Learning Structured Models for Recognizing Human Actions

Greg Mori
School of Computing Science
Simon Fraser University

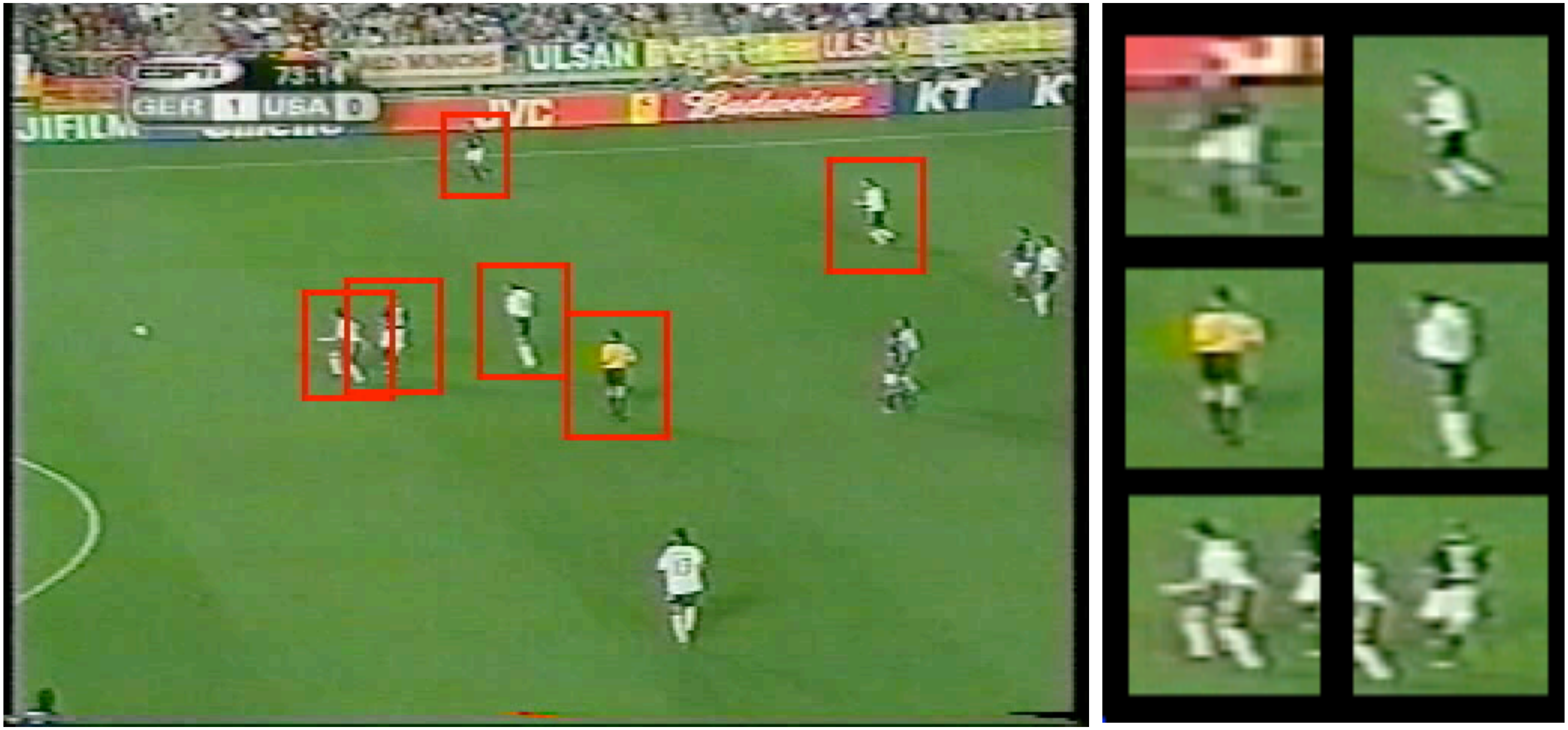
Seventh Canadian Conference on Computer and Robot Vision
June 2, 2010

SFU Vision and Media Lab

Action Recognition

- Recognize human actions from raw video data

Gathering action data

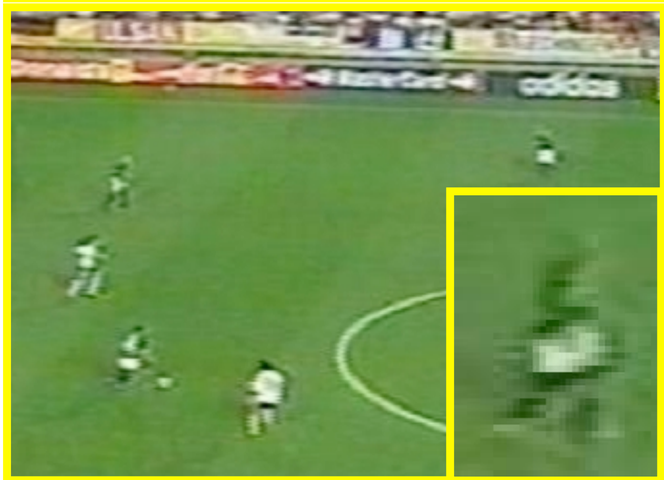


- 3 components:
  - detect humans, track, recognize action

SFU Vision and Media Lab

Far field

- 3-pixel man
- Blob tracking



Medium field

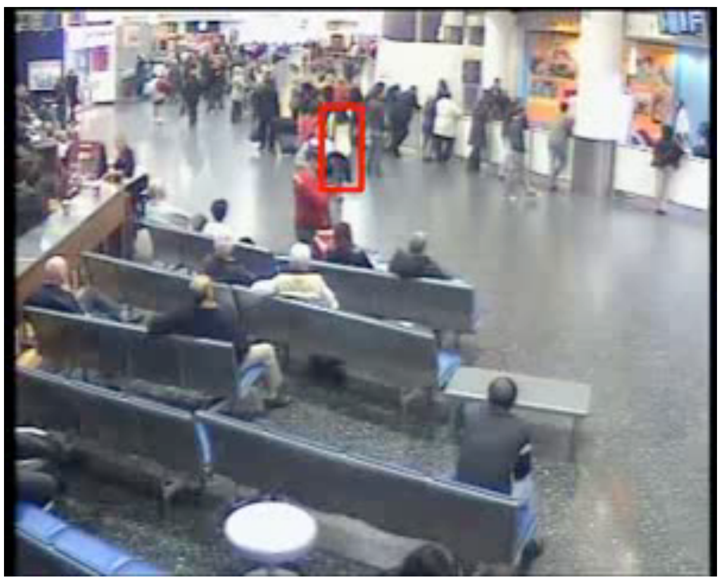
- 30-pixel man
- Coarse-level actions

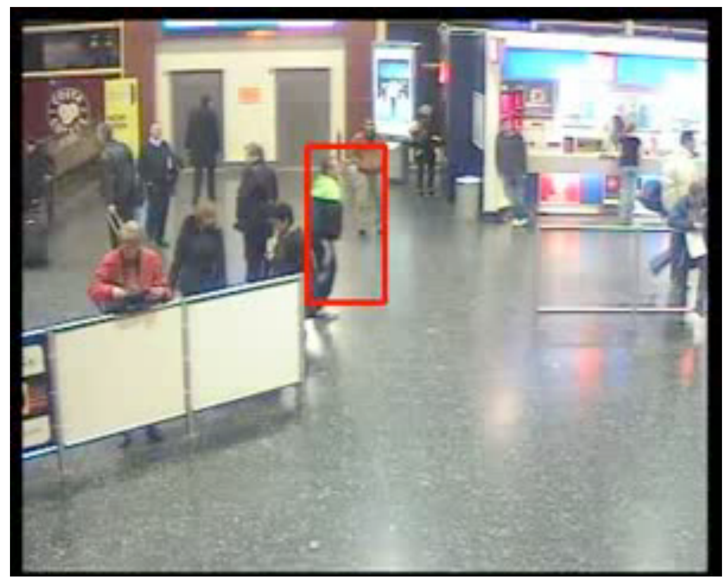
Near field

- 300-pixel man
- Find and track limbs

Applications - Surveillance

- Automated video surveillance
  - Draw attention to actions of interest
  - Save human operator time





Yang, Lan, Mori TRECVid 2009

Applications – Scientific Data Collection

Automatically detect falls, near-falls

SFU Vision and Media Lab

Applications – Road Safety

- Collect data on pedestrian behaviour
  - Collaboration with Saunier and Sayed (UBC, EPM Civil Engineering)

SFU Vision and Media Lab

Applications - HCI

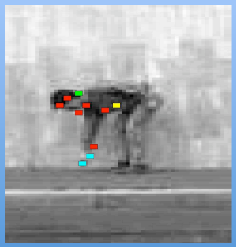
SFU Vision and Media Lab

Structured Models

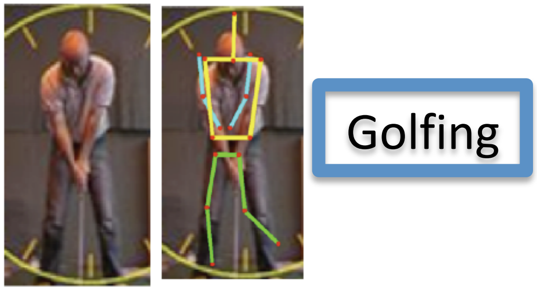
- Models that account for spatial and temporal structure of actions
  - Flexible
    - E.g. local feature models
  - Capture the Gestalt
    - E.g. template representations
- This talk: representations and algorithms for structured models of human actions

Outline

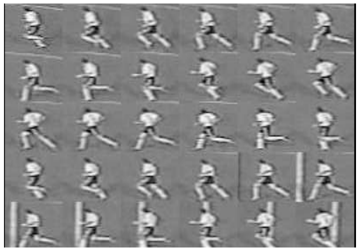
- Combined parts and whole model
  - Wang and Mori NIPS 2008, CVPR 2009



- Latent pose estimation
  – Yang et al. CVPR 2010



- "Bag-of-words" sequence model
  – Wang and Mori T-PAMI 2009



SFU Vision and Media Lab

Appearance vs. Motion

Jackson Pollock
Number 21 (detail)

Spatial Motion Descriptor

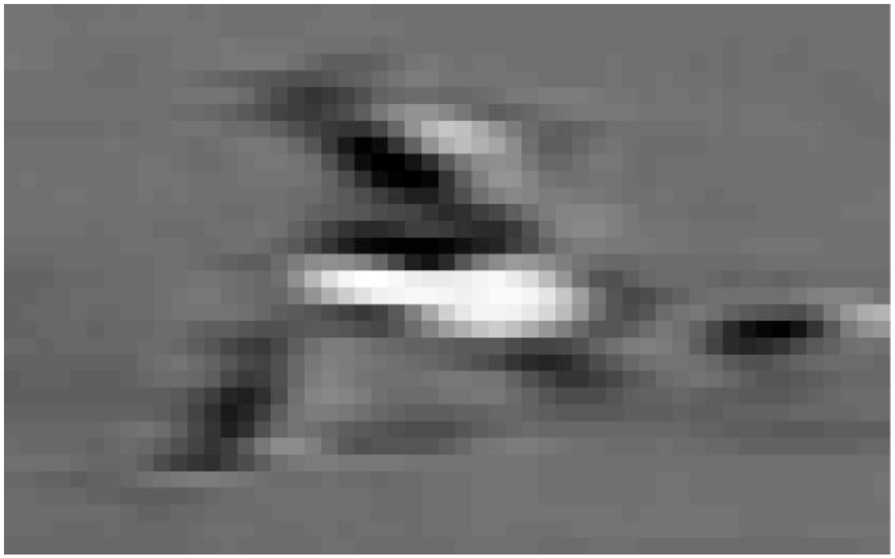
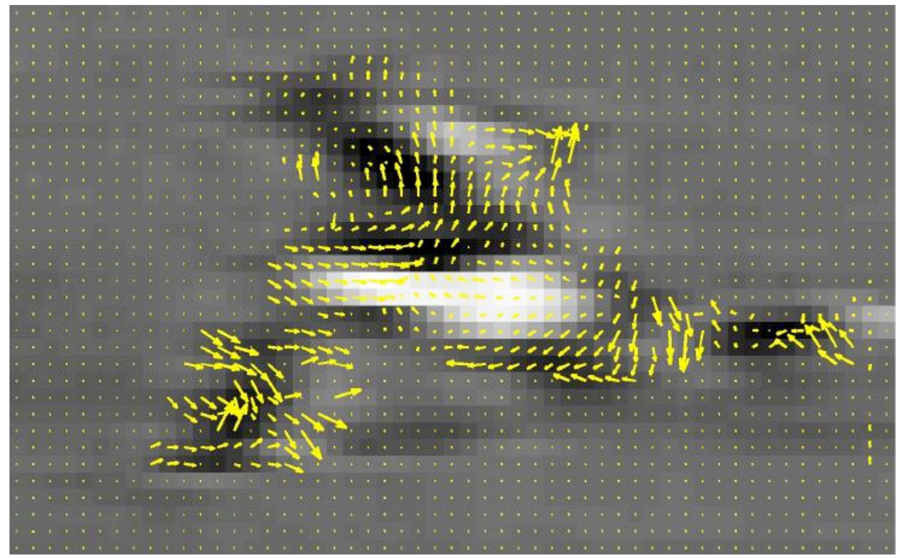
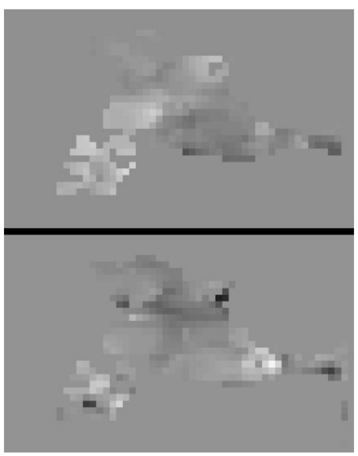


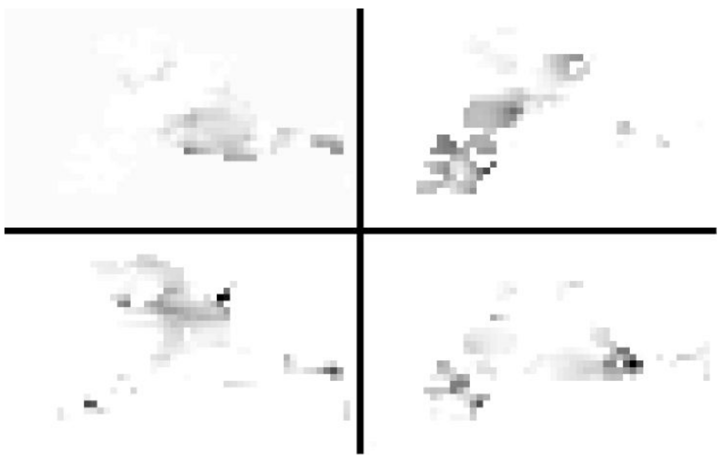
Image frame



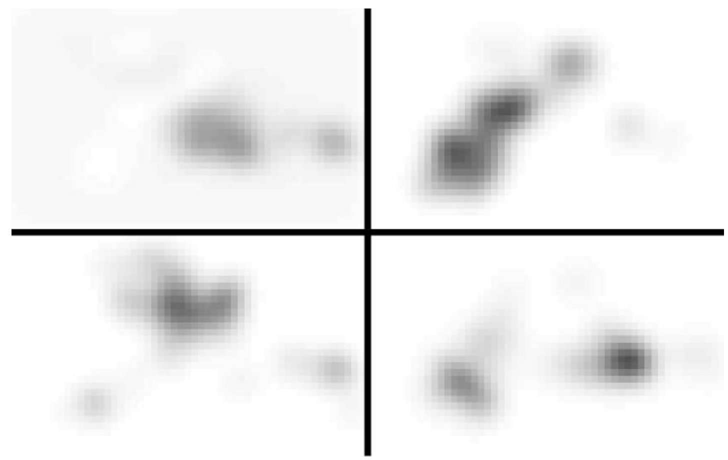
Optical flow



F_x, F_y

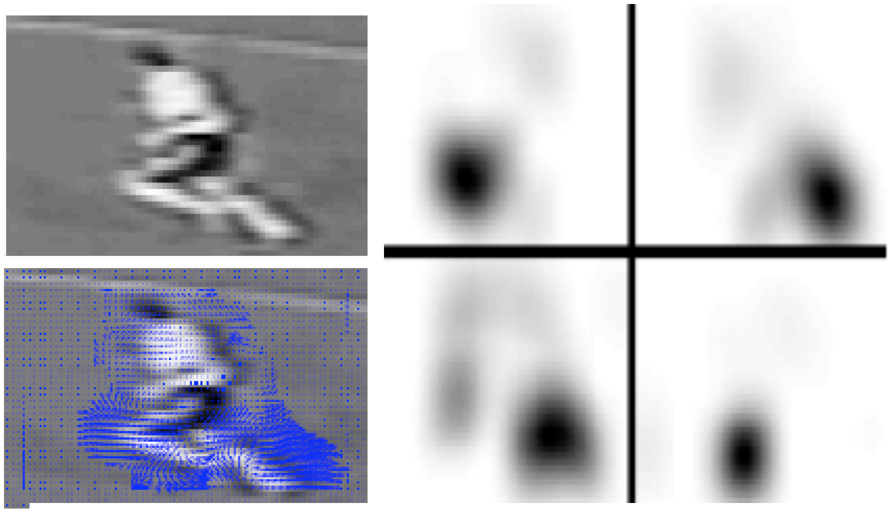


$F_x^-, F_x^+, F_y^-, F_y^+$



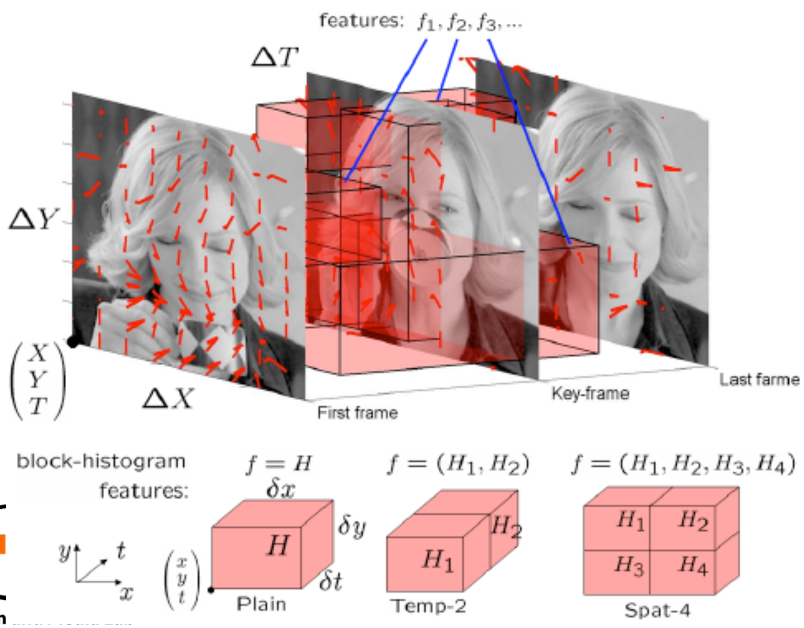
blurred $F_x^-, F_x^+, F_y^-, F_y^+$

Previous Work



Large-scale feature

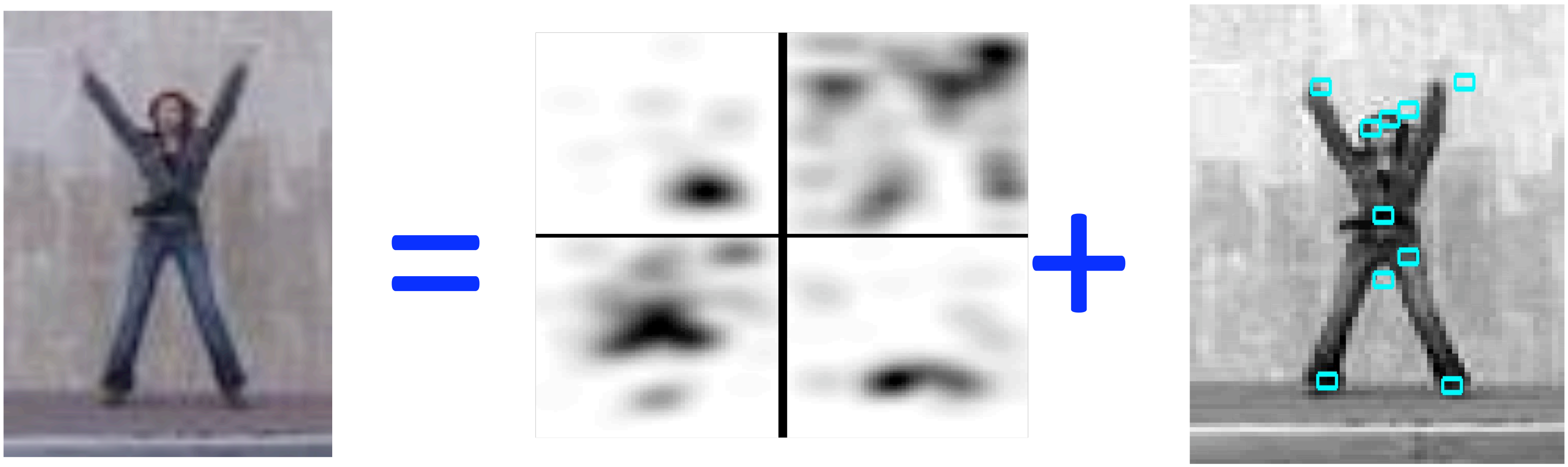
[e.g. Efros, Berg, Mori, Malik, ICCV03]



Local patches

[e.g. Laptev & Perez, ICCV07]

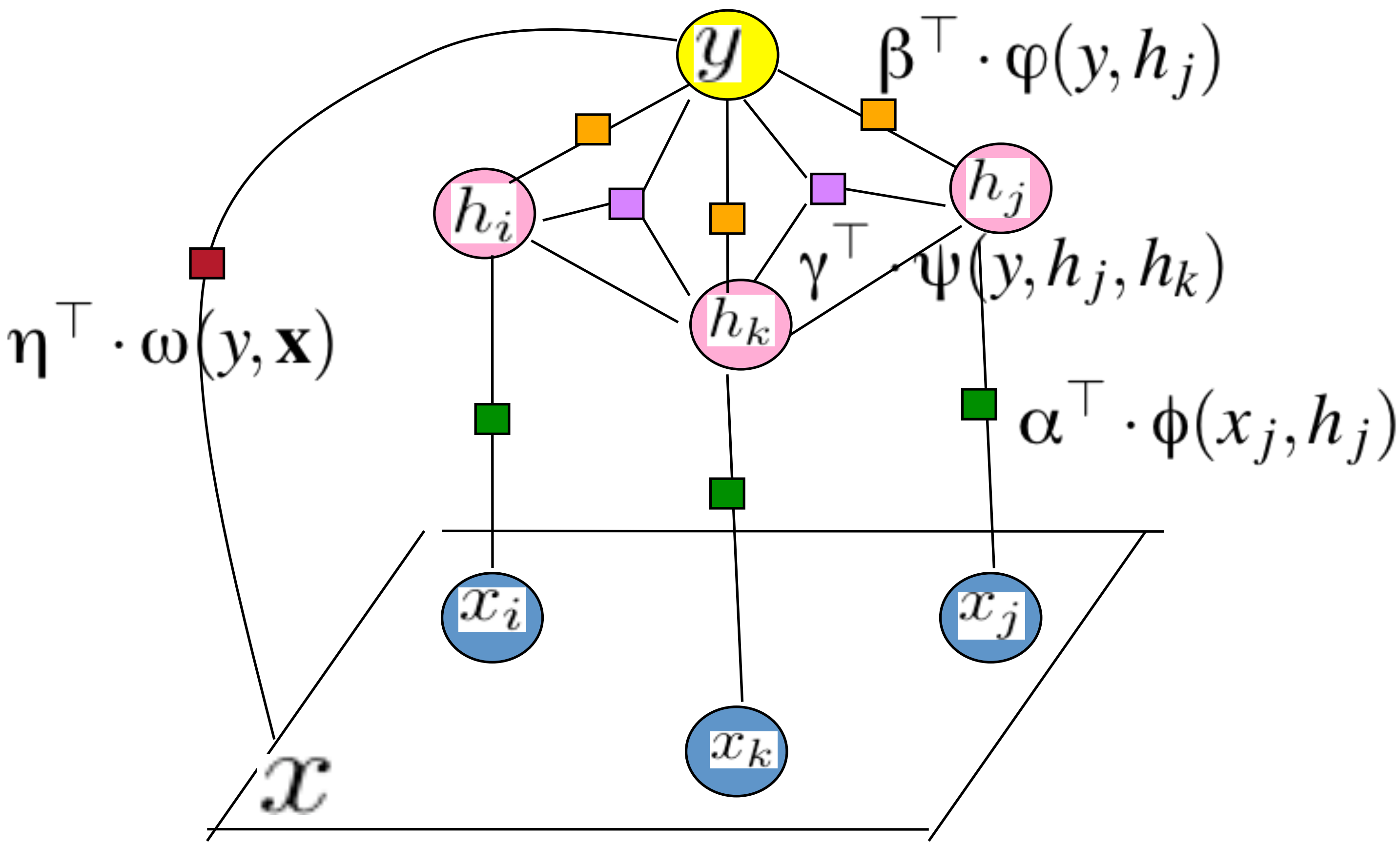
Large vs. Small Scale Features



Challenge: How to combine in a principled manner?

SFU Vision and Media Lab

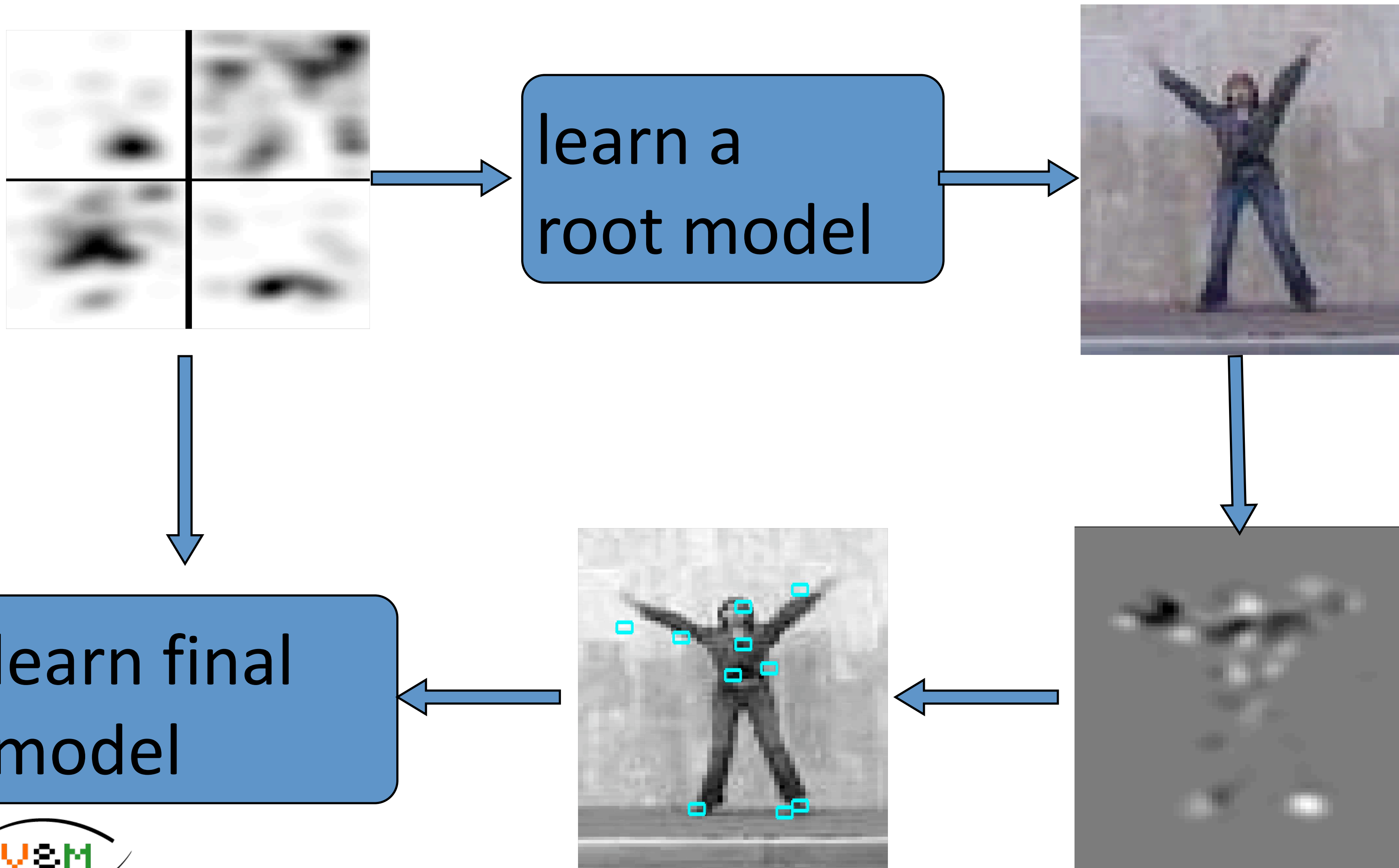
Hidden Conditional Random Field



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

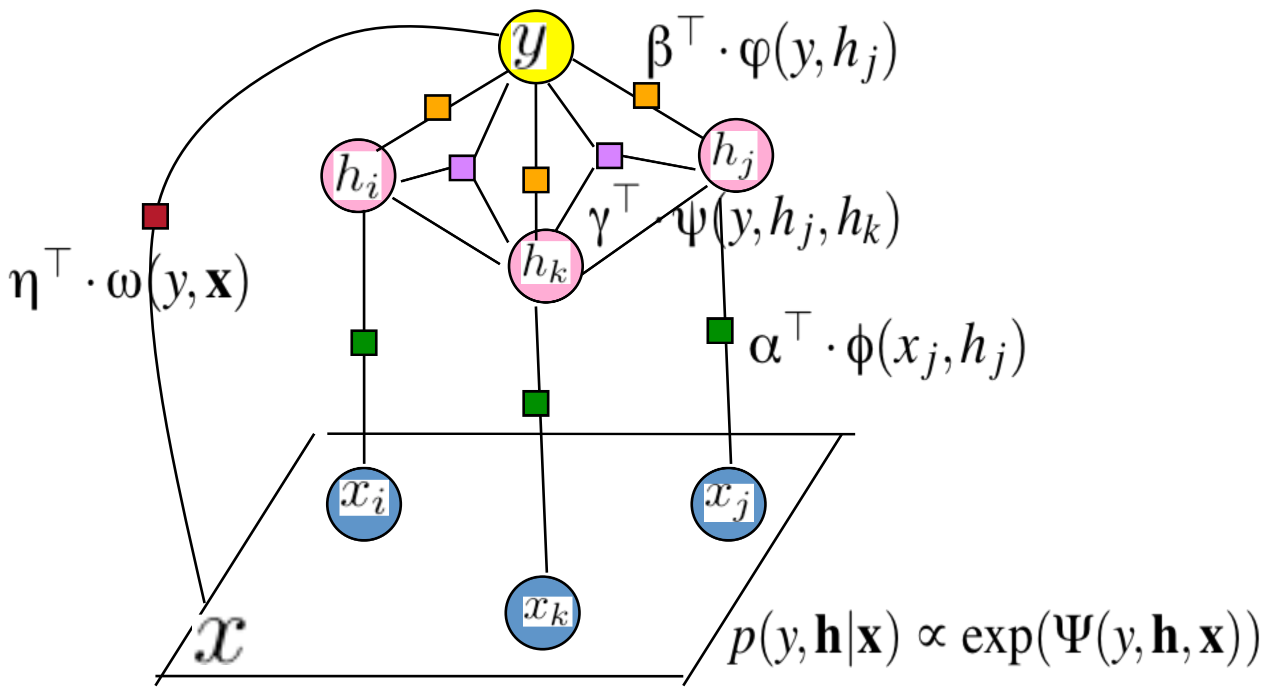
SFU Vision and Media Lab

Finding Parts



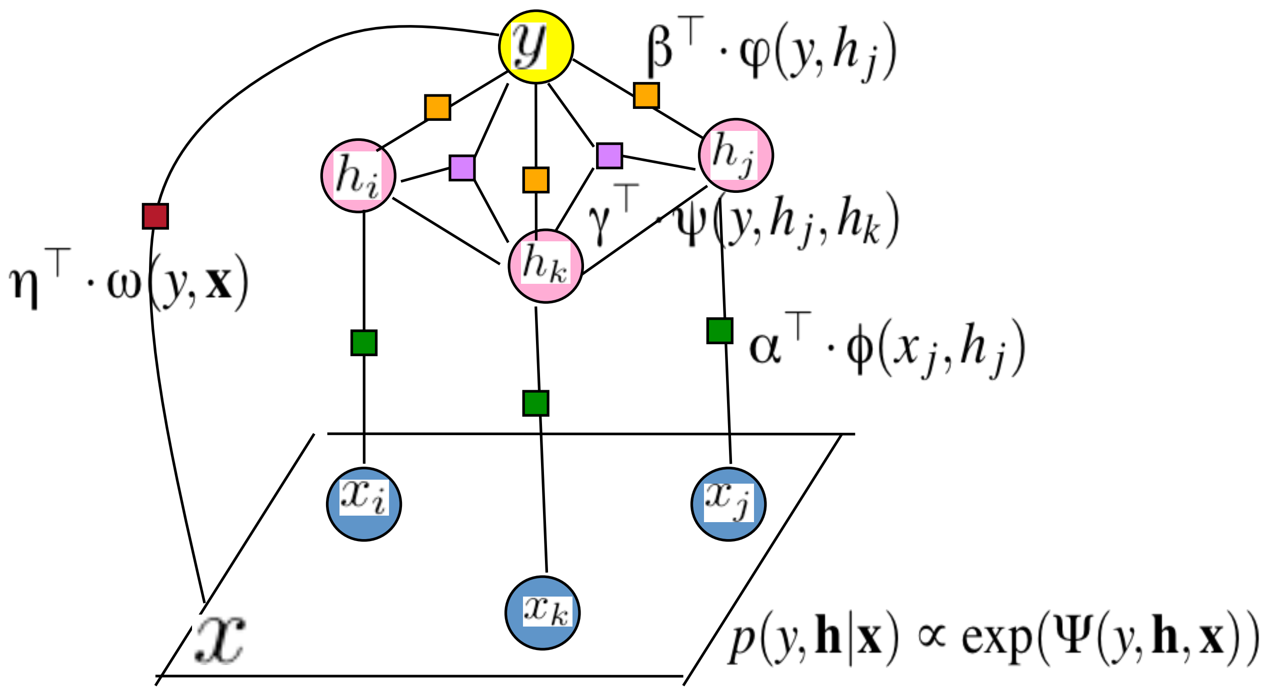
SFU Vision and Media Lab

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)

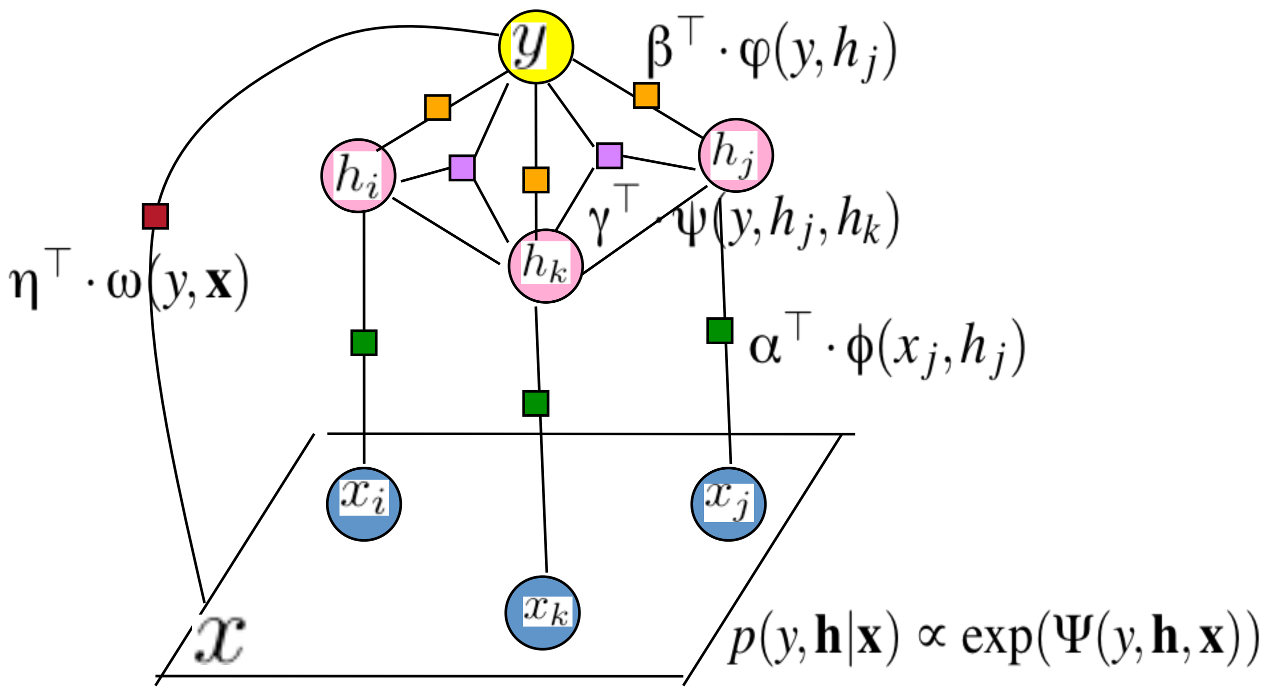
Conditional Likelihood



- Choose parameters to make likelihood on ground-truth labels as large as possible

$$\ell = \sum_t \log p(y^t|\mathbf{x}^t) = \sum_t \log\left(\boxed{\sum_{\mathbf{h}} p(y^t, \mathbf{h}|\mathbf{x}^t)}\right)$$

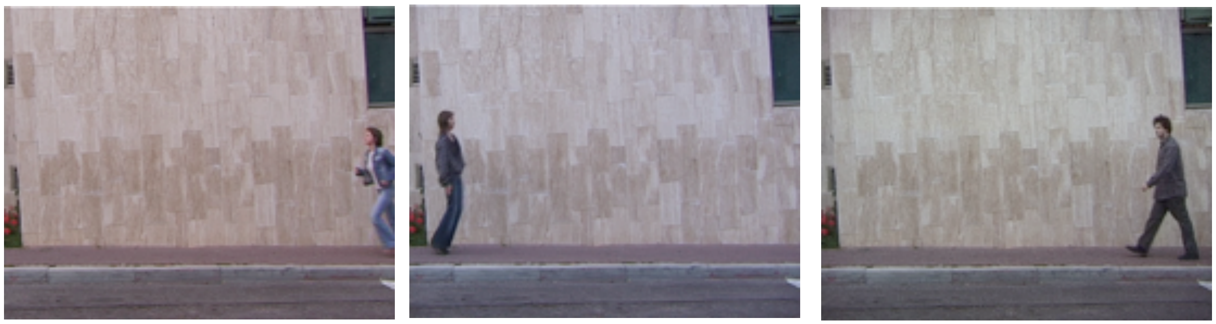
Max-Margin



- Choose parameters to make score on ground-truth label higher than any competing label

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

Experiments: Weizmann dataset

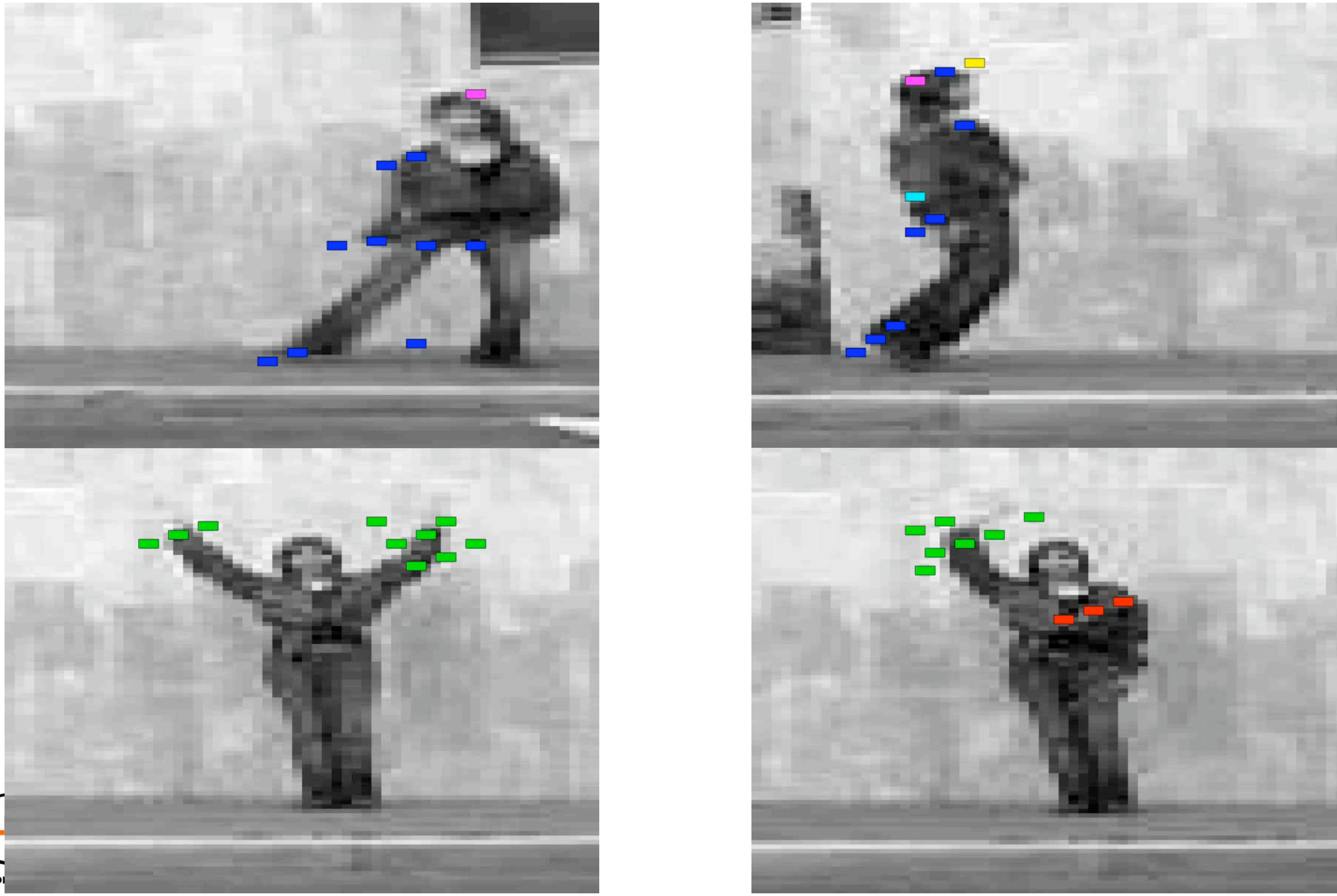


- Benchmark dataset
  - 9 actions
  - 9 subjects

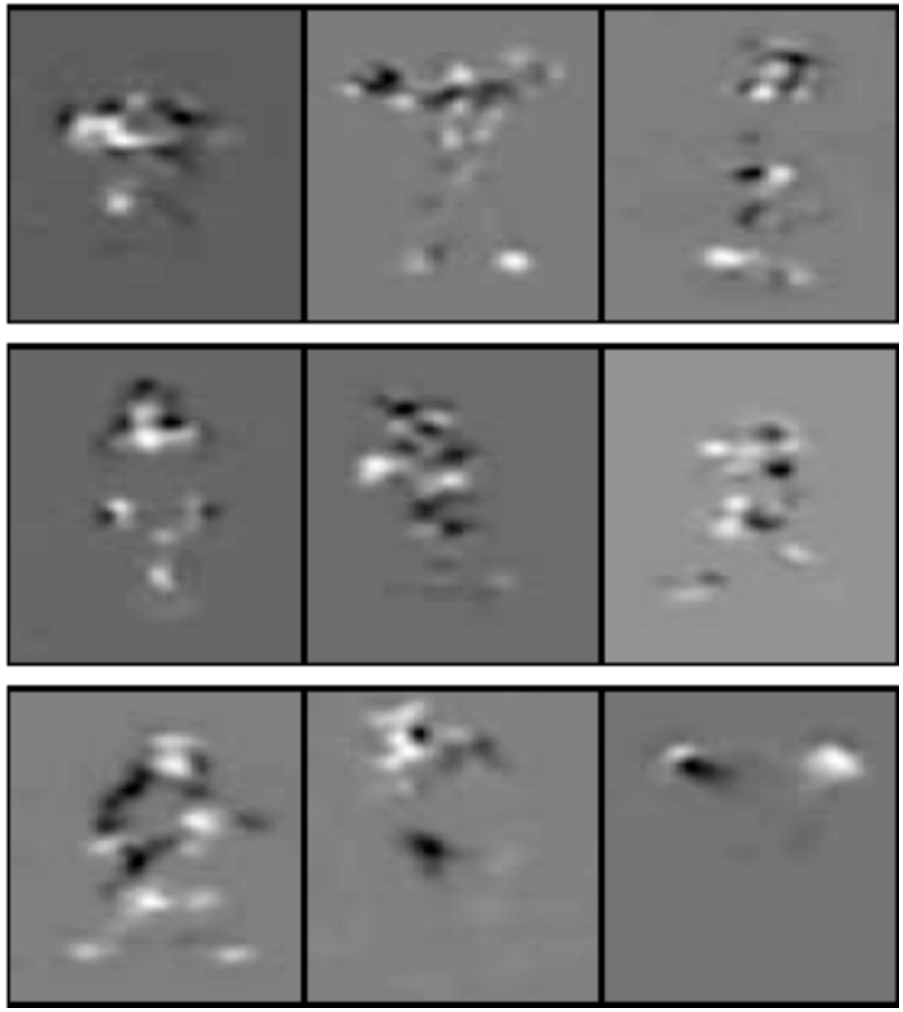
Method	Accuracy
Ours (MM-hCRF)	100%
Ours (CL-hCRF)	97.2%
Jhuang & Poggio ICCV07	98.8%
Niebles & Fei-Fei BMVC06	72.8%

SFU Vision and Media Lab

Inferred Part Labels



Visualization of Learned Model



SFU Vision and Media Lab

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

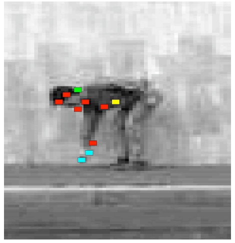
CL $$\log \sum_{\mathbf{h}} p(Y = y^t, \mathbf{h} | \mathbf{x}^t) \text{ vs. } \log \sum_{\mathbf{h}} p(Y \neq y^t, \mathbf{h} | \mathbf{x}^t)$$

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

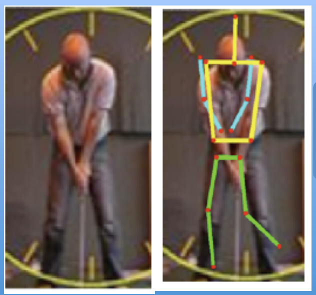
SFU Vision and Media Lab

Outline

- Combined parts and whole model
  - Wang and Mori NIPS 2008, CVPR 2009

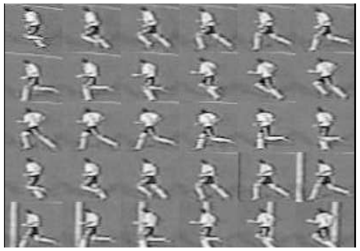


- Latent pose estimation
  – Yang et al. CVPR 2010



Golfing

- "Bag-of-words" sequence model
  – Wang and Mori T-PAMI 2009



SFU Vision and Media Lab

Goal

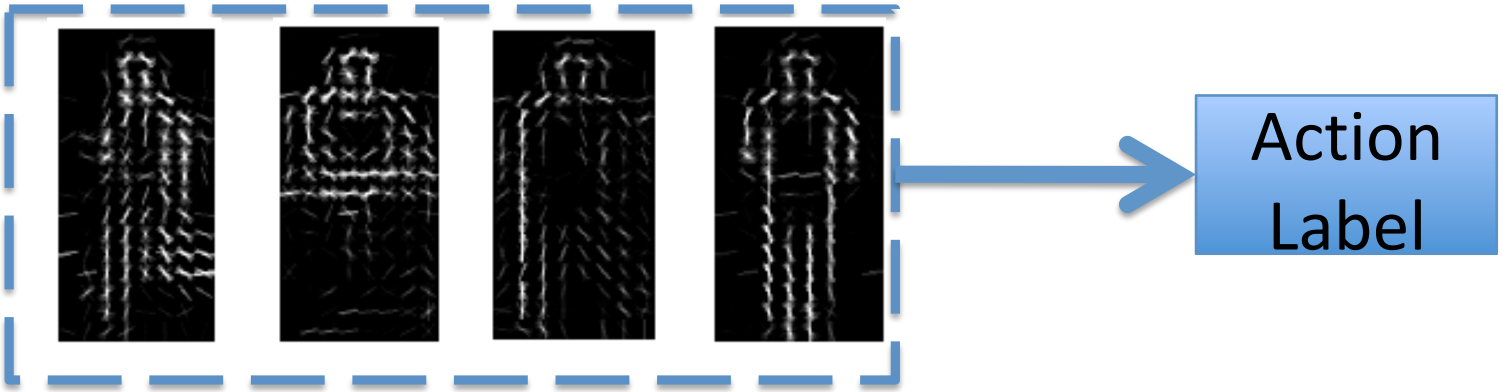
- Action recognition from still images
  - News/sports image retrieval and analysis
  - An important cue for video-based action recognition

SFU Vision and Media Lab

Previous work

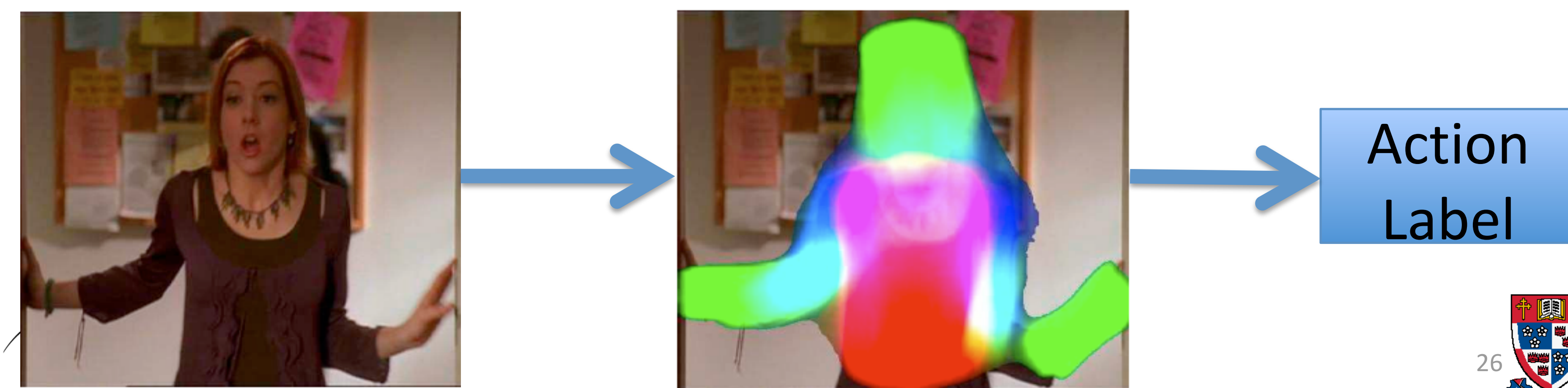
- Global template-based representation

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

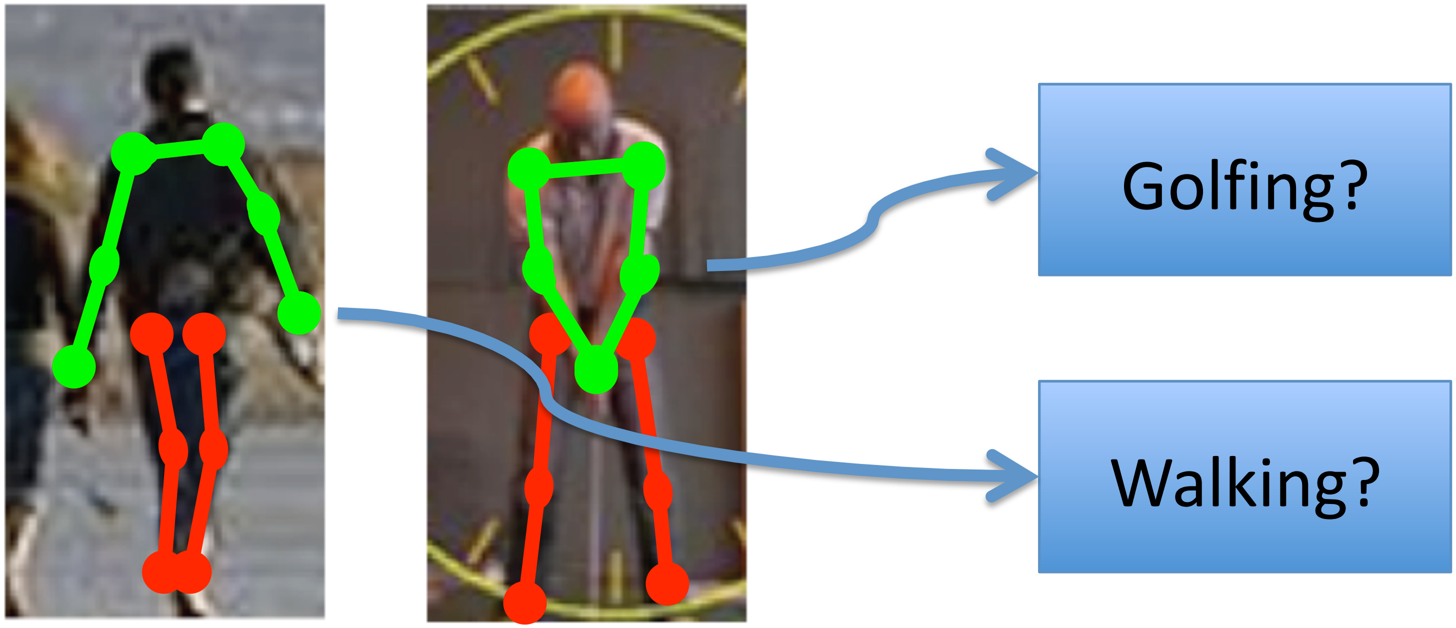


- Pose estimation + action recognition

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



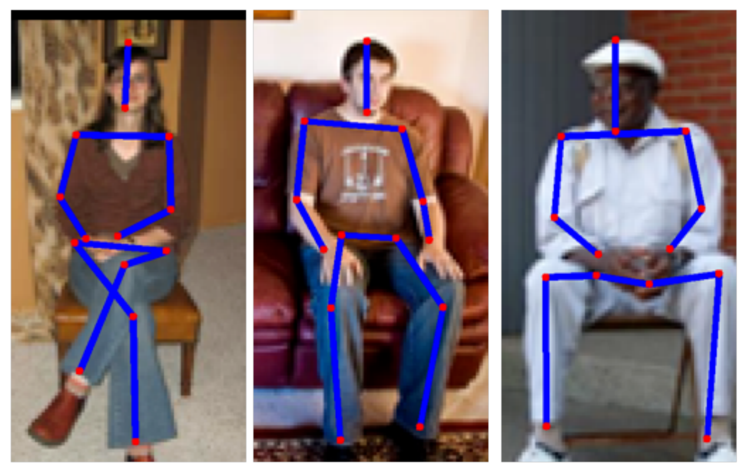
Discriminative Pose



- 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

SFU Vision and Media Lab

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

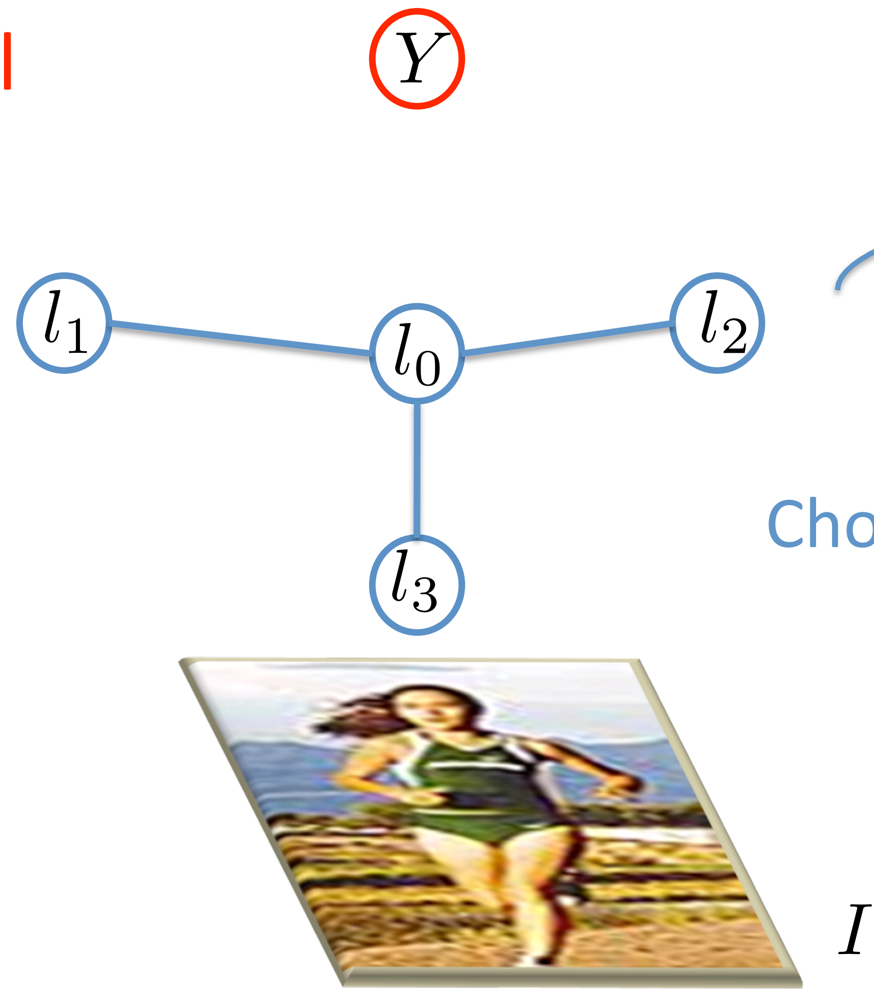
SFU Vision and Media Lab

Model Formulation

Action Label

Pose

Image



Choose best pose L

$$H(I,Y;\Theta) = \max_L \Theta^T \Psi(I,L,Y)$$

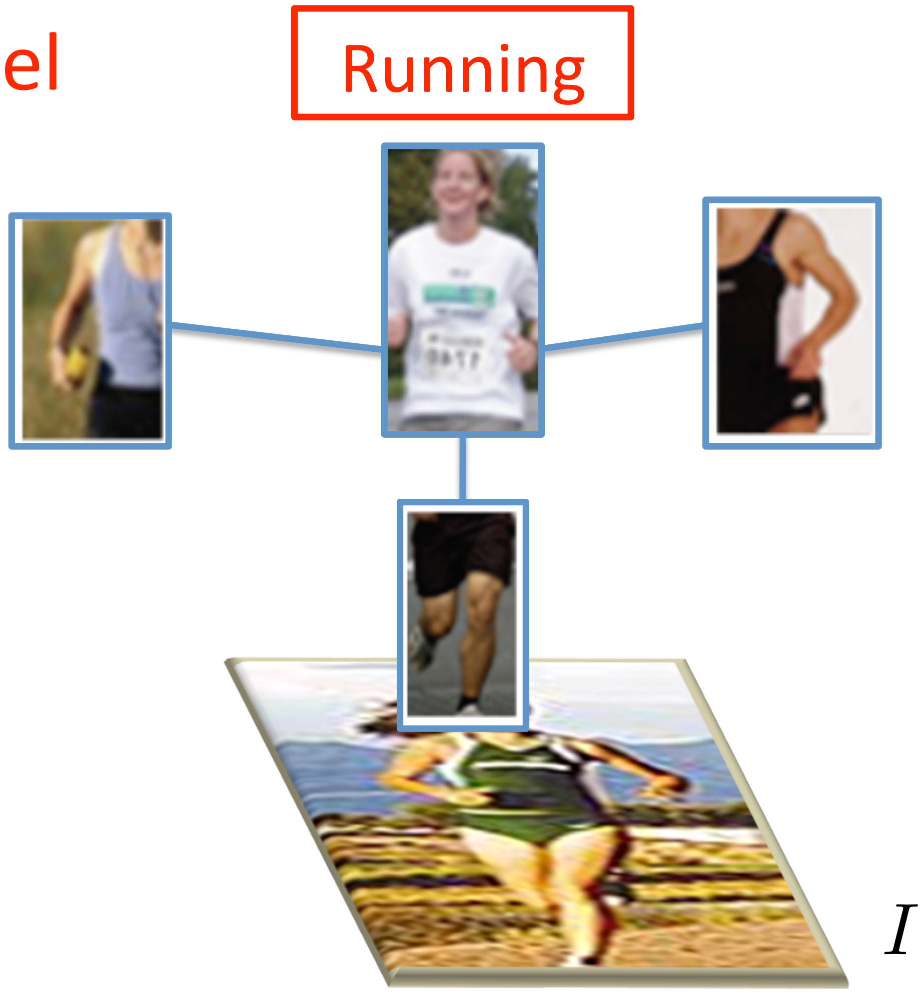
SFU Vision and Media Lab

Model Formulation

Action Label

Pose

Image



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

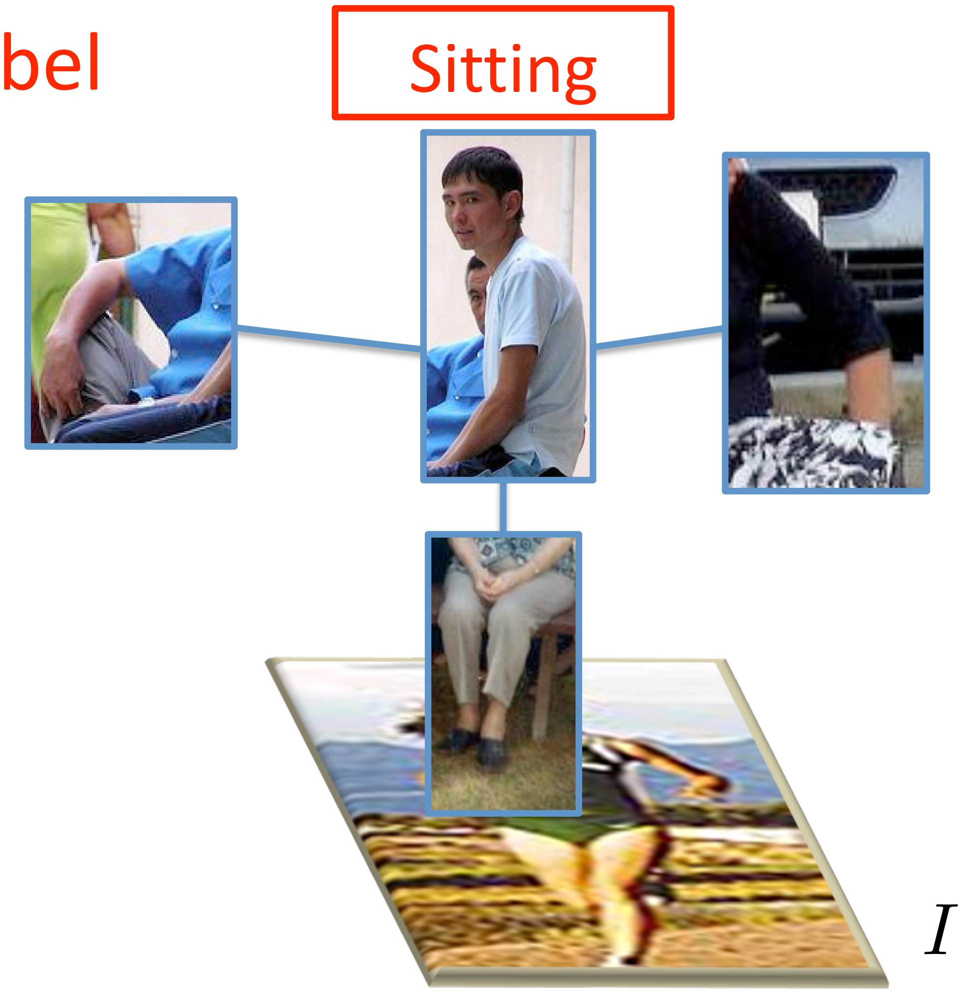
SFU Vision and Media Lab

Model Formulation

Action Label

Pose

Image



Small score for $H(I, Y = Sitting; \Theta)$

SFU Vision and Media Lab

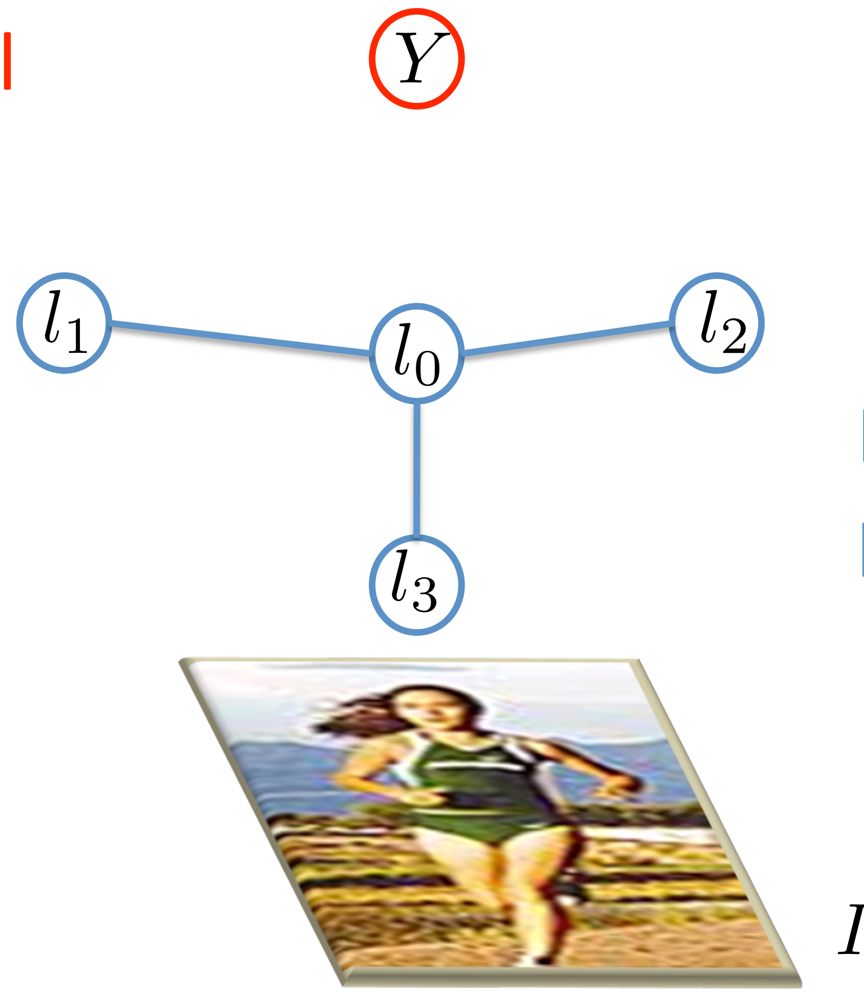
Model Details I

Action Label

Pose

Relative body part locations

Image



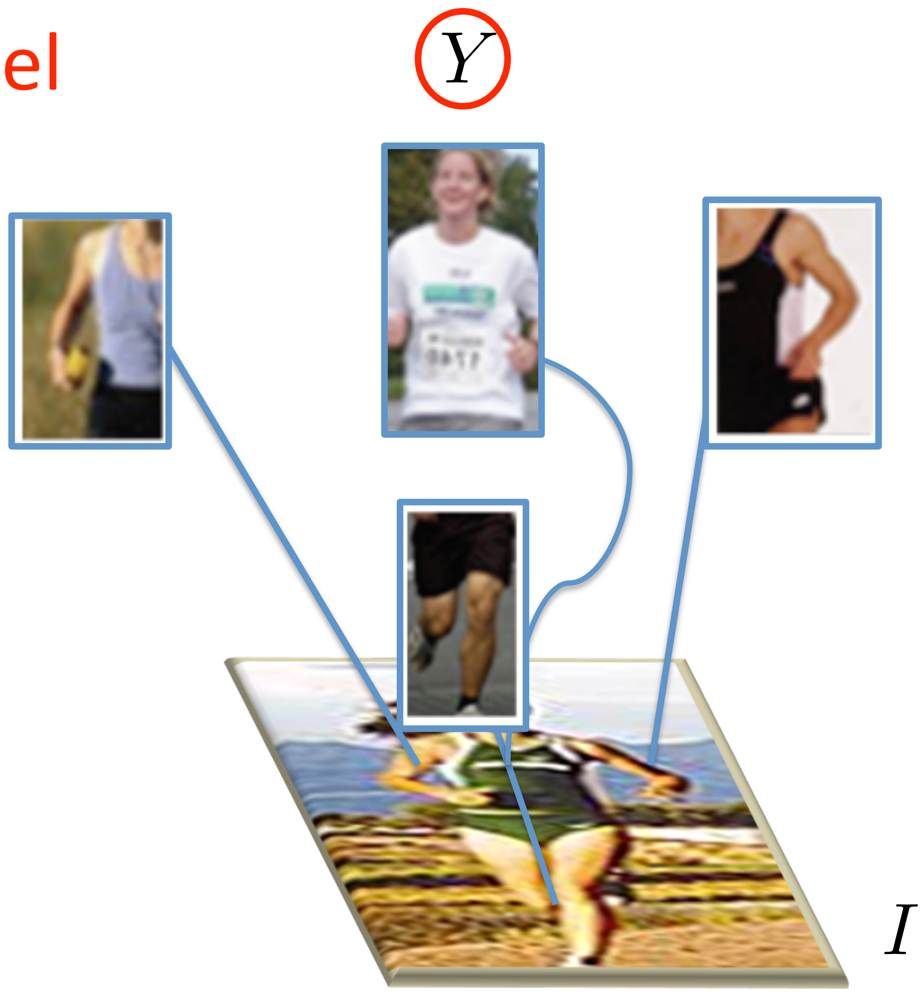
SFU Vision and Media Lab

Model Details II

Action Label

Pose

Image



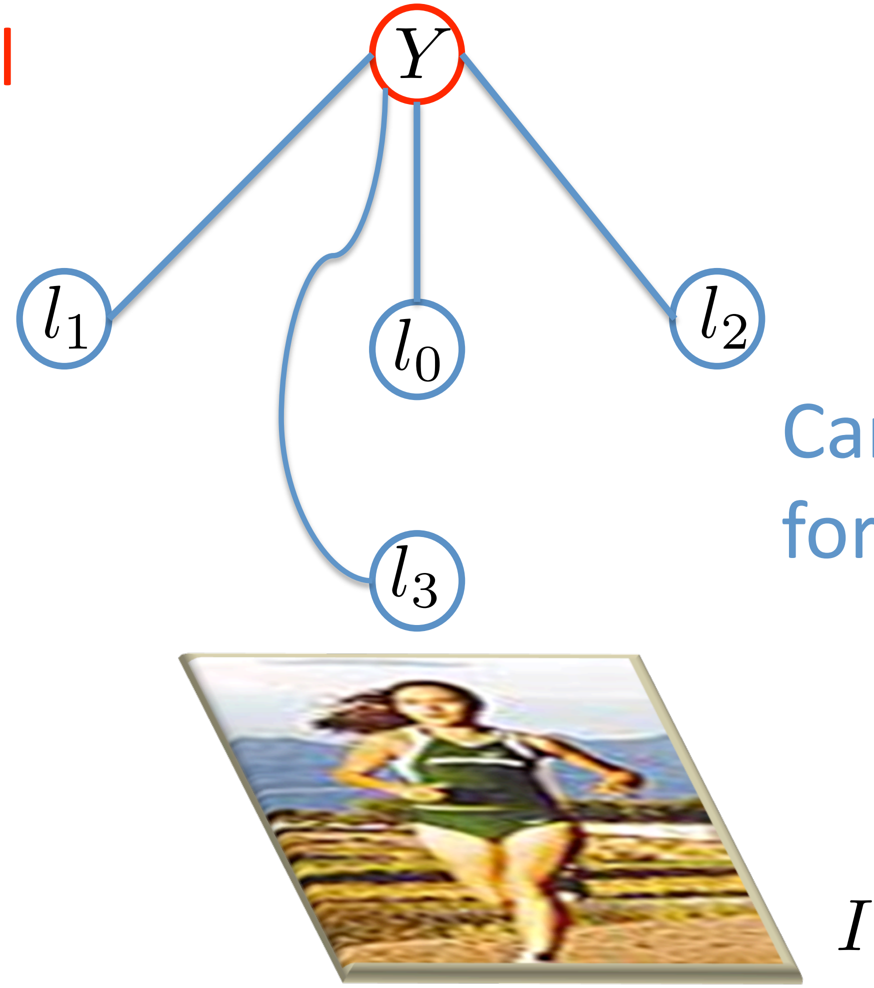
Poselet matching

SFU Vision and Media Lab

Model Details III

Action Label

Pose

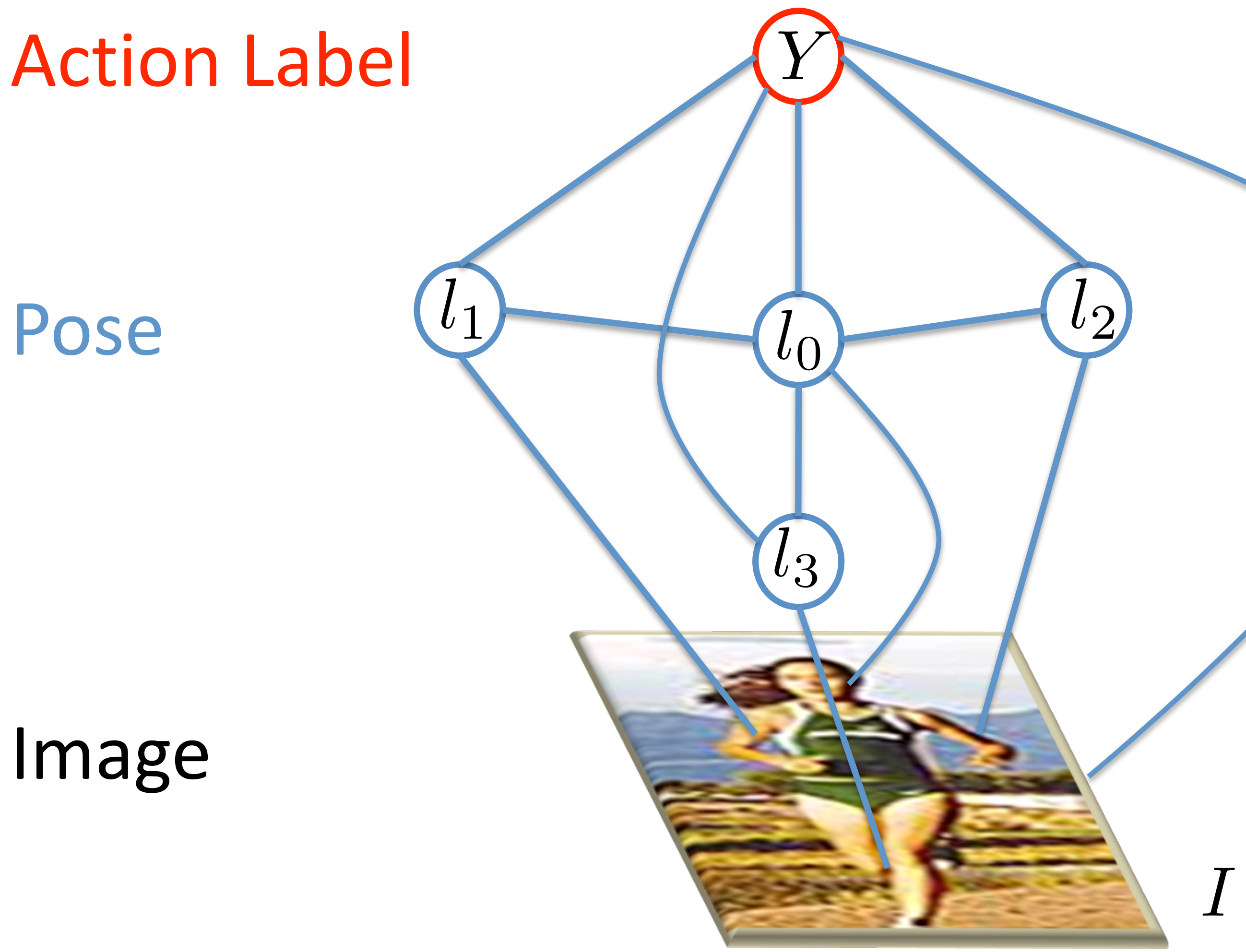


Canonical poses for an action

Image

SFU Vision and Media Lab

Full Model



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

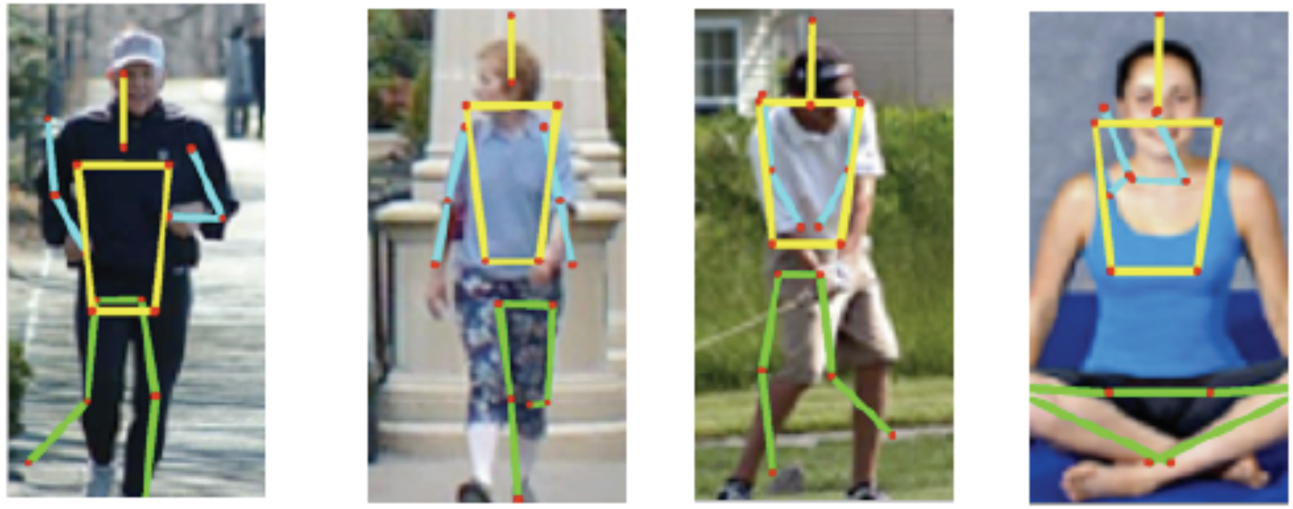
	Running	Walking	PlayGolf	Sitting	Dancing
Running	0.81	0.06	0.00	0.03	0.10
Walking	0.38	0.46	0.02	0.00	0.13
PlayGolf	0.34	0.09	0.27	0.04	0.25
Sitting	0.11	0.05	0.02	0.61	0.22
Dancing	0.31	0.13	0.02	0.07	0.47

Baseline – HOG/SVM:
52% per class accuracy

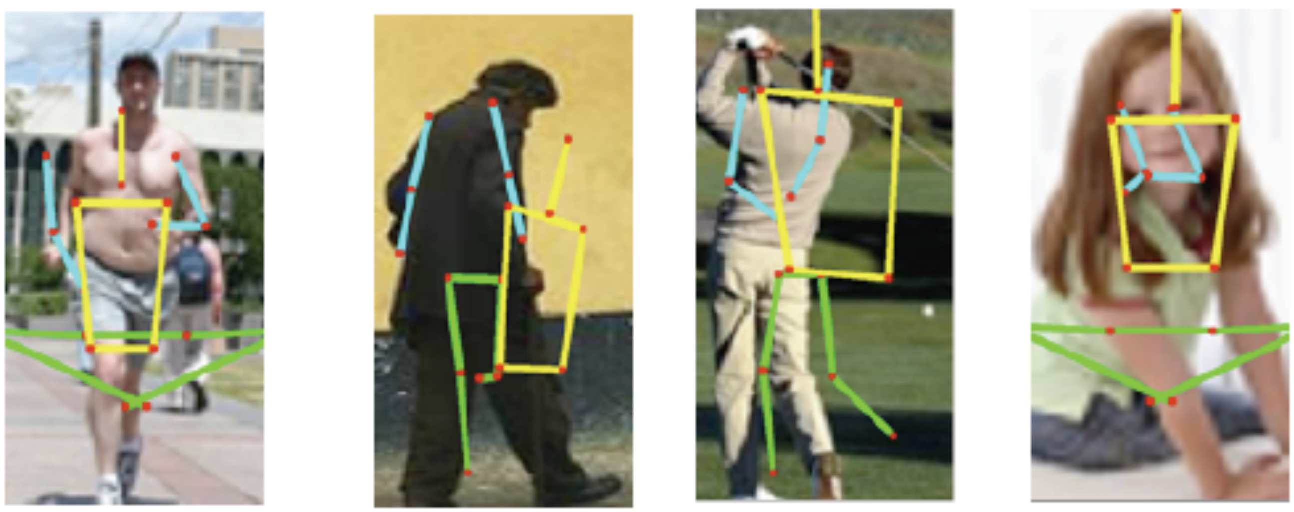
	Running	Walking	PlayGolf	Sitting	Dancing
Running	0.66	0.08	0.07	0.07	0.13
Walking	0.24	0.48	0.12	0.01	0.15
PlayGolf	0.10	0.03	0.65	0.03	0.18
Sitting	0.02	0.01	0.06	0.79	0.13
Dancing	0.15	0.08	0.12	0.12	0.53

Ours – Latent Pose:
62% per class accuracy

Visualization of latent pose



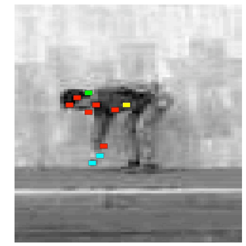
Successful classification examples



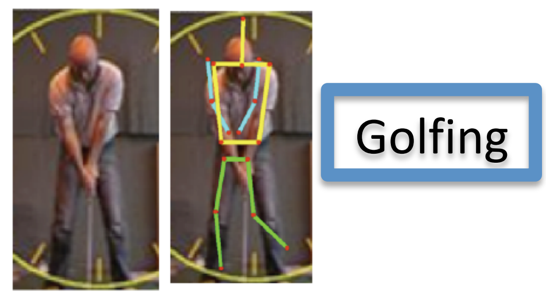
Unsuccessful classification examples

Outline

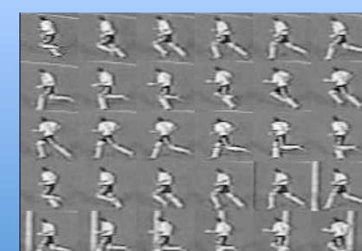
- Combined parts and whole model
  - Wang and Mori NIPS 2008, CVPR 2009



- Latent pose estimation
  – Yang et al. CVPR 2010



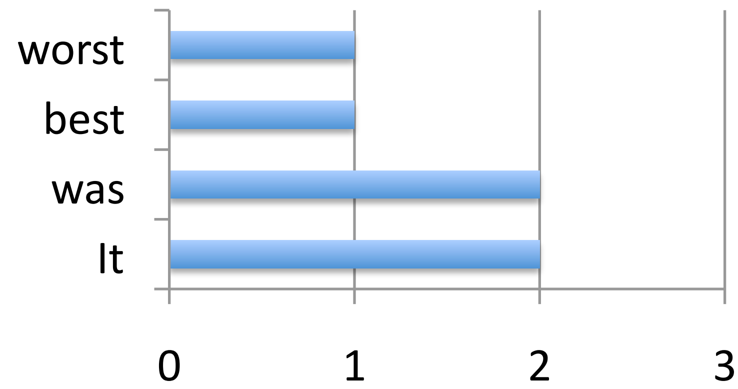
- “Bag-of-words” sequence model
  – Wang and Mori T-PAMI 2009



SFU Vision and Media Lab

"Bag-of-Words" Models

- Text document models
  - "It was the best of times, it was the worst of times."



- Bag of Words + Topic Models in Computer Vision
  - Scenes: Fei-Fei & Perona CVPR'05
  - Objects: Sivic et al. ICCV'05, Fergus et al. ICCV'05, Russell et al. CVPR'06
  - Actions: Niebles et al. BMVC'06
  - Human Poses: Bissaco et al. NIPS'06

SFU Vision and Media Lab

Role of Temporal Information

- No temporal info
  - Classify each video frame independently
  - e.g., Efros et al. 03, Shechtman & Irani 05, Fathi & Mori 08

SFU Vision and Media Lab

Role of Temporal Information

? — ? — ? — ? — ? — ? — ? — ? — ? — ?

- Strong temporal info
  - Use hidden Markov Model or grammar on top of video frames
  - e.g. Bobick & Ivanov 98

SFU Vision and Media Lab

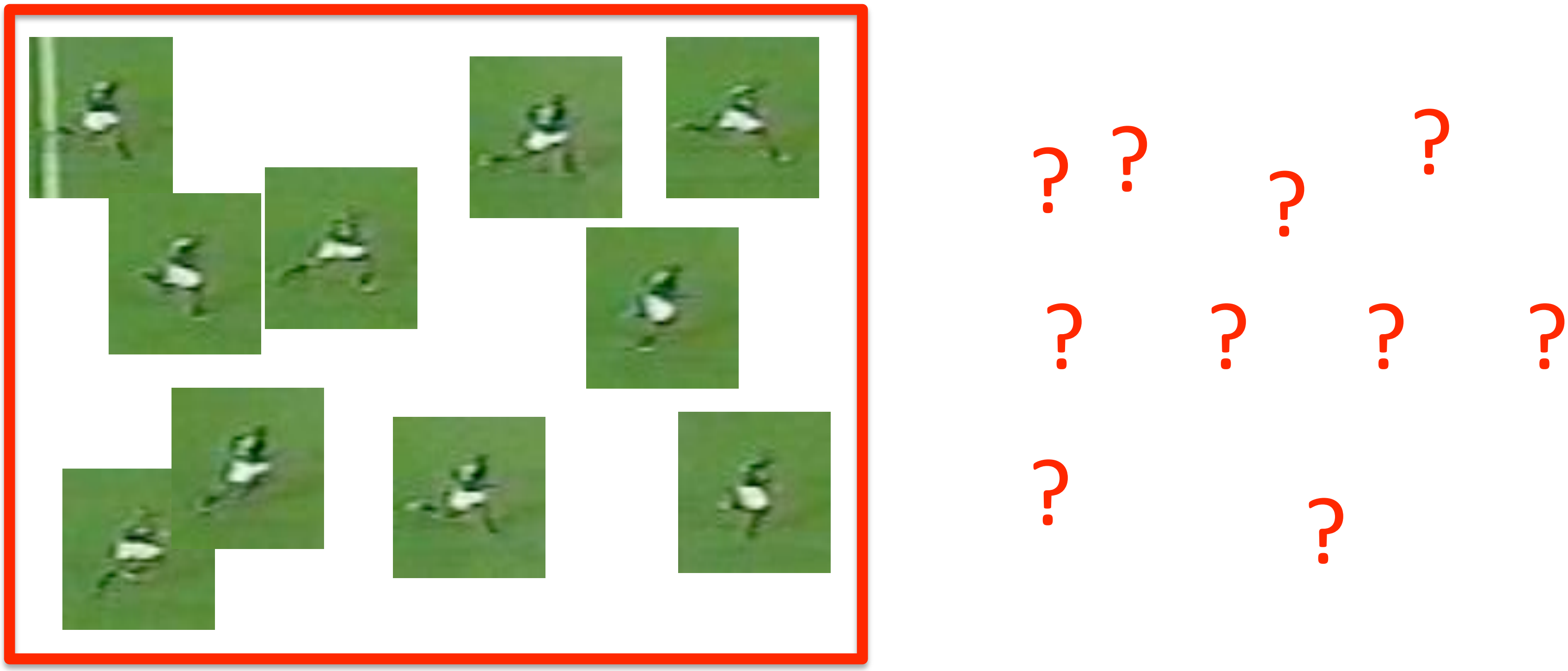
Role of Temporal Information

? ? ? ? ? ? ? ? ? ?

- Our work is somewhere in between
  - Use bag of frames representation
  - Capture some temporal structure (co-occurrences of actions)
  - Simpler than full temporal models

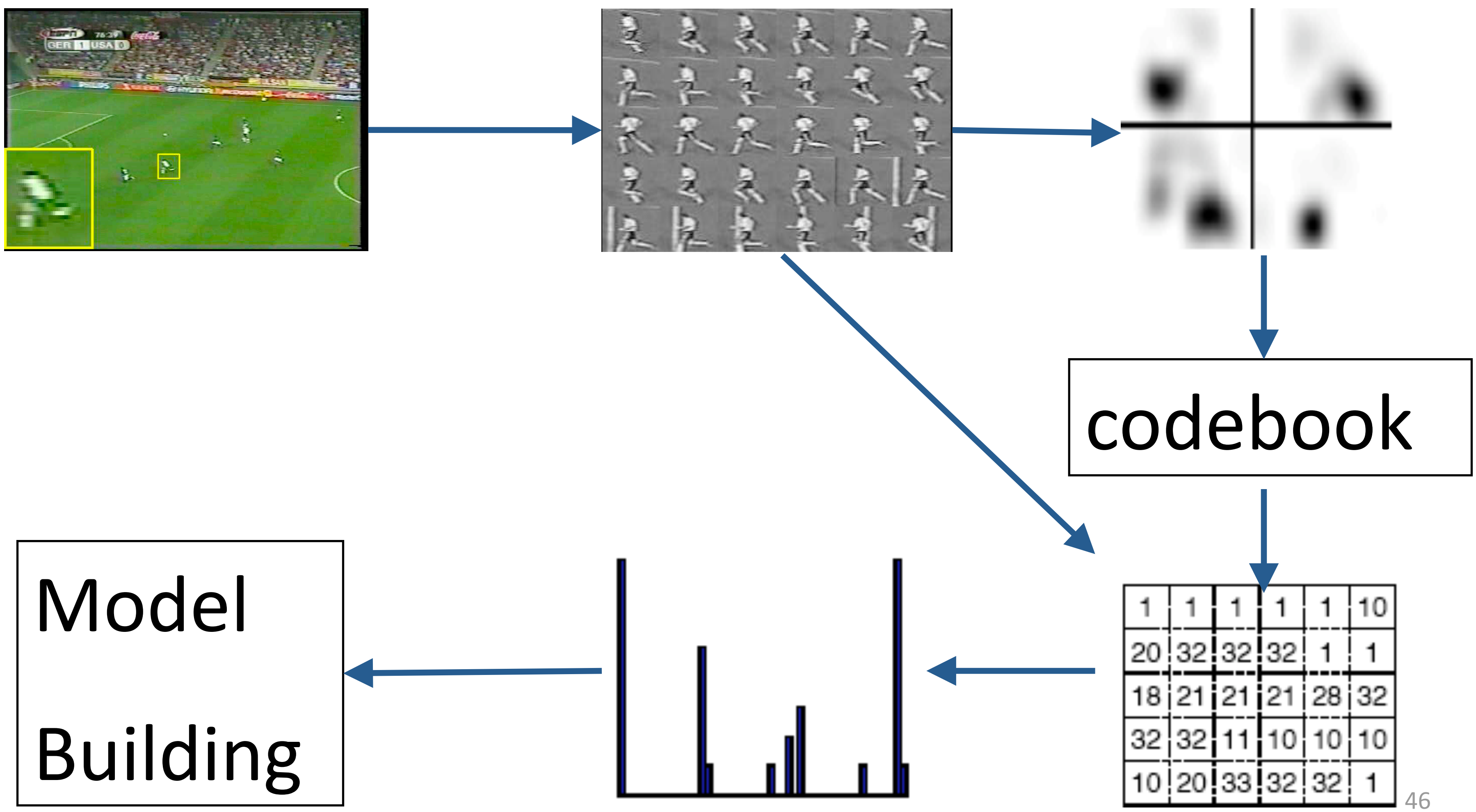
SFU Vision and Media Lab

Role of Temporal Information

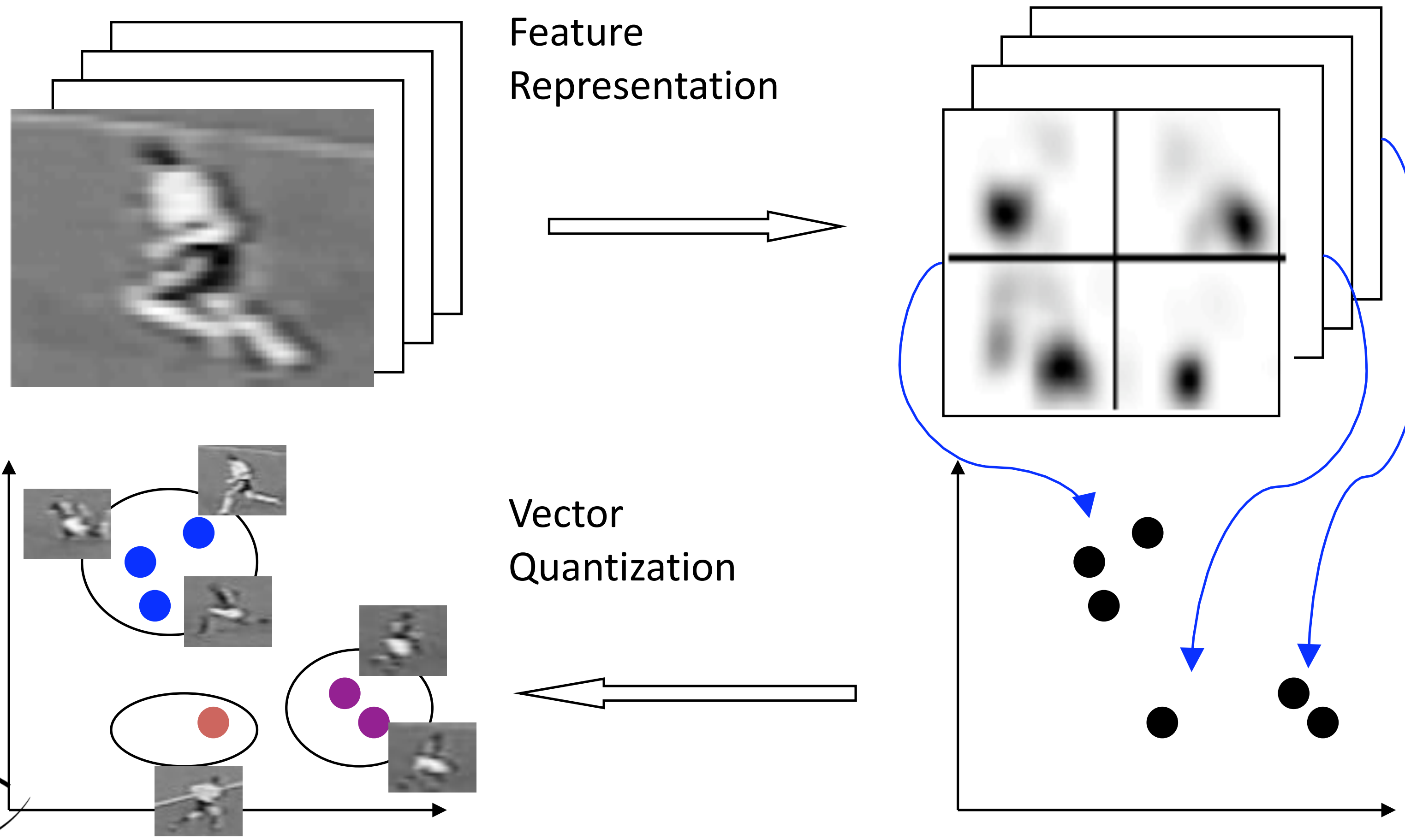


- Our work is somewhere in between
  - Use bag of frames representation
  - Capture some temporal structure (co-occurrences of actions)
  - Simpler than full temporal models

Bag-of-Words Sequence Model

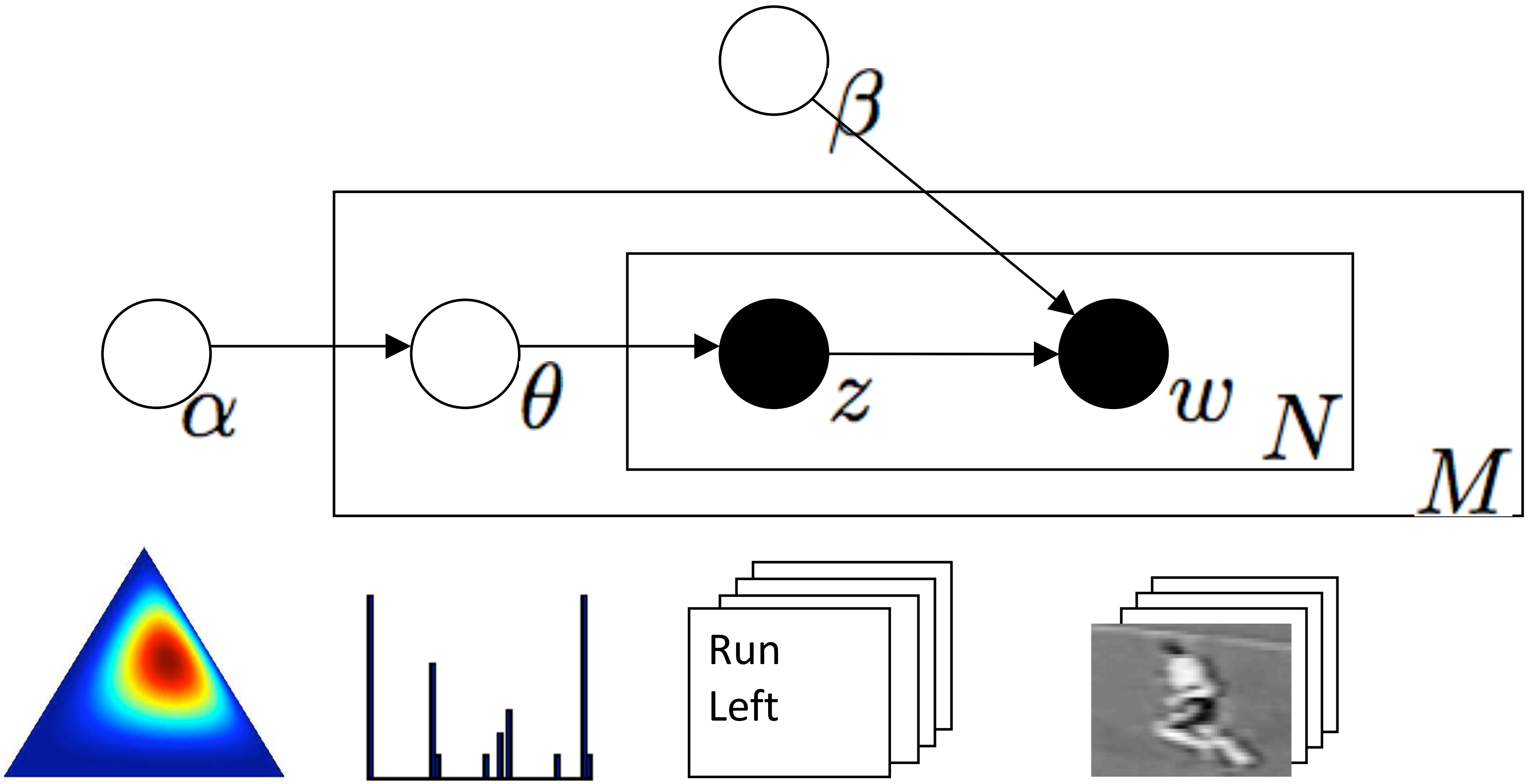


Codebook Formation



SFU Vision and Media Lab

Semi-Latent Dirichlet Allocation



Learning is easier due to decoupling of model parameters
cf. Blei et al. JMLR 2003

Experiments: KTH dataset

- Benchmark dataset
  - 6 actions
  - 25 subjects
  - 4 scenarios

Method	Accuracy
Ours (sLDA)	91.2%
Liu & Shah CVPR08	94.2%
Jhuang and Poggio ICCV07	91.7%
Niebles & Fei-Fei BMVC06	81.5%
Schuldt & Laptev ICPR04	71.7%

	boxing	handclapping	handwaving	jogging	running	walking
boxing	0.94	0.02	0.02	0.00	0.00	0.01
handclapping	0.00	0.98	0.02	0.00	0.00	0.00
handwaving	0.00	0.00	1.00	0.00	0.00	0.00
jogging	0.00	0.00	0.00	0.86	0.11	0.03
running	0.01	0.00	0.00	0.26	0.71	0.02
walking	0.00	0.00	0.00	0.01	0.01	0.98

Experiments: Soccer Dataset

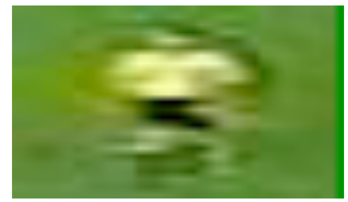
- Real actions, moving camera, poor video
- 8 classes of actions
- 4500 frames of labeled data

Action	Our method (sLDA)	Efros et al. (k-NN)
Run left 45	0.64	0.67
Run left	0.77	0.58
Walk left	1.00	0.68
Walk in/out	0.86	0.79
Run in/out	0.81	0.59
Walk right	0.86	0.68
Run right	0.71	0.58
Run right 45	0.66	0.66

SFU Vision and Media Lab

Experiments: Irregularity detection



- sLDA is full probabilistic model
- Can detect most unusual sequences via likelihood
  - Sequences with lowest likelihood under model shown

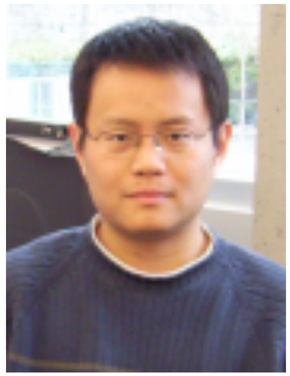
SFU Vision and Media Lab

Conclusion

- Structured models
  - Whole versus parts
    - Learning criterion: conditional likelihood vs. max-margin learning
  - Semantically meaningful parts
    - Latent human pose estimation for action recognition
  - Temporal structure
    - Bag-of-frames
    - Probabilistic model

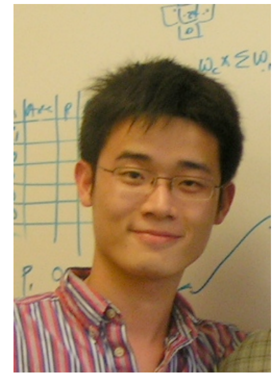
Acknowledgements

Mani Ranjbar



Yang Wang

Tian Lan

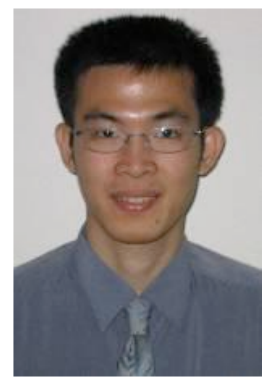


Weilong Yang

Bahman Yari Saeed Khanloo

Mark Bayazit

Alex Couture-Beil



Ferdinand Stefanus

Thank you!

SFU Vision and Media Lab