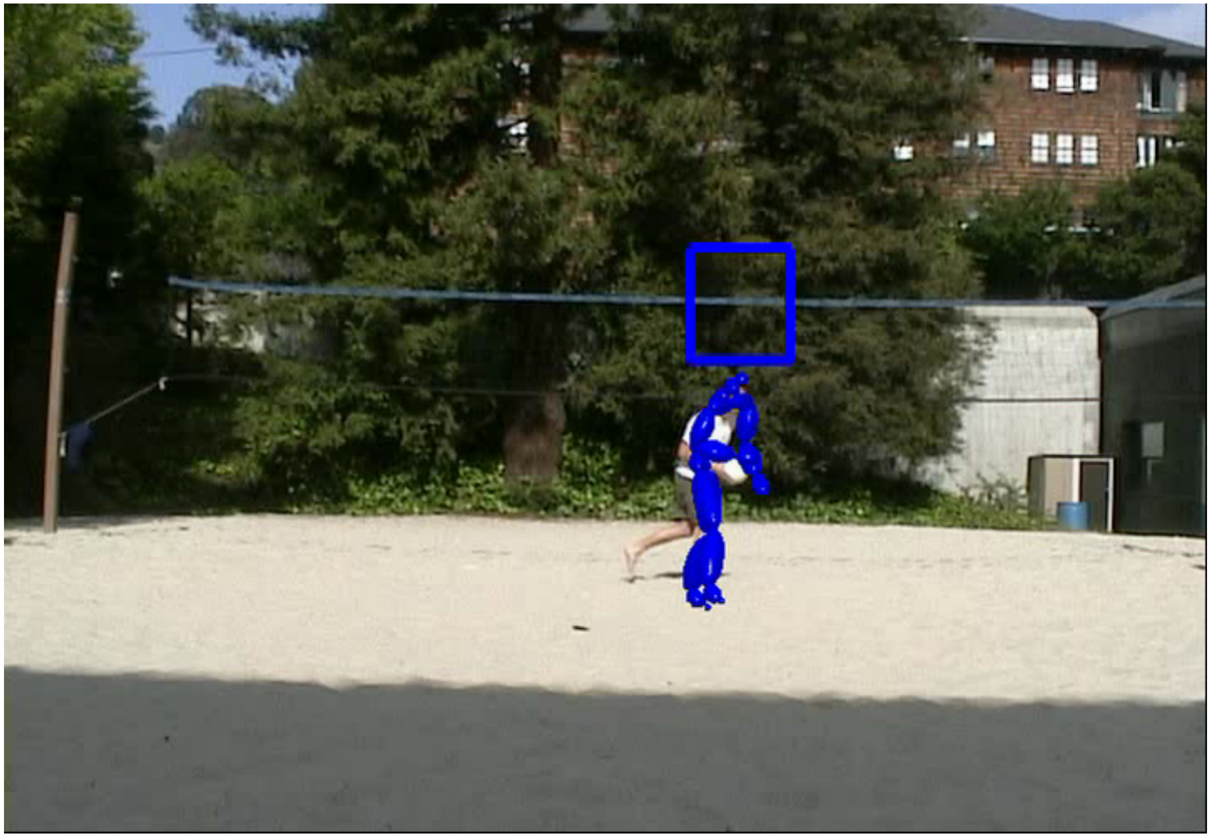
Action from Pose II: Key frame Templates

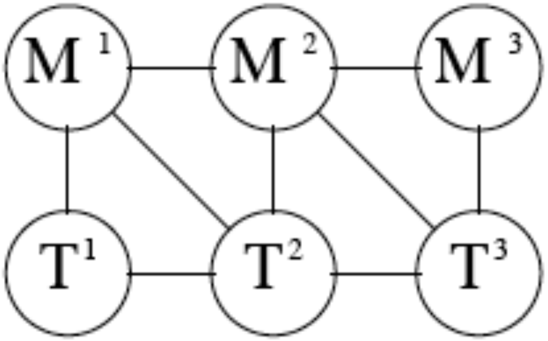


- Key frame matching to test sequence to find similar poses
  - Shape matching on edge maps using *order structure*

Sullivan & Carlsson ECCV 2002

Action from Pose II: Classifier on Pose





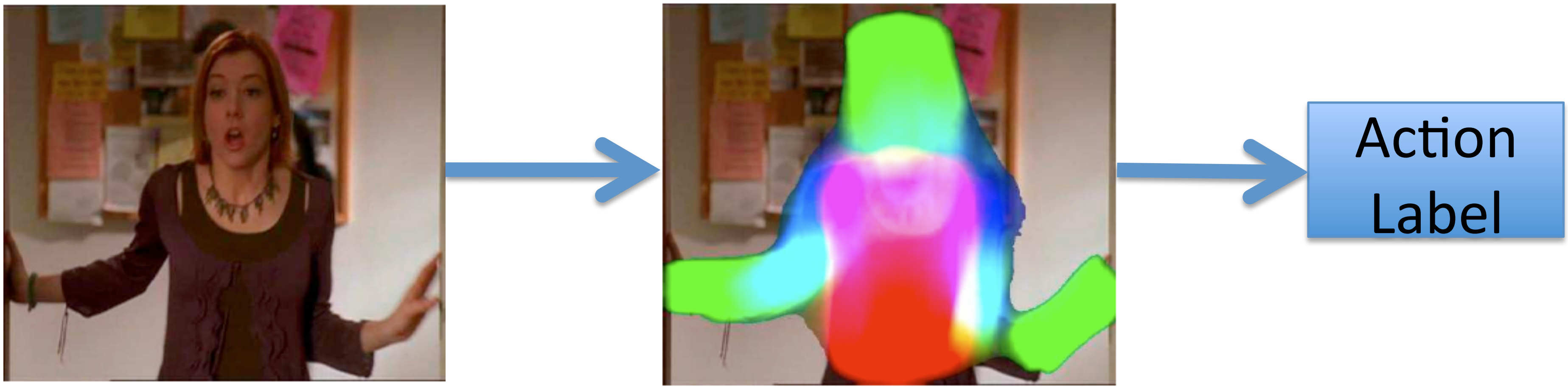
- M is quantized 3d pose
- T is root orientation

- Automatic person detection-tracking
- Compare quantized pose to labeled training poses
  - Smooth over time

Ramanan & Forsyth NIPS 2003

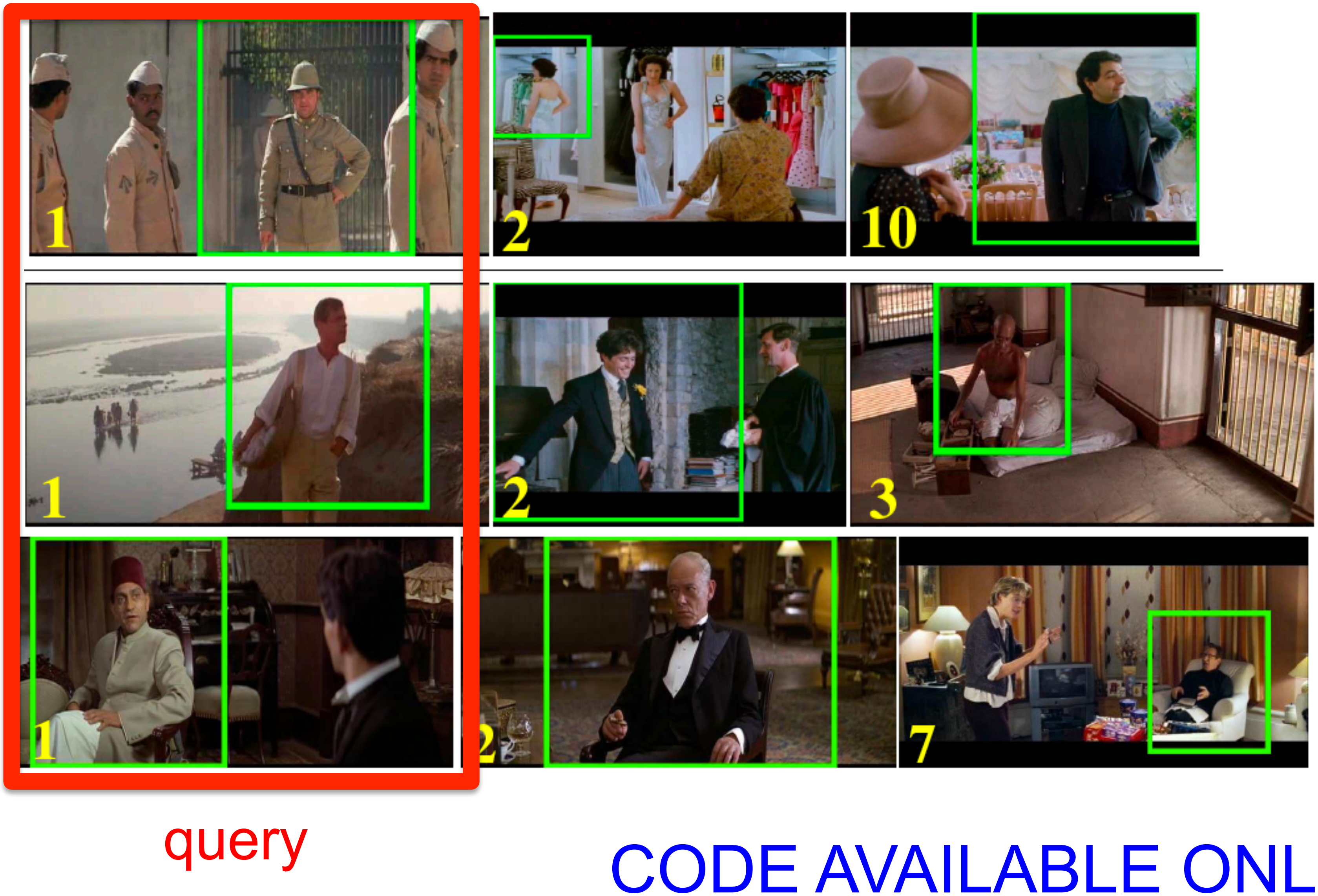
Action from Pose III: Pose Search

- Video shot retrieval from pose
  - Either *query-by-example* or classification
  - Focus on upper body pose
    - Pictorial structures model



Action Label

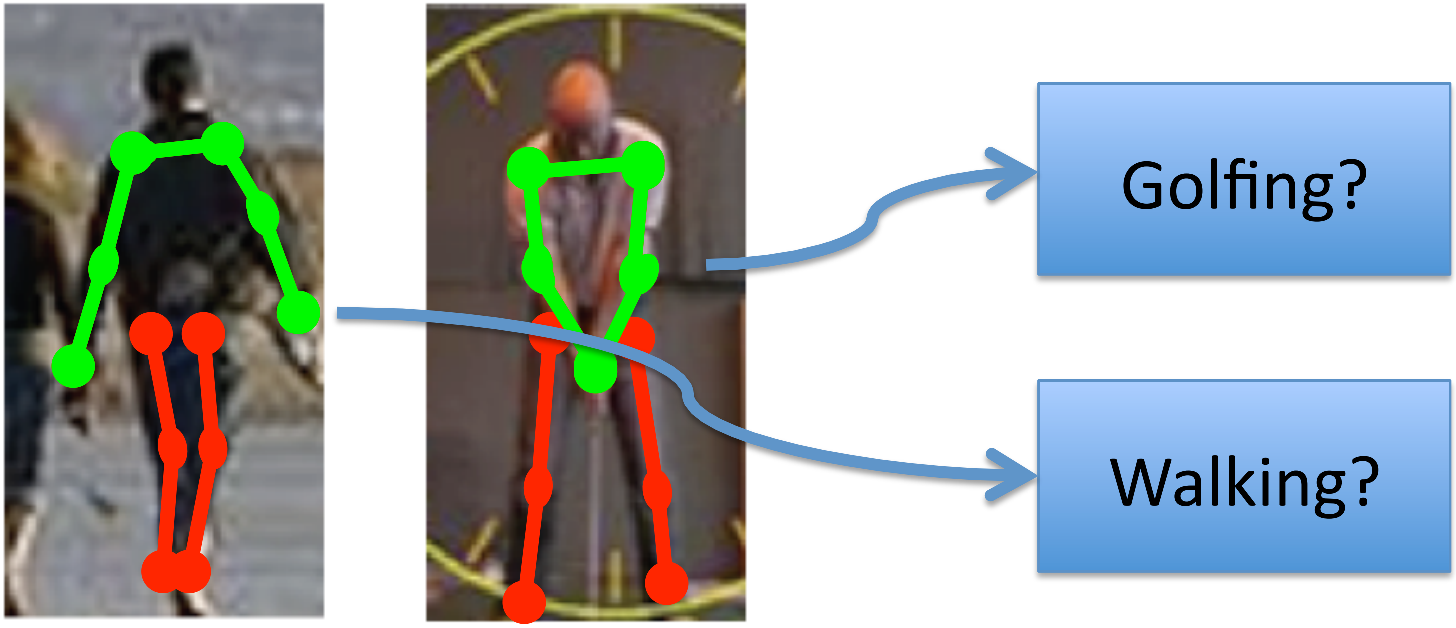
Ferrari, Marin & Zisserman CVPR 2009



CODE AVAILABLE ONLINE

- SVM on descriptors of absolute & relative part locations, segmentations
  - Include short tracks for robustness

Action from Pose IV: Discriminative Pose

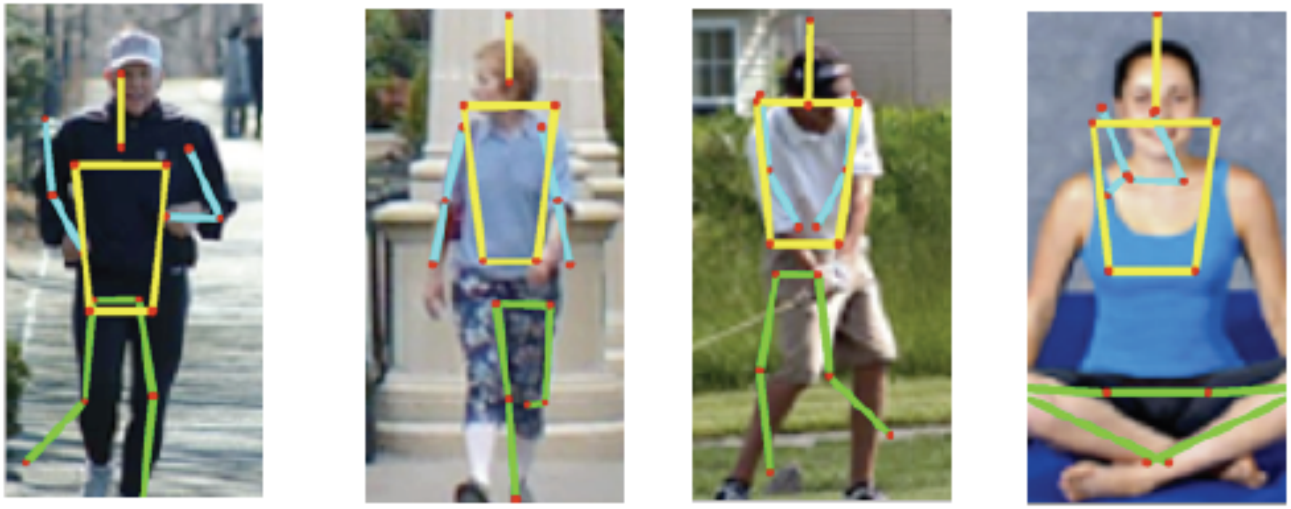


Golfing?

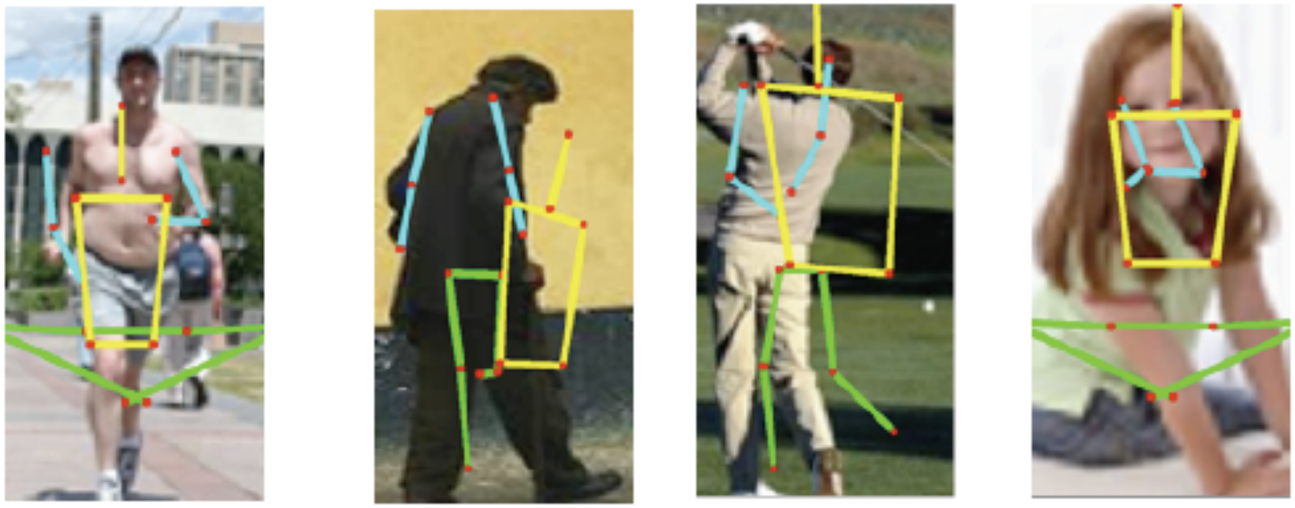
Walking?

- Focus on discriminative elements of pose for action classification
- Use exemplar-based “poselet” representation

Yang, Wang & Mori CVPR 2010



Successful classification examples



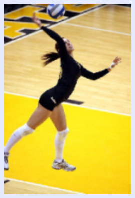
Unsuccessful classification examples

Action from Pose V: Poses and Objects

A:

Tennis forehand

Croquet shot



Volleyball smash

. . .

O:

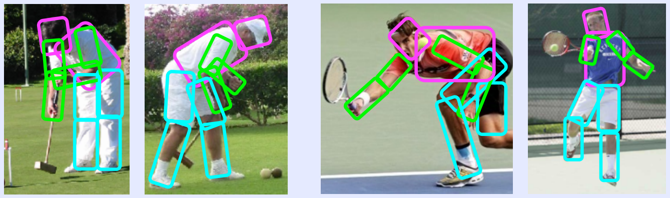
Tennis racket

Croquet mallet

Volleyball

. . .

H:

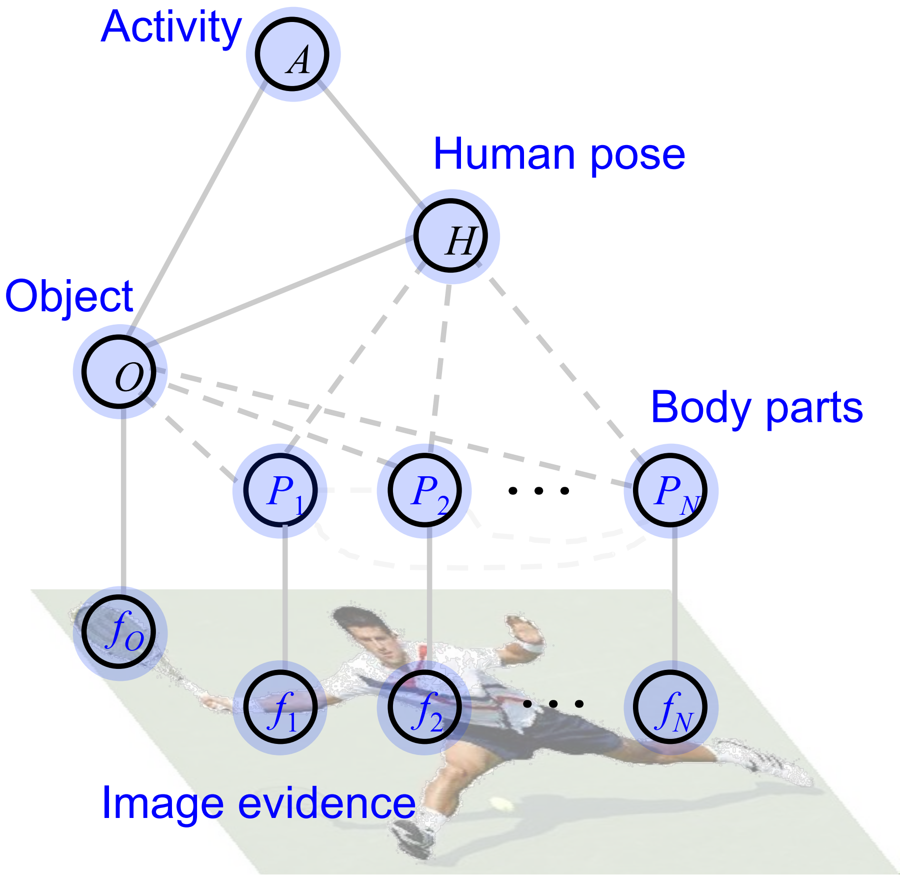


Intra-class variations

- More than one H for each A;
- **Unobserved** during training.

P: l_P: location; θ_P: orientation; s_P: scale.

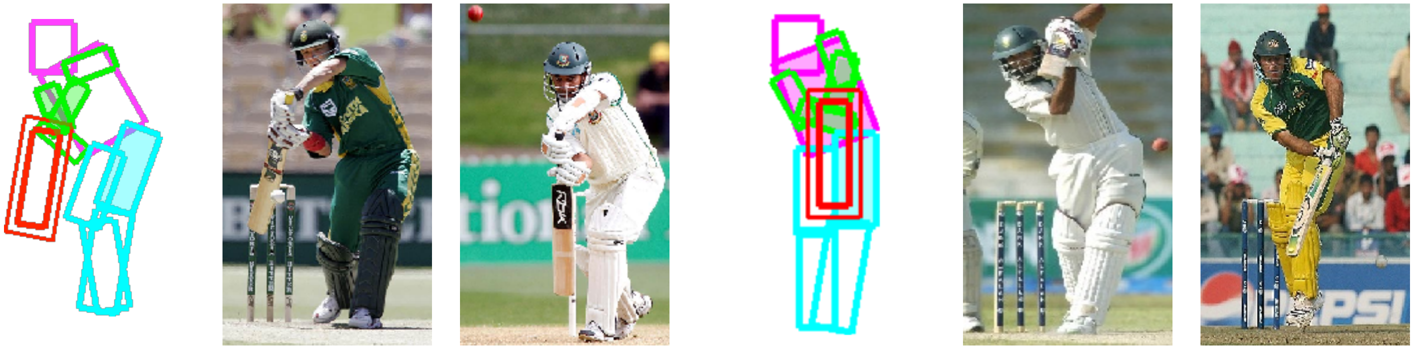
f: Shape context. [Belongie et al, 2002]



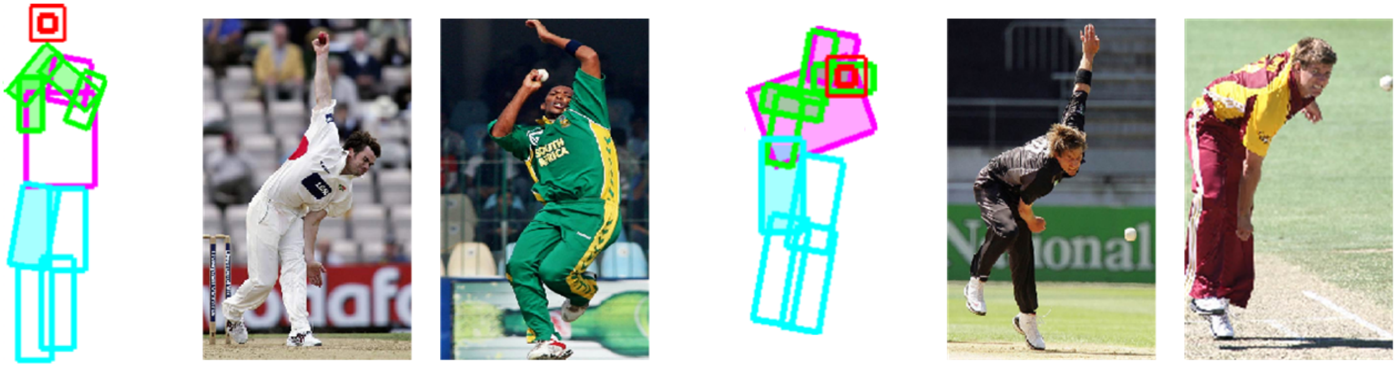
Yao & Fei-Fei CVPR 2010

Learning Results

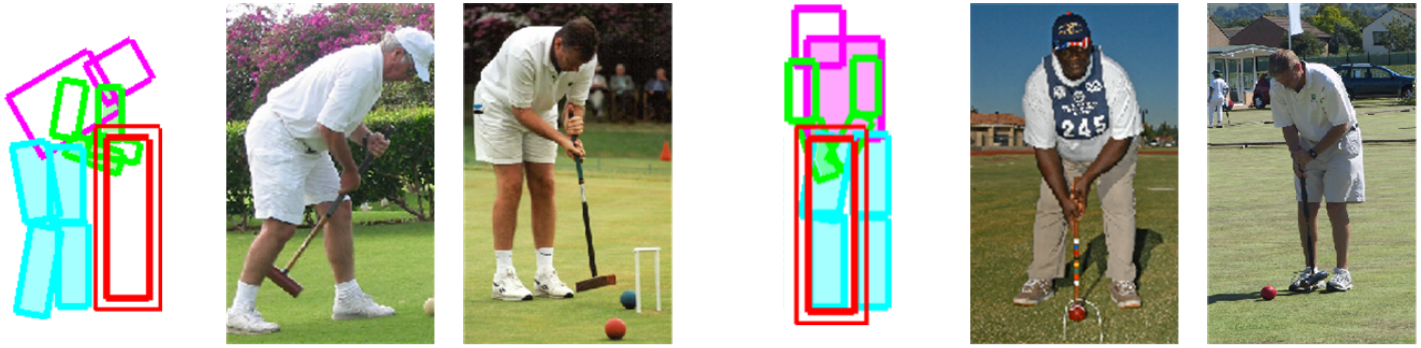
Cricket defensive shot



Cricket bowling



Croquet shot

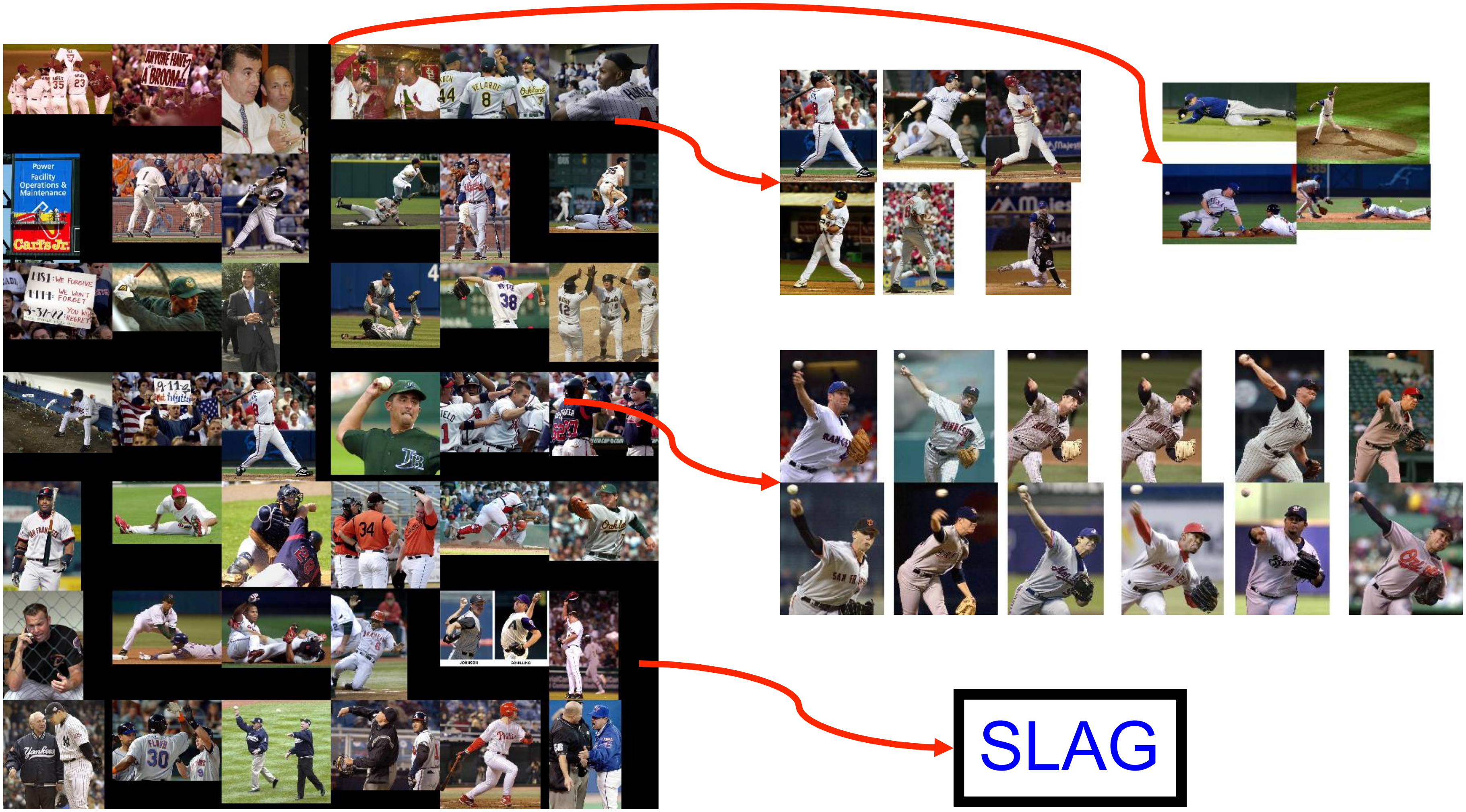


Analyzing Image Collections

- Build action models from web search results

Ikizler-Cinbis, Cinbis, Sclaroff ICCV 2009

Clustering Actions



SLAG

- Find repeated poses in a dataset

Wang, Jiang, Drew, Li, Mori CVPR 2006

Dataset: PASCAL VOC Action Classification

Riding horse

Reading book

Taking photo

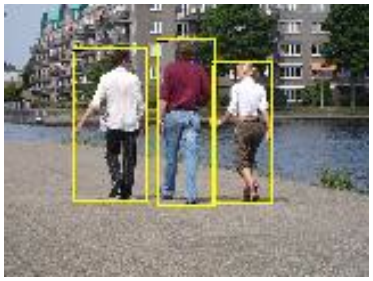
Riding bike

Play instrument

Running

Phoning

Use computer



Walking

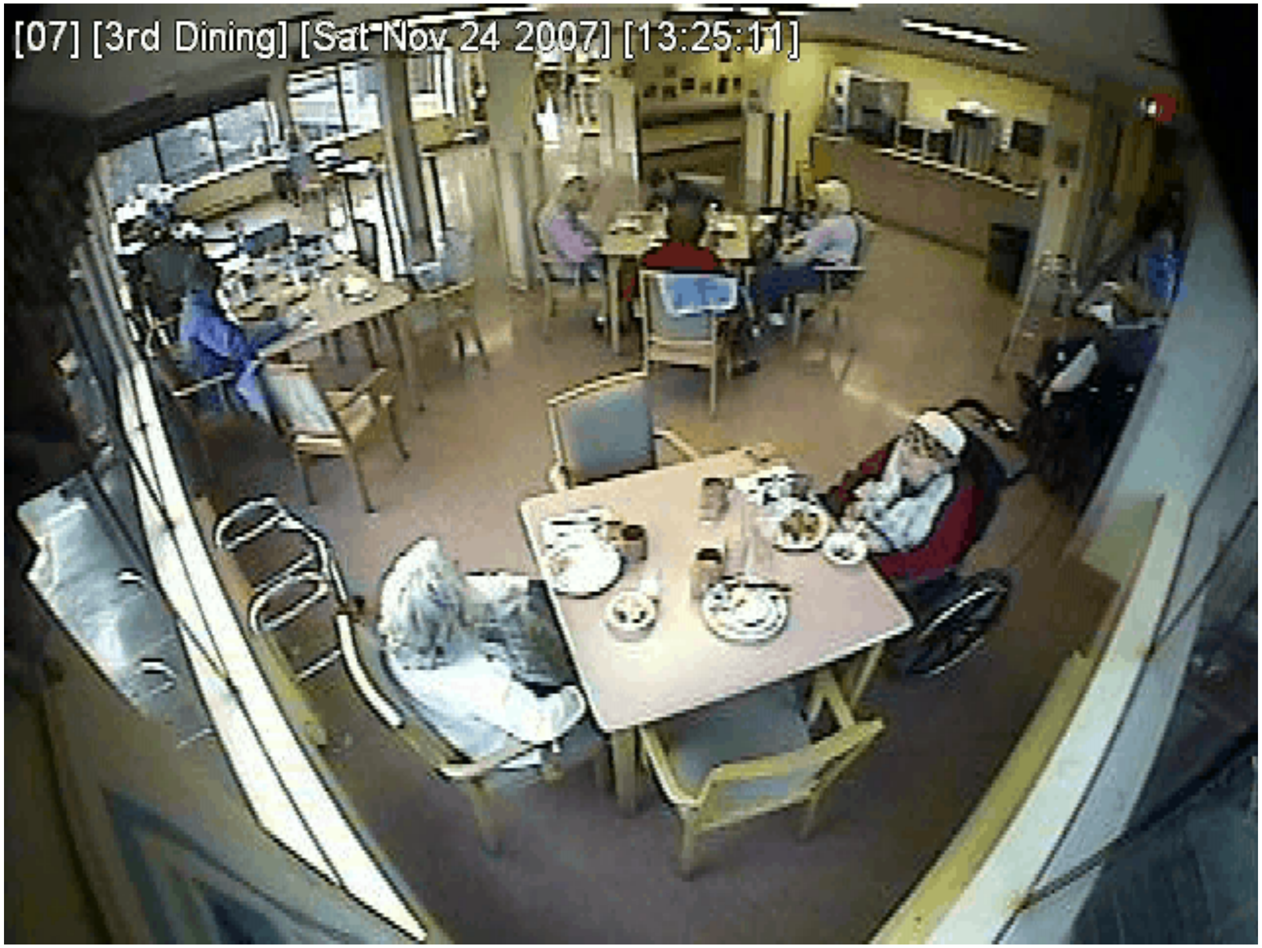
- Person location given
- Classify into one of 9 categories

Summary

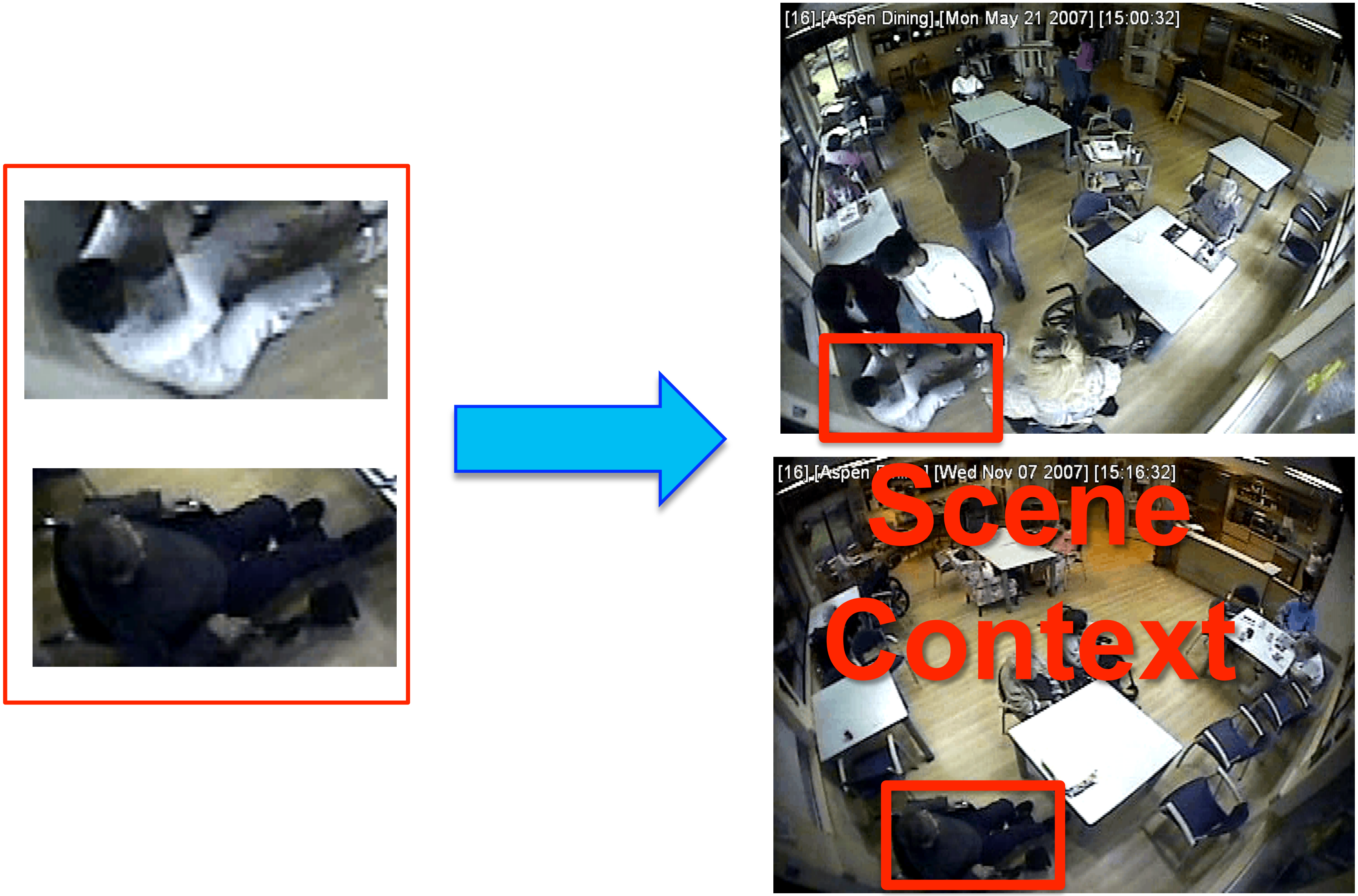
- Pose as representation for action recognition
  - Captures much information about action
  - Invariance to clothing / lighting effects
  - Model and exemplar based representations
- New direction: Action recognition from still images
  - Image retrieval and analysis
  - An important cue for video-based action recognition
  - Pose seems essential

SCENE MODELS

Getting the Whole Picture



Getting the Whole Picture


[16].[Aspen Dining].[Mon May 21 2007] [15:00:32]
[16].[Aspen
[Wed Nov 07 2007] [15:16:32]

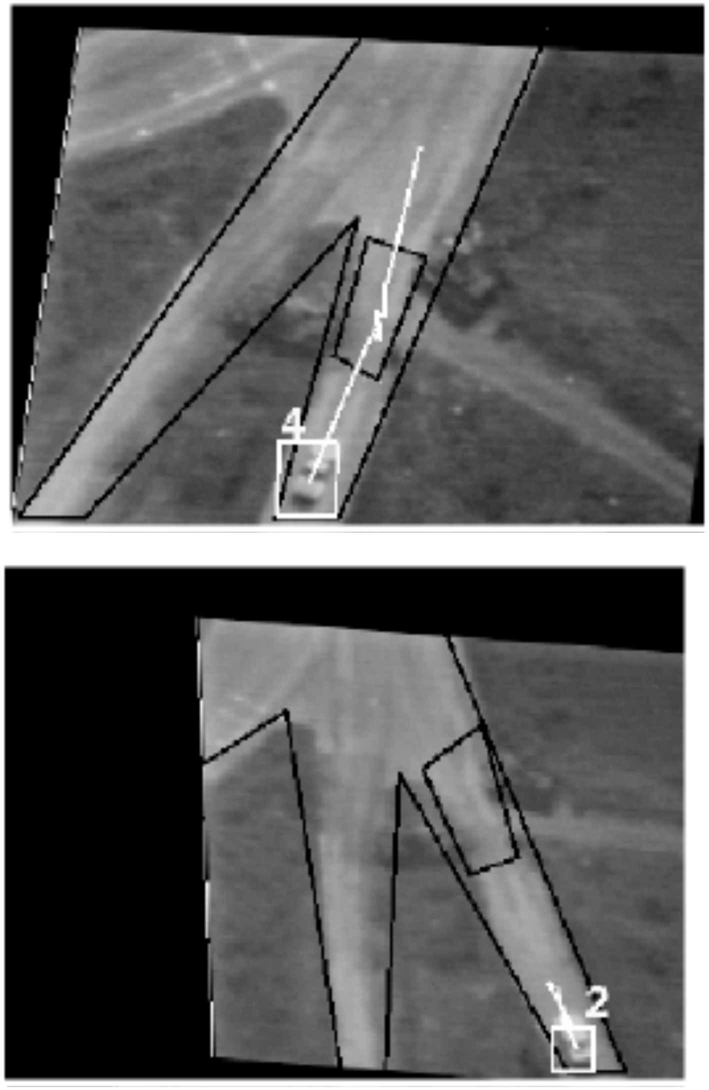
Scene Context

Scene Model Ingredients

- Describe low-level components
  - Actions of individual people
  - Movement of pixels
- Identify key objects or locations in scene
  - Buildings, roads, etc.
- Model interactions between people, objects, and locations

Scene Models I: Rule-based System

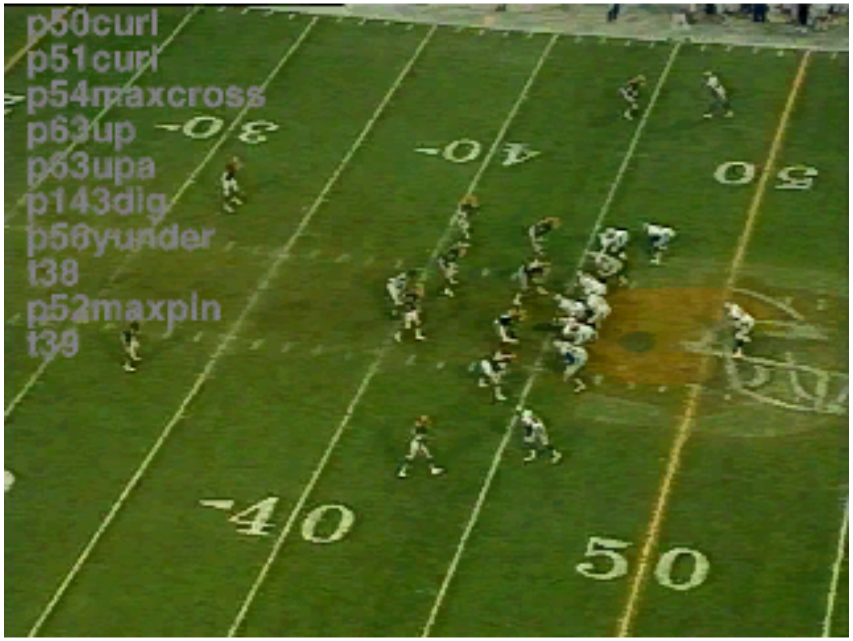
- Detect and track moving objects
- Manually identify key regions in scene
  - E.g. road, checkpoint
- **Scenarios** describe relative arrangements of objects in scene
  - E.g. proximity of car to checkpoint
  - Notions of scene **context**

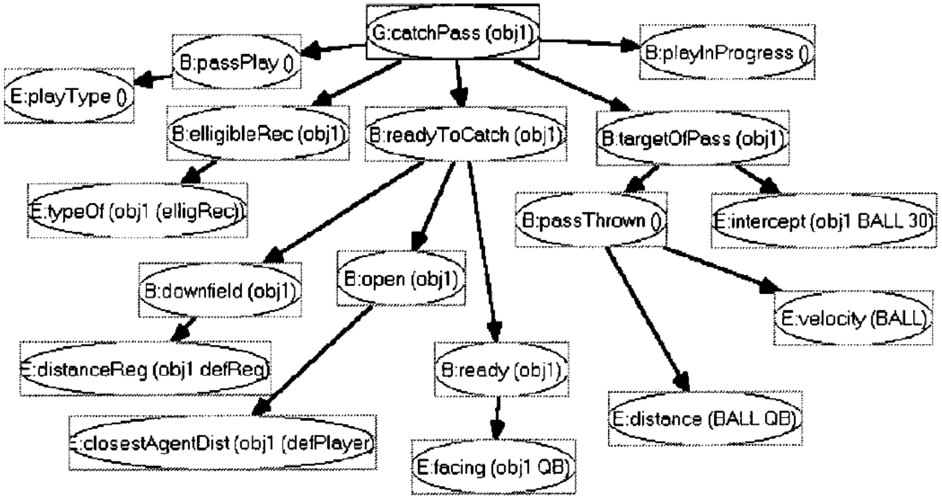


Medioni, Cohen, Bremond, Hongeng, Nevatia PAMI 2001

Scene Models II: Bayes Nets

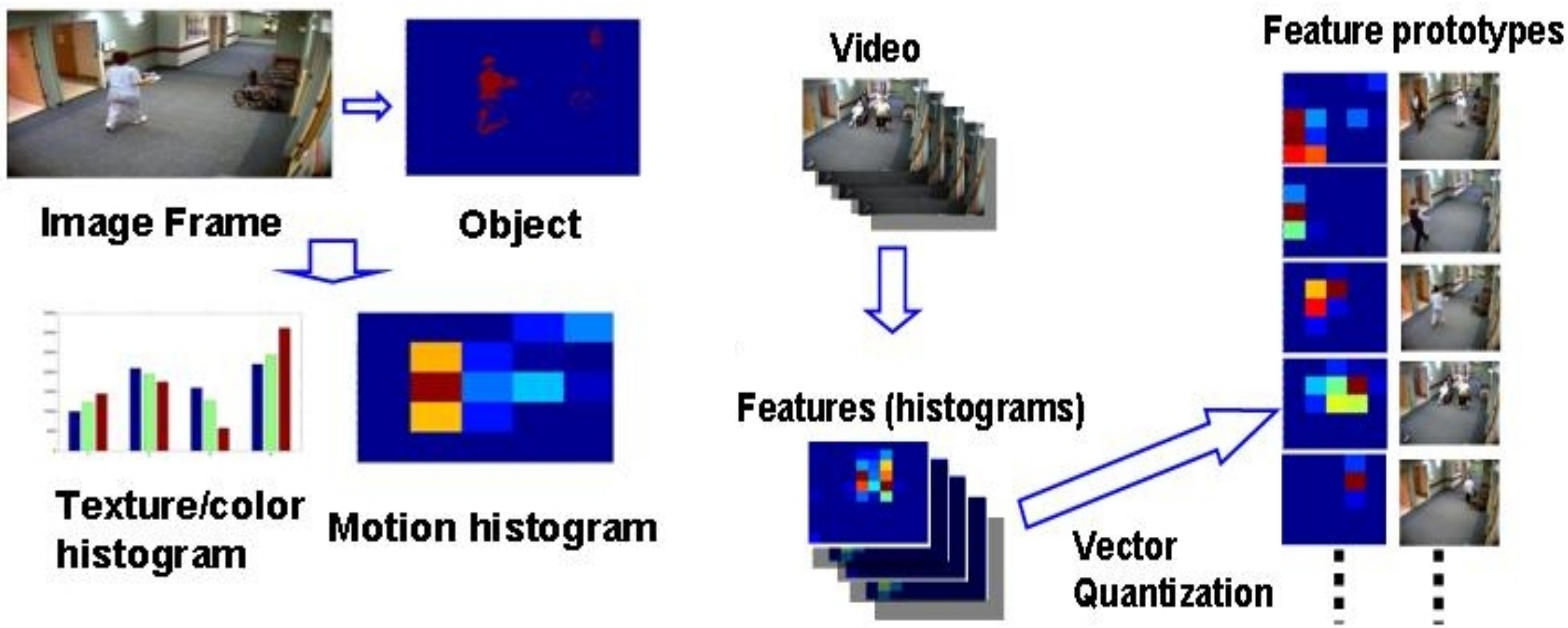
- Detect and track players, ball
- Low-level action detectors for individual players
- Hand-constructed Bayes net for each activity
  - – Spatial and temporal relations between low-level actions





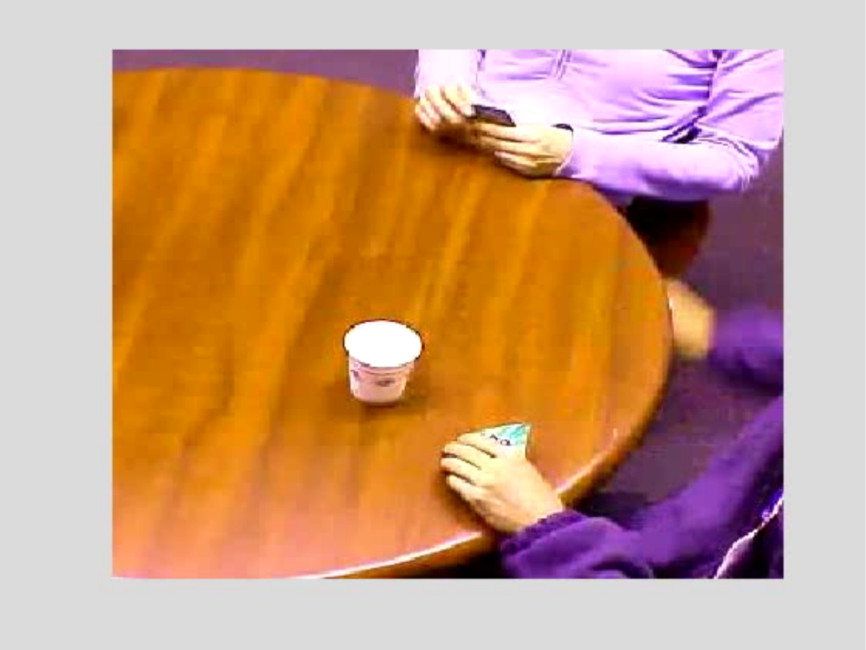
Intille & Bobick CVPR 1999

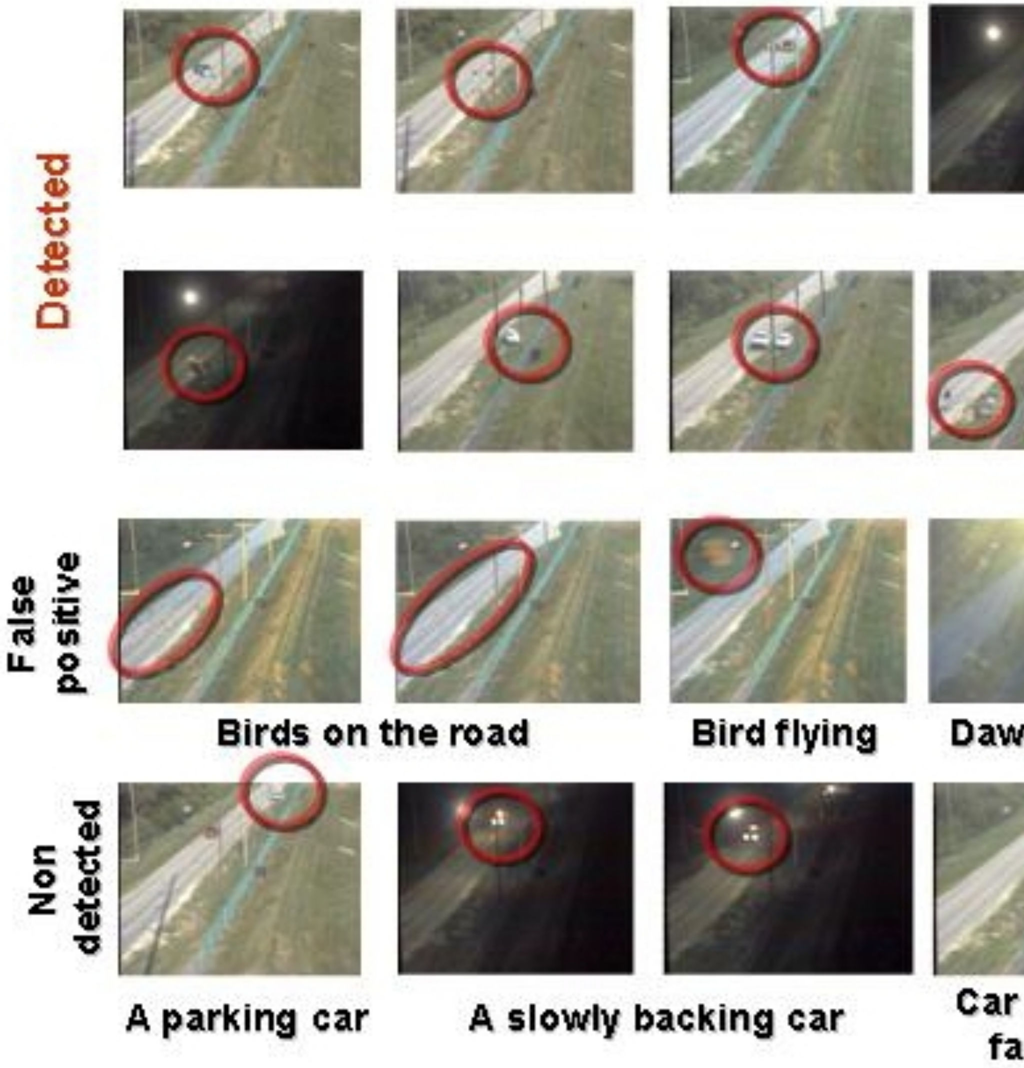
Scene Models III: Unsupervised Learning of Unusual Events



- Global, frame-level feature
  - Bag-of-words representation
- Detect unusual events by clustering
  - Isolated, varied clusters are unusual

Zhong, Shi & Visontai CVPR 2004



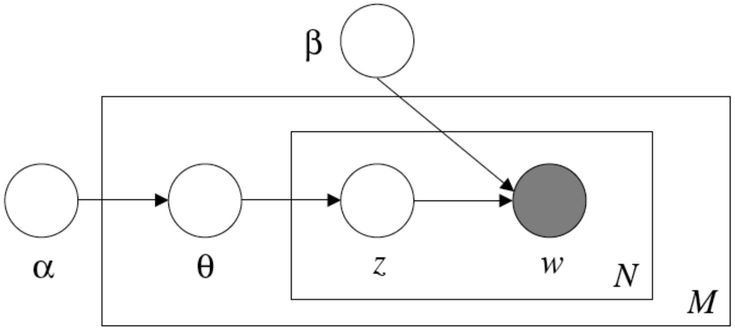


- Cheating detection in simulated card game

- Real-world highway dataset
  - Cars pulling off road, backing up, U-turns

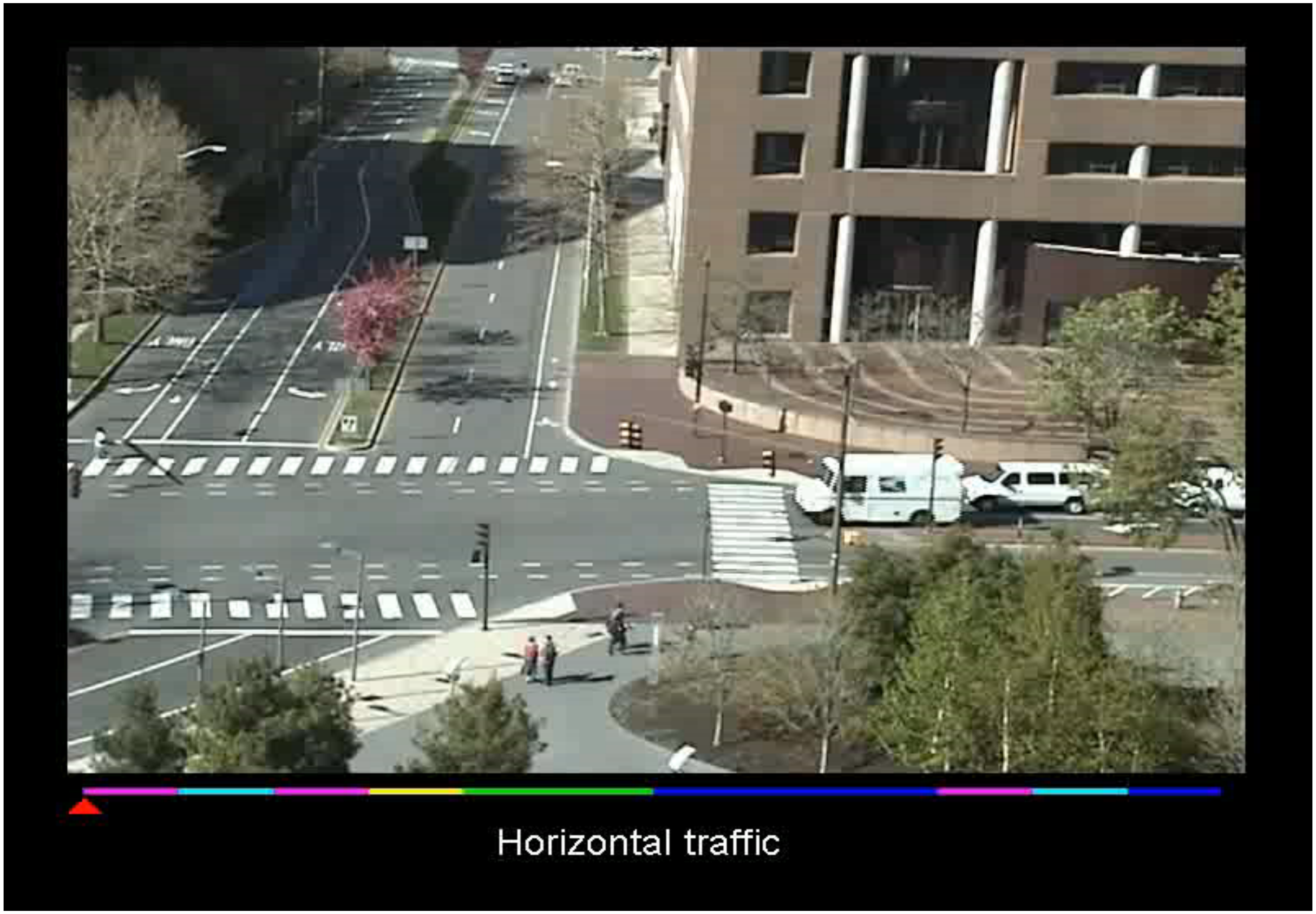
Scene Models IV: Unsupervised Hierarchical Scene Model

- Describe moving pixels by location and motion direction
  - No object detection
- Use as visual words in Latent Dirichlet Allocation (LDA) type model
  - Infer low-level actions from words



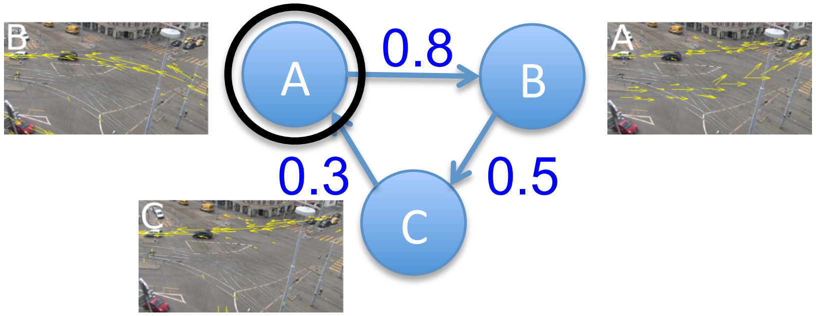
Wang, Ma, Grimson PAMI 2009

Blei, Ng, Jordan JMLR 2003

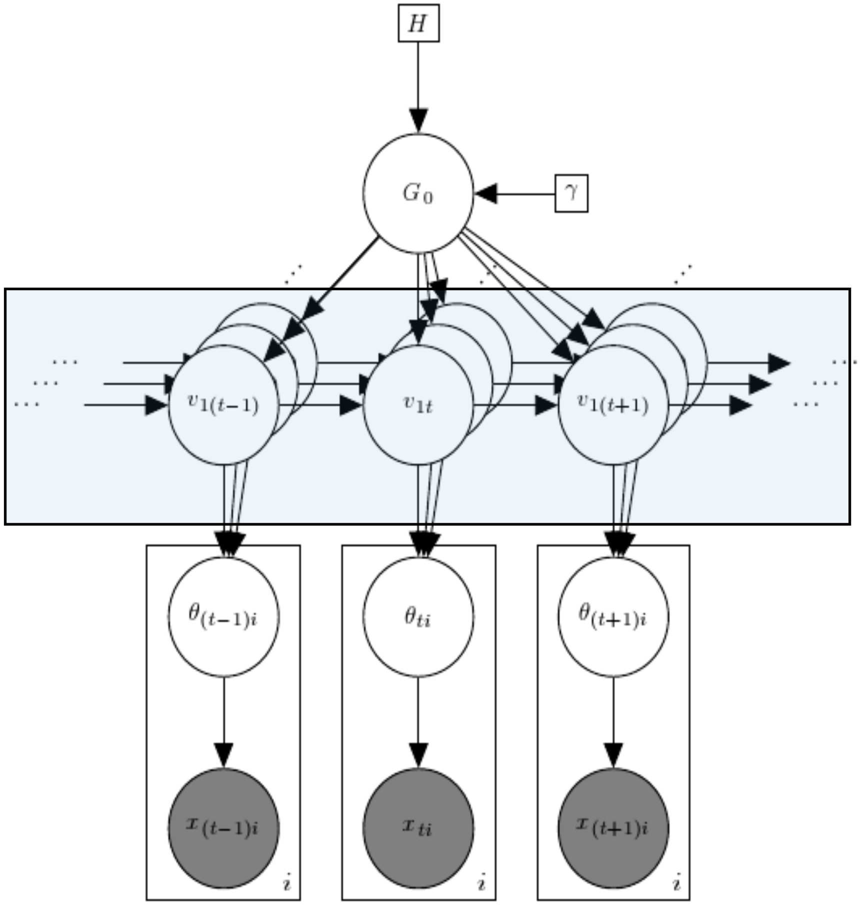


- Higher-level activity analysis
  - Distribution of low-level actions over entire scene
- Applications
  - Temporal segmentation by activity
  - Abnormality detection

Scene Models V: Hierarchical with Temporal Dependencies



- Hierarchical Dirichlet Process model
  - Learn number of activities automatically

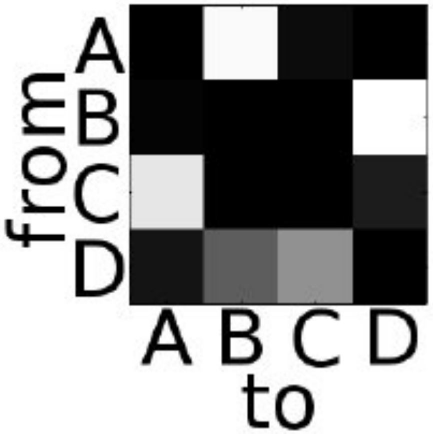


Kuettel, Breitenstein, van Gool & Ferrari CVPR 2010

traffic light controlled scene

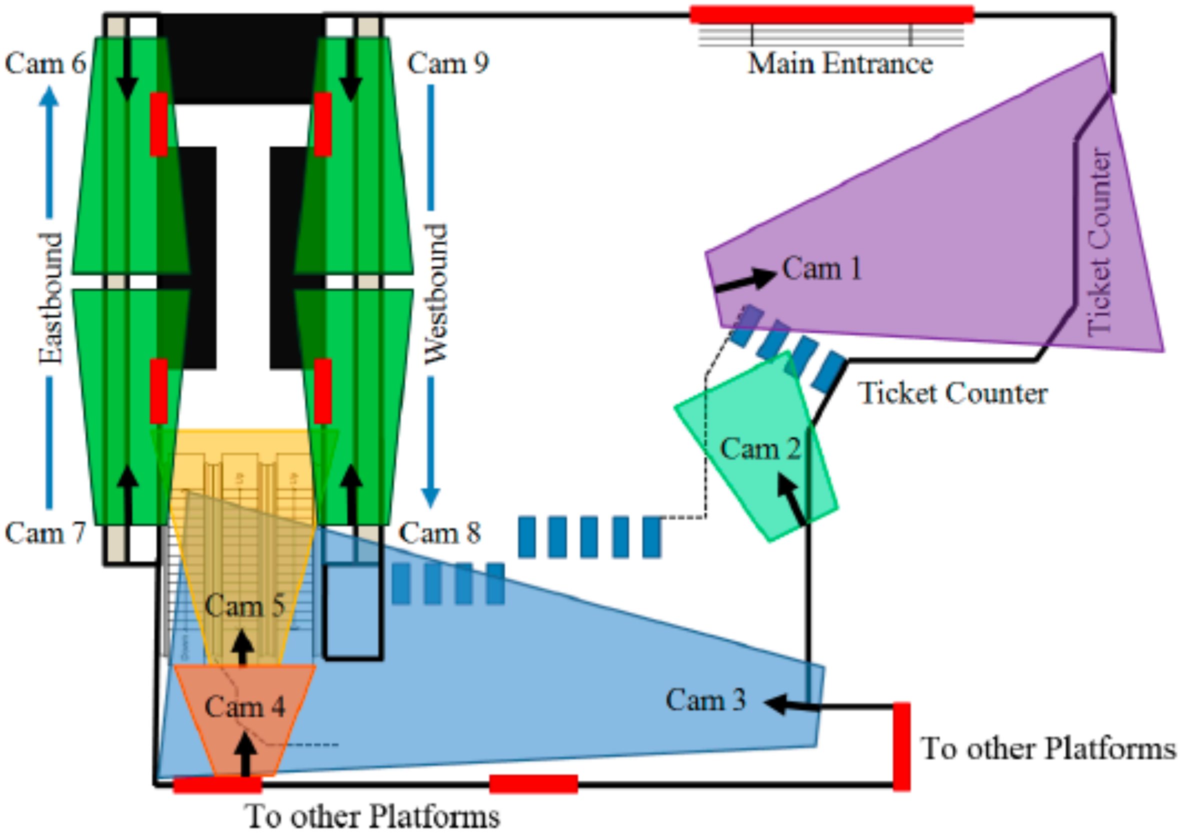
current state A (history A)

current state A (history A)



- continuous video
- annotated with states and history
- 3x speed

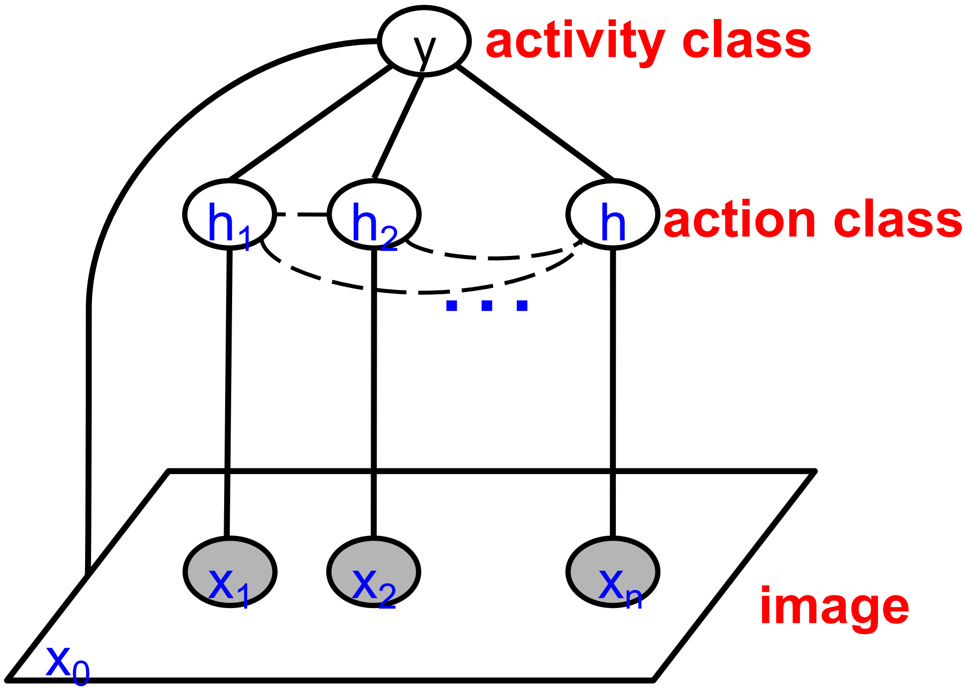
Scene Models VI: Multi-Camera Scene Decomposition

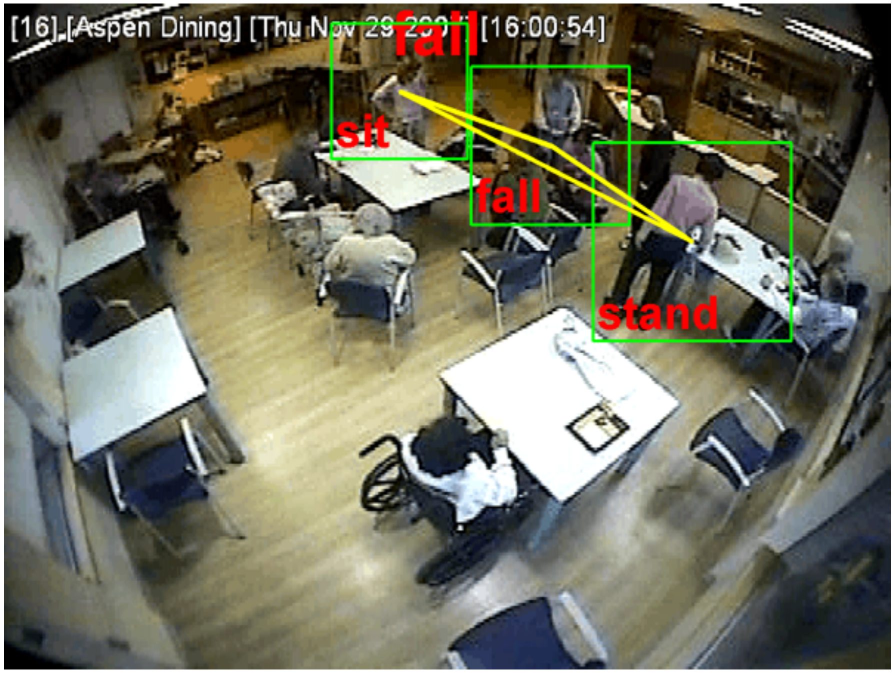


Loy, Xiang & Gong CVPR, ICCV 2009

- Consider time-delayed correlations between regions
  - Applications to irregularity detection

Scene Models VII: Person-Person Context

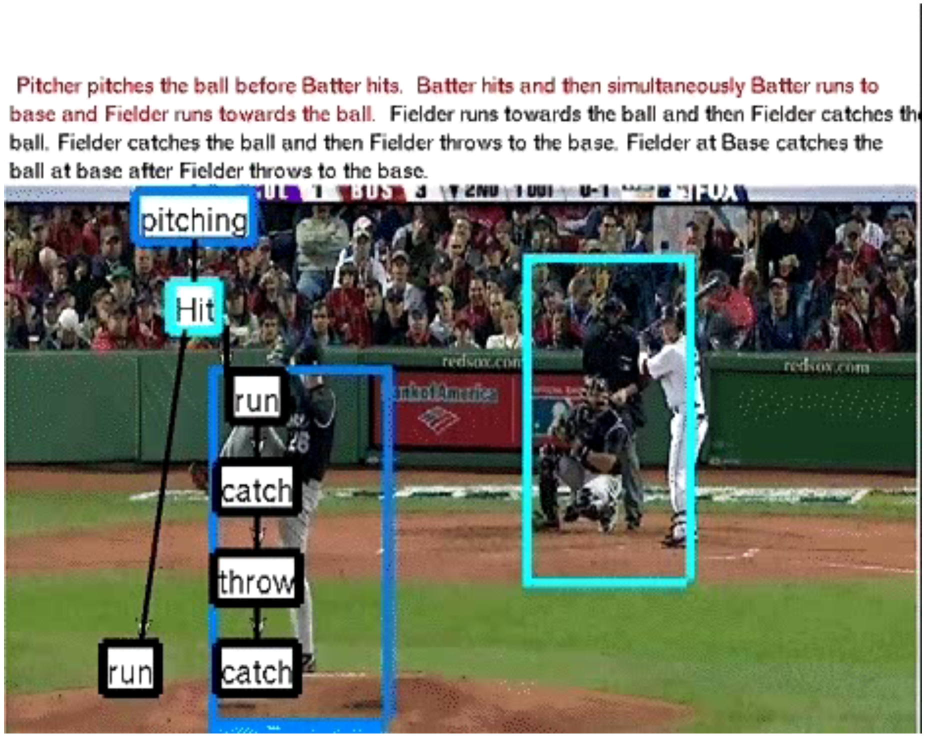




Choi, Shahid, & Savarese VS 2009
Lan, Wang, Yang, & Mori SGA 2010, NIPS 2010

Scene Models VIII: Storyline Model

- Captioned baseball videos in training
- Build AND-OR graph representation of activities
  - AND specifies elements of an activity that must occur
  - OR allows variation in how an element appears
- Describe low-level tracks using STIPs
- Match tracks to actions in AND-OR graph



Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder runs towards the ball and then Fielder catches th ball. Fielder catches the ball and then Fielder throws to the base. Fielder at Base catches the ball at base after Fielder throws to the base.

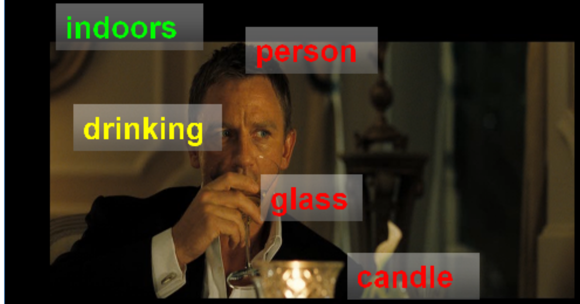
Gupta, Srinivasan, Shi, Davis CVPR 2009

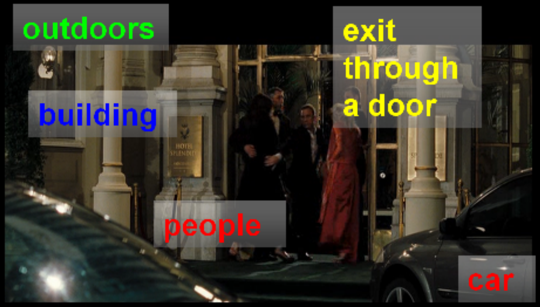
Summary

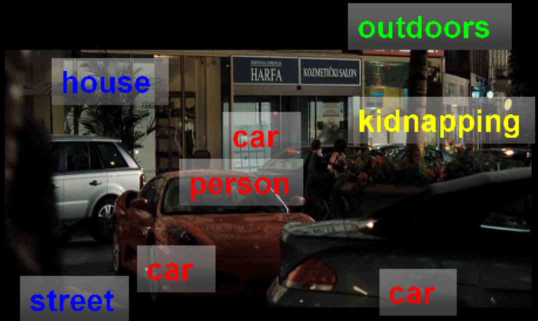
- Scene modeling to look at the big picture
- Feature representations
  - Holistic: describe entire scene, irrespective of individuals
  - Local: describe actions of individuals
- Structure of activities
  - Model free: clustering-type approaches
  - Strong models: grammars, probabilistic models

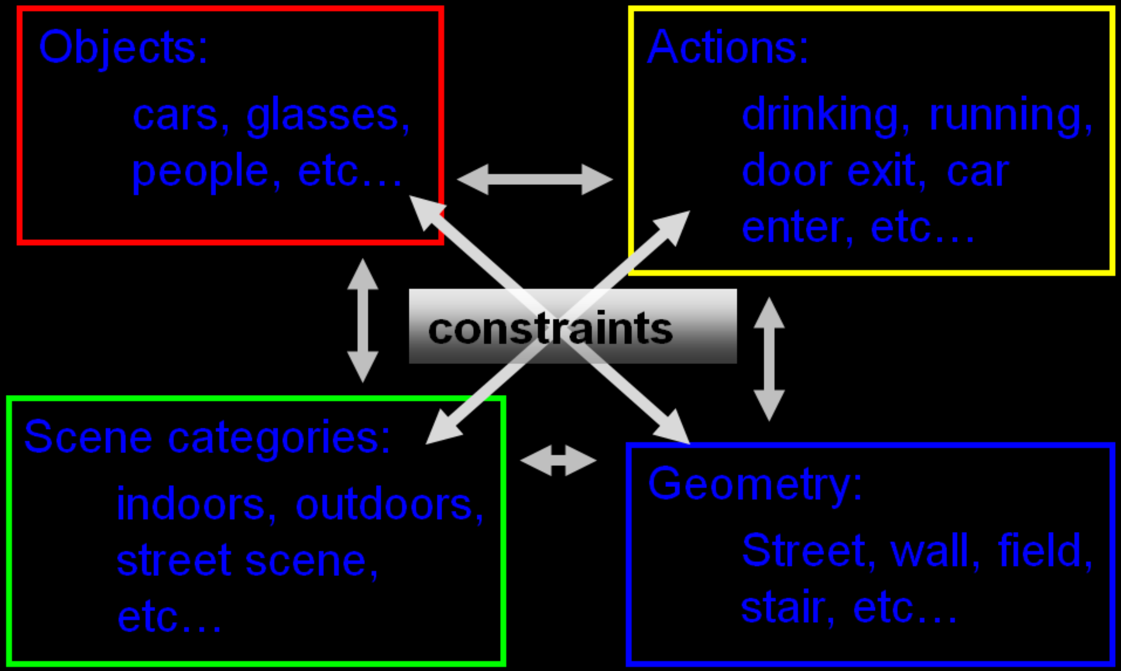
CONCLUSIONS

Computer vision grand challenge: Video understanding









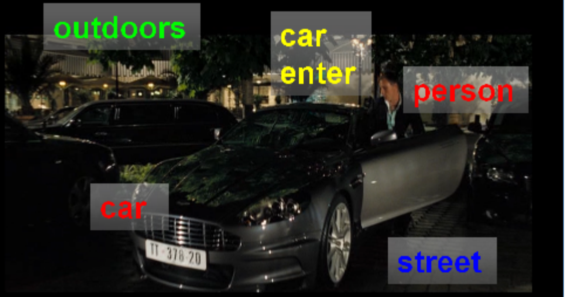
Objects: cars, glasses, people, etc…

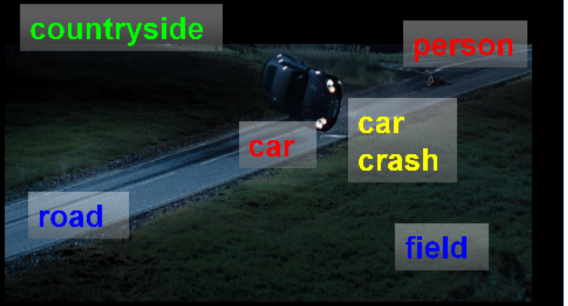
Actions: drinking, running, door exit, car enter, etc…

constraints

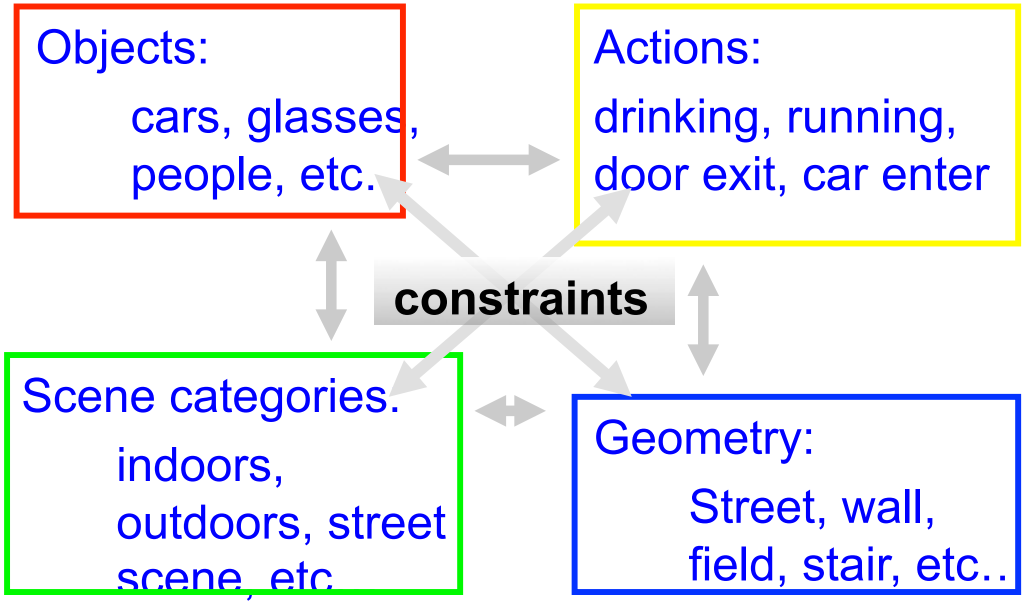
Scene categories: indoors, outdoors, street scene, etc…

Geometry: Street, wall, field, stair, etc…





Original Aim



Objects:
cars, glasses, people, etc.

Actions:
drinking, running, door exit, car enter

constraints

Scene categories:
indoors, outdoors, street scene, etc

Geometry:
Street, wall, field, stair, etc..

- 🟨 **Early silhouette and tracking-based methods**
- 🟨 **Motion-based similarity measures**
- 🟨 **Template-based methods**
- 🟨🟥🟩 **Local space-time features**
- 🟨🟥🟩 **Bag-of-Features action recognition**
- 🟨🟥🟩🟦 **Weakly-supervised methods**
- 🟨🟥 **Pose estimation and action recognition**
- 🟨🟥 **Action recognition in still images**
- 🟨🟥🟩🟦 **Human interactions and dynamic scene models**

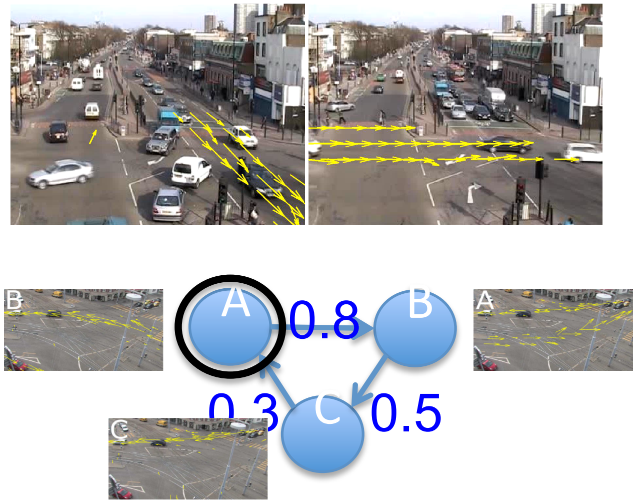
Future Directions I: Problem Definitions

Riding horse

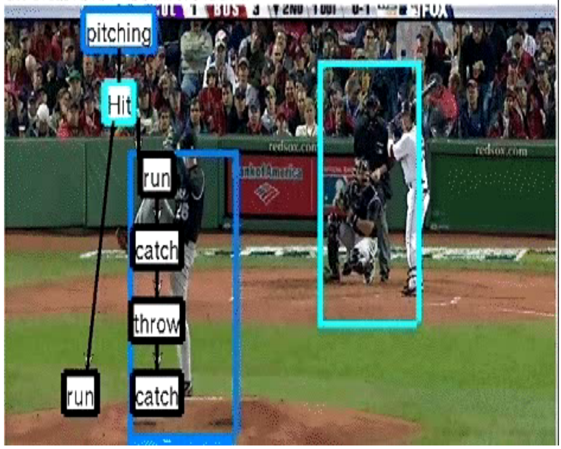
Reading book

Riding bike

Play instrument



Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder runs towards the ball and then Fielder catches th ball. Fielder catches the ball and then Fielder throws to the base. Fielder at Base catches the ball at base after Fielder throws to the base.



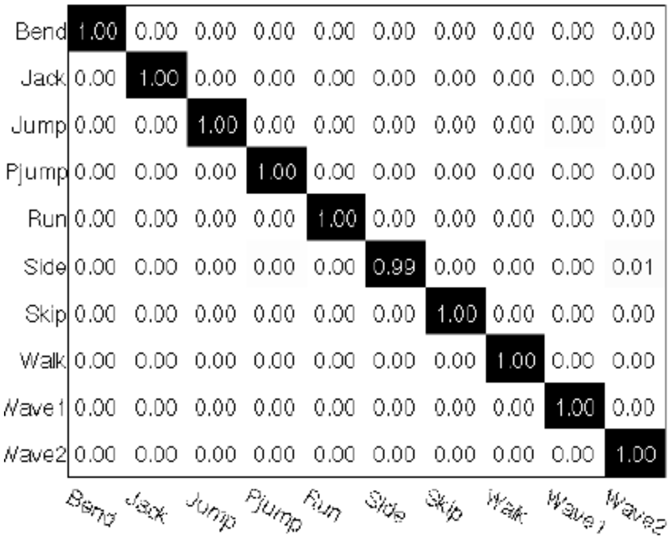
Datasets & Baselines

- Standardization of datasets for field
  - Allow comparison of algorithms
    - E.g. KTH for low-level features, atomic actions
  - Fair tuning of model parameters
- New algorithms compare to baselines
  - Bag-of-words on densely sampled STIPs
  - Pose estimation (Ferrari et al. code)
  - HOG SVM (Dalal & Triggs code, Ramanan code)

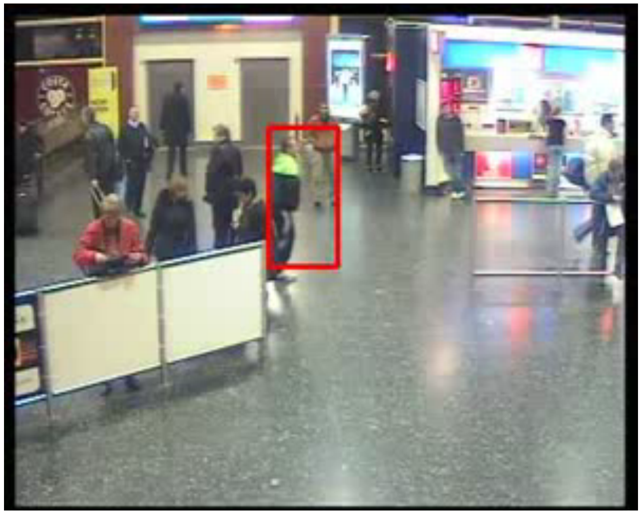
Datasets & Baselines

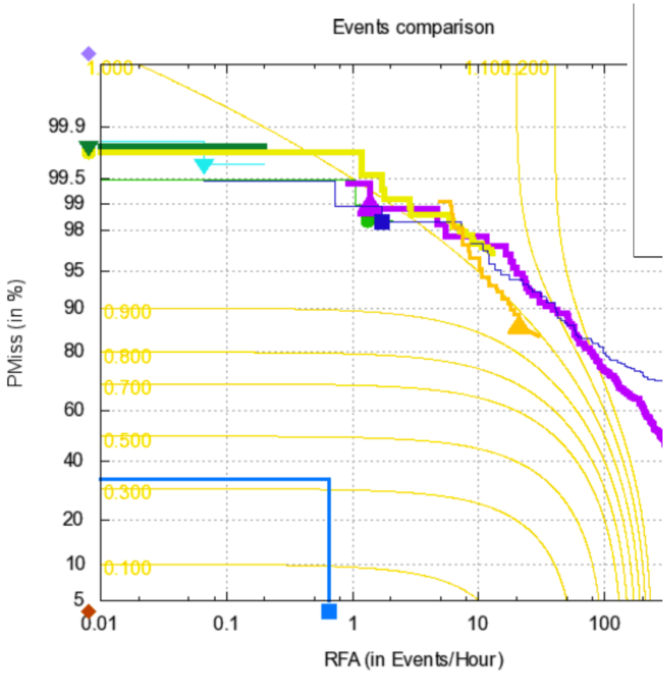
- Standardization of datasets for field
  - Don't feel constrained by the existing problem definitions
  - Do make your new dataset available
    - Should clearly specify separate training and test sets
- New algorithms compare to baselines
  - Do use reasonable variant of standard baselines for your new problem

Future Directions II: Back to Basics



	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pjump	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Side	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.01
Skip	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00





Future Directions II: Back to Basics

- Even atomic low-level actions are very difficult to detect reliably
  - Far more work needed on representations for the action of a single person
  - Features
  - Temporal representation, smoothing
  - Tracking
  - …

Future Directions III: Obtaining Data

1. Cameras and bandwidth are cheap
2. Lots of training data is potentially available

+

=

Training data

⇨ Potential for the huge progress

… if we can get the data

Readily available video annotation

	Aligned with video	Describes visual content	Source
Subtitles	Yes	No	DVD, Internet
Scripts for TV series, movies and sport games	No	Yes	Internet, e.g. www.dailyscript.com
Plot summaries and synopses	No	Yes, sparsely	Internet (e.g. IMDB)
Instruction videos	No	Yes	Internet, e.g. www.videojug.com
Descriptive Video Service	Yes	Yes	DVD, rare
Word tags	No	Yes, sparsely	Internet (e.g. YouTube)
Manual labelling, Human Computation	??	??	Mechanical Turk, ESP Game, ~~Grad~~ undergrad students

Open questions:

- How to benefit from the structure of the human body in complex situations, e.g. heavy occlusions, uniformly colored clothing?
- Will action classification generalize over different video domains: Movies, TV, YouTube, Surveillance video?
- What is the *useful* action vocabulary? Are we trying to solve the right problem? How can we visualize/display the results?

Interesting novel directions:

- Use actions for recognizing functional and physical object properties, e.g. “sitable”, “eatable”, “heavy”, “solid” objects…
- Action prediction, i.e. what can happen in the given situation: e.g. is it dangerous to cross this road?
- Explore more sources of strong and weak supervision: Manual surveillance, Descriptive Video Service (DVS), YouTube tags; Transcripts of sports games; Instruction videos.

References - Preliminaries

- P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In Proc. 9th Int. Conf. Computer Vision, pages 734–741, 2003.
- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 2005.
- Bo Wu and Ram Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In Proc. 10th Int. Conf. Computer Vision, 2005.
- Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.
- Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real- time tracking. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 1999.
- Kentaro Toyama, John Krumm, Barry Brumitt, and Brian Meyers. Wallflower: Principles and practice of background maintenance. In Proc. 7th Int. Conf. Computer Vision, 1999.
- J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. Int. Journal of Computer Vision, 12(1):43–77, 1994.
- T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 2009.
- M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. Int. Journal of Computer Vision, 29(1):5–28, 1998.
- Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 2009.

References – Motion/shape Templates

- W. T. Freeman, K. Tanaka, J.Ohta, and K. Kyuma. Computer vision for computer games. In IEEE 2nd Intl. Conf. on Automatic Face and Gesture Recognition, 1996.
- J. Sullivan and S. Carlsson. Recognizing and tracking human action. In ECCV 2002
- A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In ICCV 2003
- A. Bobick and J. Davis. The recognition of human movement using temporal templates. IEEE Trans. PAMI, 23(3):257–267, 2001.
- L. Zelnik-Manor and M. Irani. Event-based video analysis. In CVPR 2001
- E. Shechtman and M. Irani. Space-time behavior based correlation. In CVPR 2005
- O. Boiman and M. Irani. Detecting irregularities in images and in video. In Proc. ICCV, 2005.
- M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In Proc. ICCV, 2005.
- Y. Ke, R. Sukthankar, and M. Hebert. Efficient Visual Event Detection using Volumetric Features . In Proc. ICCV 2005.
- Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In Proc. ICCV, 2007.
- I. Laptev and P. Pérez. Retrieving actions in movies. In Proc. ICCV 2007
- D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In Proc. CVPR, 2008.
- Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In Proc. ICCV, 2009.

References – Local Features

- I. Laptev and T. Lindeberg. Space-time interest points. In Proc. ICCV 2003.
- C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In Proc. ICPR, 2004.
- P. Dollar, V. Rabaud, G. Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In VS-PETS, 2005.
- H. Jhuang, T. Serre, L. Wolf and T. Poggio. A Biologically Inspired System for Action Recognition. In Proc. ICCV 2007
- P. Scovanner, S. Ali, and M. Shah, A 3-Dimensional SIFT descriptor and its application to action recognition, ACM MM 2007.
- J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In IJCV 2008.
- I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In Proc. CVPR 2008.
- A. Klaeser, M. Marszałek and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In Proc. BMVC 2008
- G. Willems, T. Tuytelaars and L. Van Gool. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In Proc. ECCV 2008
- H. Wang, M. M. Ullah, A. Kläser, I. Laptev and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In Proc. BMVC 2009.
- L. Yeffet and L. Wolf. Local Trinary Patterns for Human Action Recognition. In Proc. ICCV 2009.
- A. Gilbert, J. Illingworth, R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features, In Proc. ICCV 2009.
- P. Matikainen, M. Hebert, R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. ICCV workshop on Video-oriented Object and Event Classification, 2009
- M. M. Ullah, S. N. Parizi, I. Laptev. Improving bag-of-features action recognition with non-local cues. In Proc. BMVC 2010

References – Human Pose and Actions

- Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. IEEE Trans. PAMI, 25 (7):814–827, 2003.
- D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In Advances in Neural Information Processing Systems 16, 2003.
- V. Ferrari, M. Marin, and A. Zisserman. Pose search: retrieving people using their pose. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 2009.
- Yang Wang, Hao Jiang, Mark S. Drew, Ze-Nian Li, and Greg Mori. Unsupervised discovery of action classes. In CVPR, 2006.
- Nazli Ikizler-Cinbis, R. Gokberk Cinbis, and Stan Sclaroff. Learning actions from the web. In IEEE International Conference on Computer Vision, 2009.
- Weilong Yang, Yang Wang, and Greg Mori. Recognizing human actions from still images with latent poses. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 2010.
- Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 2010.

References – Periodic motion

- R. Polana and R.C. Nelson. Detection and recognition of periodic, nonrigid motion. In IJCV 1997.
- S.M. Seitz and C.R. Dyer. View invariant analysis of cyclic motion. In IJCV 1997
- A. Thangali and S. Sclaroff. Periodic motion detection and estimation via space-time sampling. In IEEE Workshop on Motion and Video Computing, 2005.
- I. Laptev, S.J. Belongie, P. Pérez and J. Wills. Periodic motion detection and segmentation via approximate sequence alignment, In Proc. ICCV 2005
- P. Wang, G.D. Abowd and J.M. Rehg. Quasi-periodic event analysis for social game retrieval. In Proc ICCV 2009

References – View invariance

- D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. in Proc. ICCV 2007.
- A. Farhadi and M. Tabrizi. Learning to recognize activities from the wrong view point. In Proc. ECCV 2008.
- I. Junejo, E. Dexter, I. Laptev and Patrick Pérez. Cross-view action recognition from temporal self-similarities. In Proc. ECCV 2008
- A. Farhadi, M. Kamali, I. Endres, D. Forsyth. A latent model of discriminative aspect. In Proc. ICCV 2009.

References – Scene Models

- X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans. PAMI, 31(3):539– 555, 2009.
- Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. Understanding videos, constructing plots - learning a visually grounded storyline model from annotated videos. In CVPR, 2009.
- T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. Int. Journal of Computer Vision, 67(1):21–51, 2006.
- G. Medioni, I. Cohen, F. Bre´mond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. IEEE Trans. PAMI, 23(8):873–889, 2001.
- Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. PAMI, 22(8):852–872, 2000.
- D. Moore and I. Essa. Recognizing multitasked activities using stochastic context-free grammar using video. In AAAI, 2002.
- Chen Change Loy, Tao Xiang, and Shaogang Gong. Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In ICCV, 2009.
- Xiaogang Wang, Keng Teck Ma, Gee Wah Ng, and W. Eric L. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., 2008.
- W. Choi, K. Shahid, and S. Savarese. ”what are they doing? : Collective activity classification using spatio-temporal relationship among people”. In 9th International Workshop on Visual Surveillance, 2009.
- Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In CVPR, 2009.

Thank you!

Workshop materials available:

https://sites.google.com/site/humanactionstutorialeccv10/