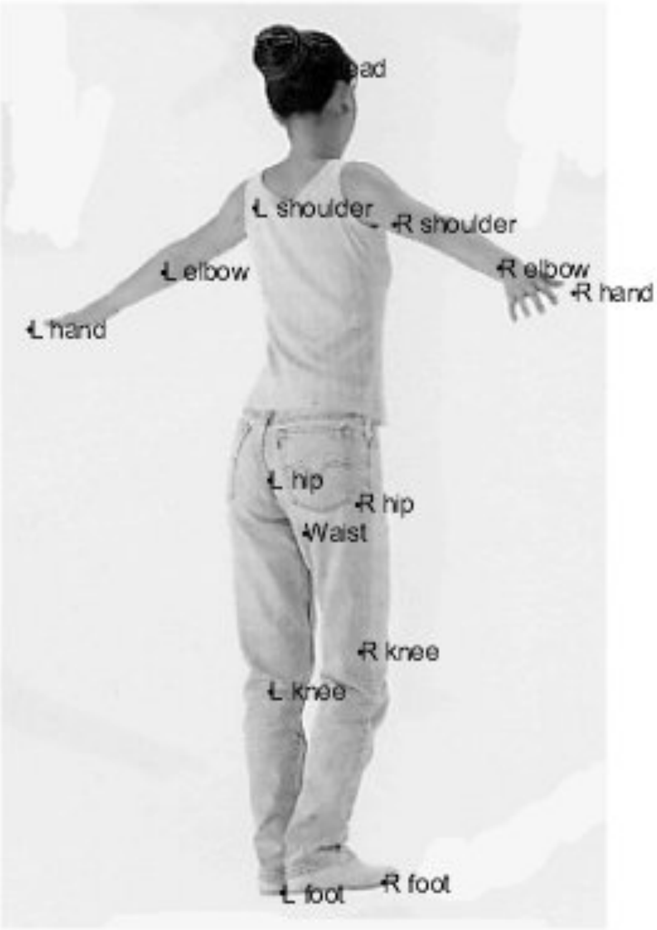
Human Pose Estimation

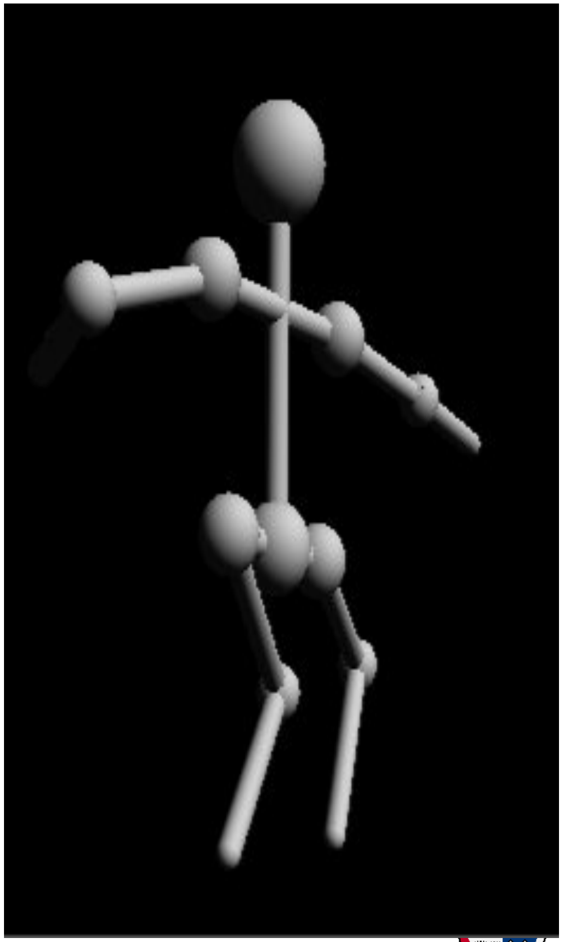
Greg Mori

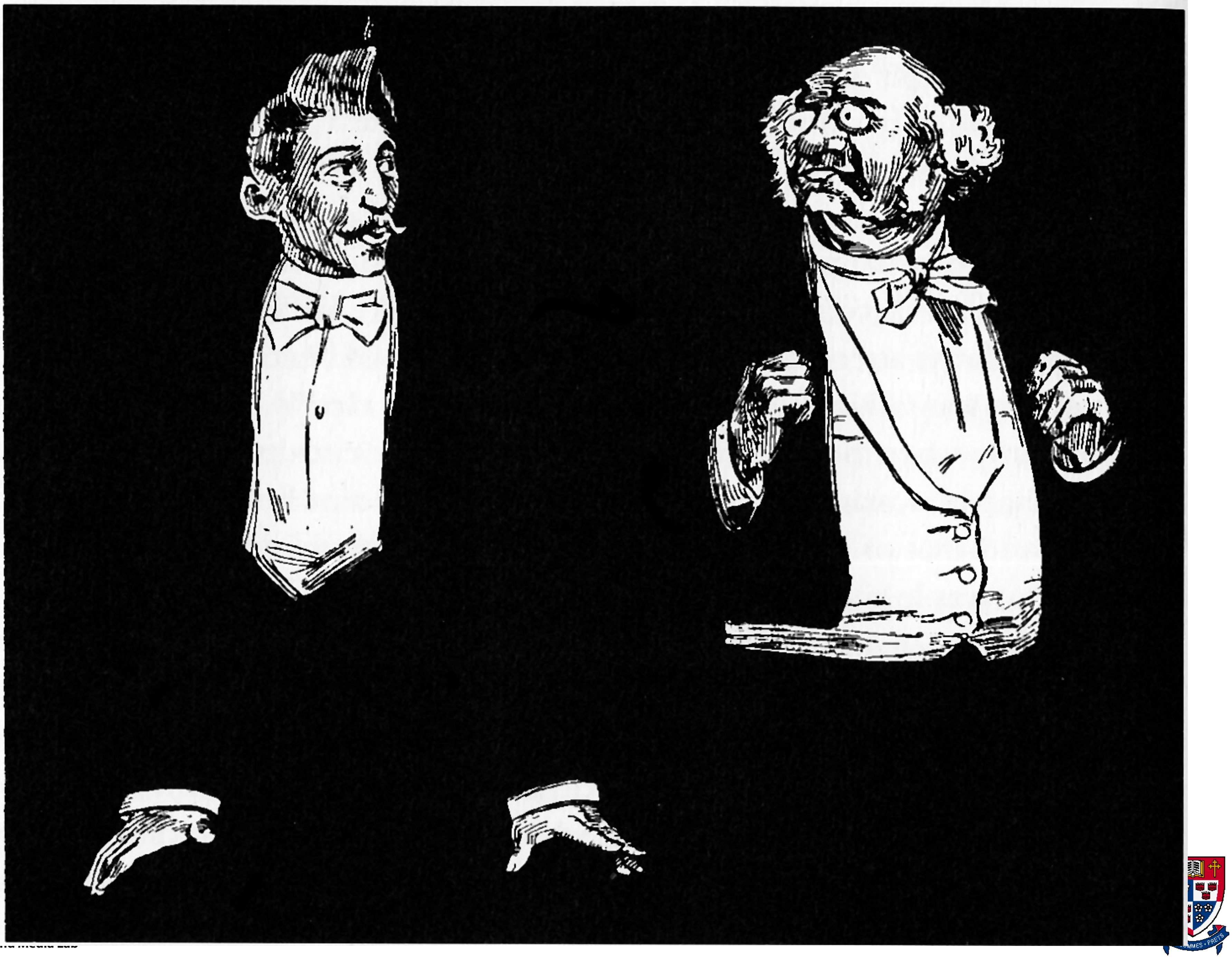
CMPT 888

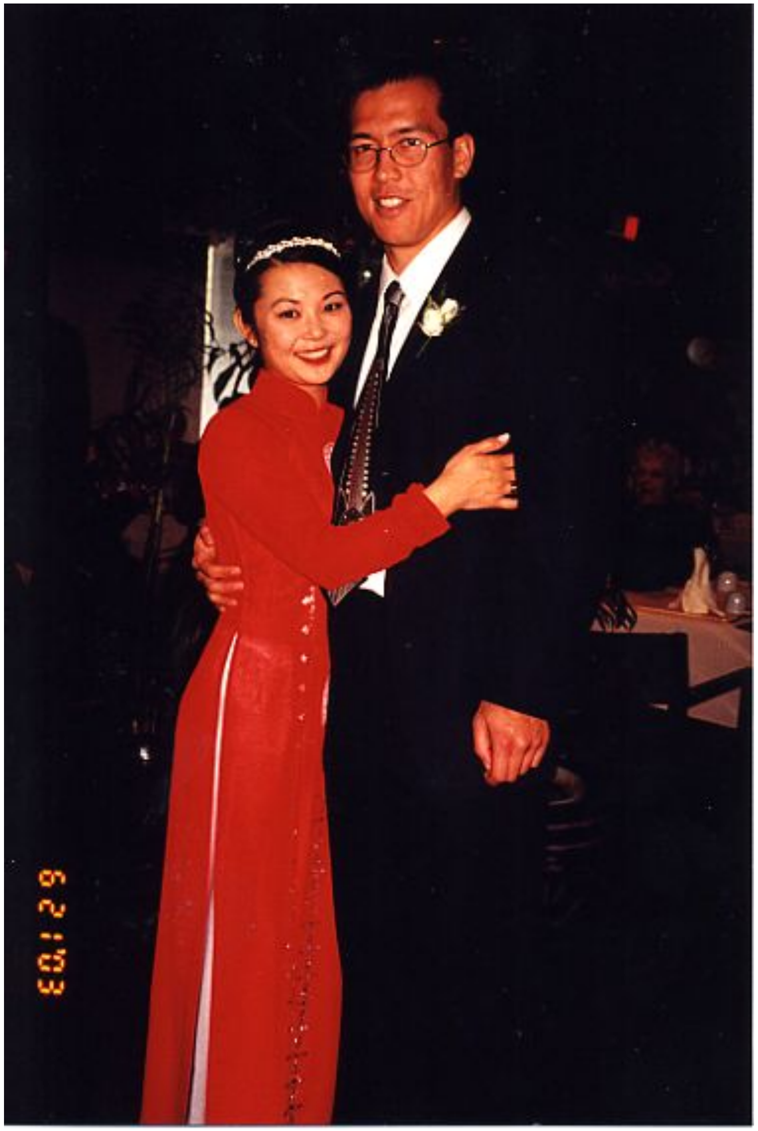
SFU Vision and Media Lab

Problem


ead
L shoulder
R shoulder
L elbow
R elbow
R hand
L hand
L hip
R hip
Waist
R knee
L knee
L foot
R foot



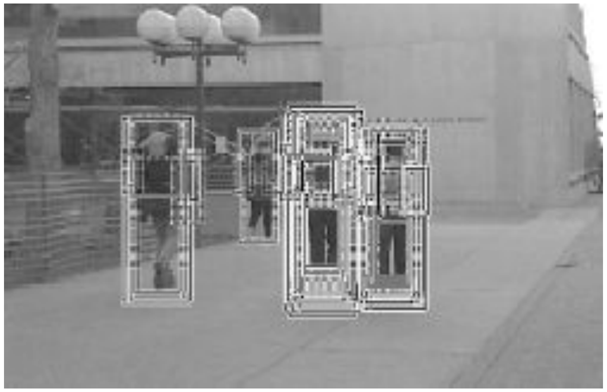




SFU Vision and Media Lab

Human Figures in Still Images

- Detection of humans is possible for stereotypical poses
  - Standing
  - Walking
  - (Viola et al., Dalal & Triggs)
- But we want to do more
  - Wider variety of poses
  - Localize joint positions



SFU Vision and Media Lab

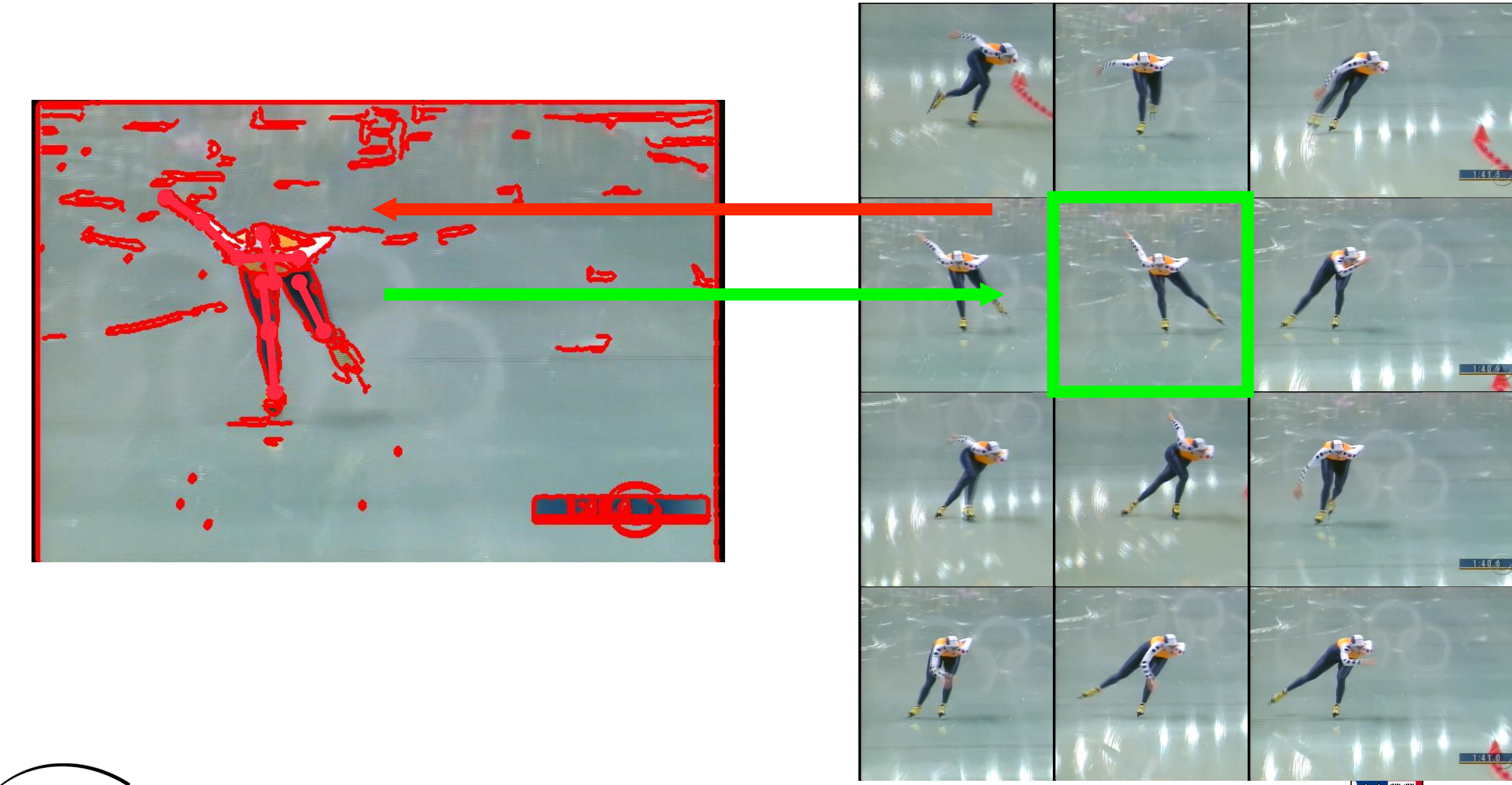
Problem

Models vs. Exemplars

- Two broad classes of approaches
  - – Match templates (exemplar-based)
  - – Fit model

EXEMPLAR METHODS

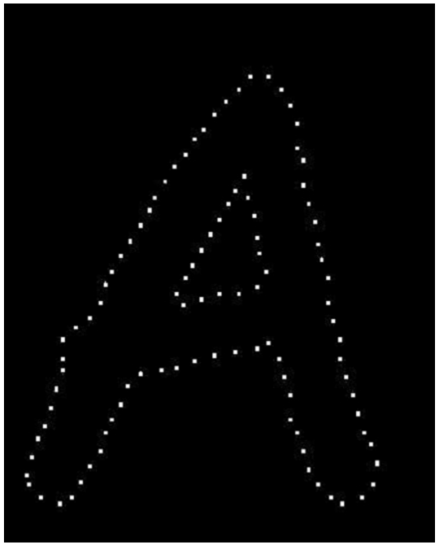
Shape Matching For Finding People



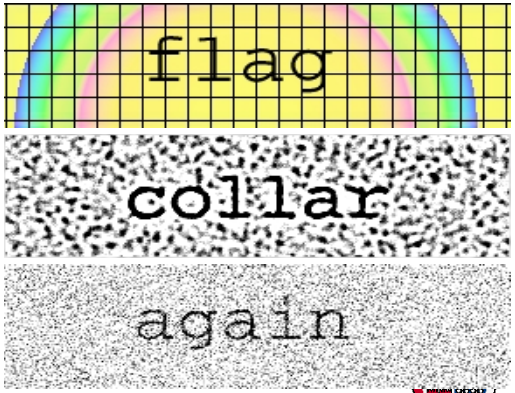
Database of Exemplars

SFU Vision and Media Lab

Shape Contexts

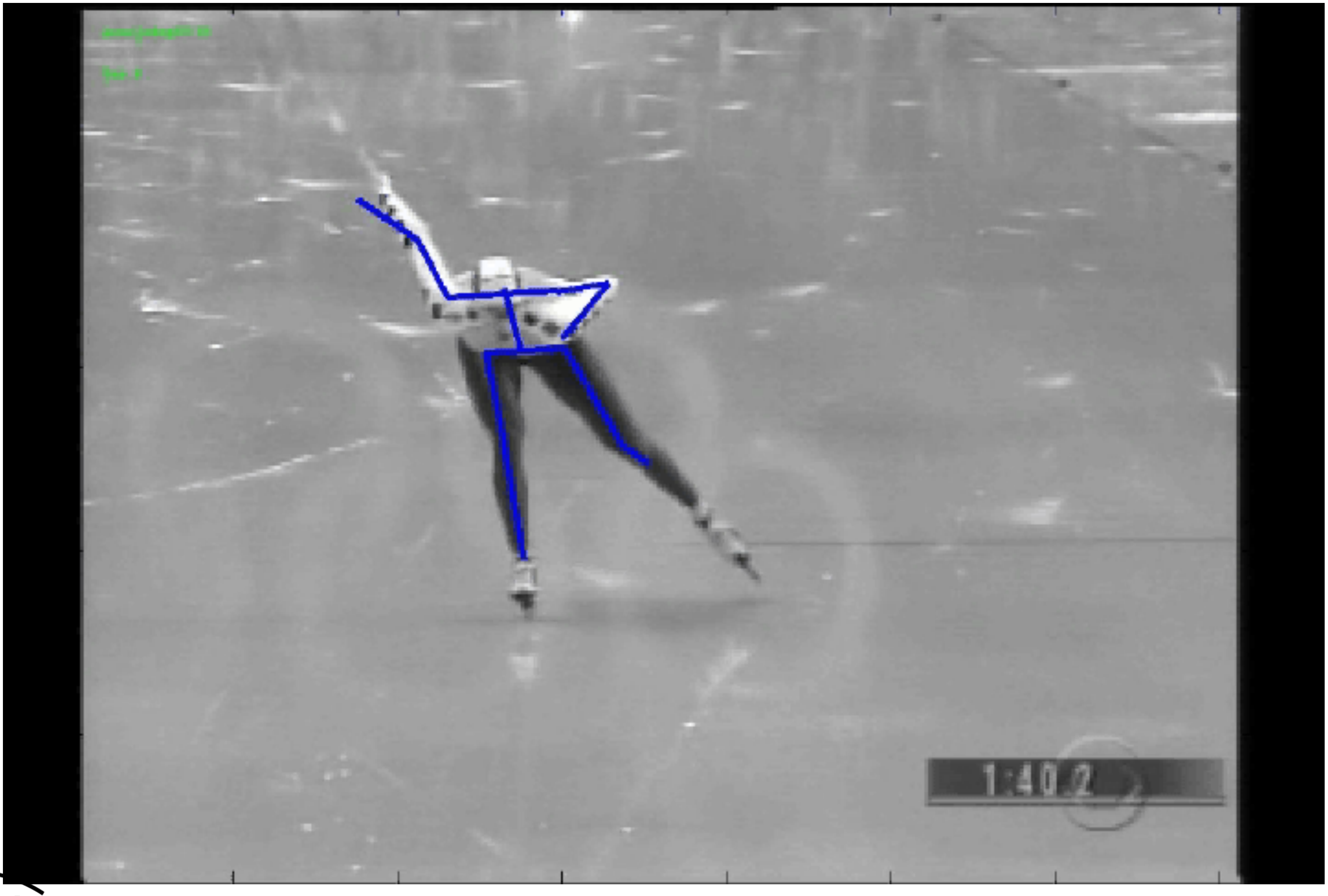


- Deformable template approach
  - Shapes represented as a collection of edge points
- Two stages
  - Fast pruning
    - Quick tests to construct a shortlist of candidate objects
    - Database of known objects could be large
  - Detailed matching
    - Perform computationally expensive comparisons on only the few shapes in the shortlist
- Publications
  - *Mori et al., CVPR 2001*
  - *Mori and Malik, CVPR 2003*
    - Featured in New York Times Science section



SFU Vision and Media Lab

Results: Tracking by Repeated Finding

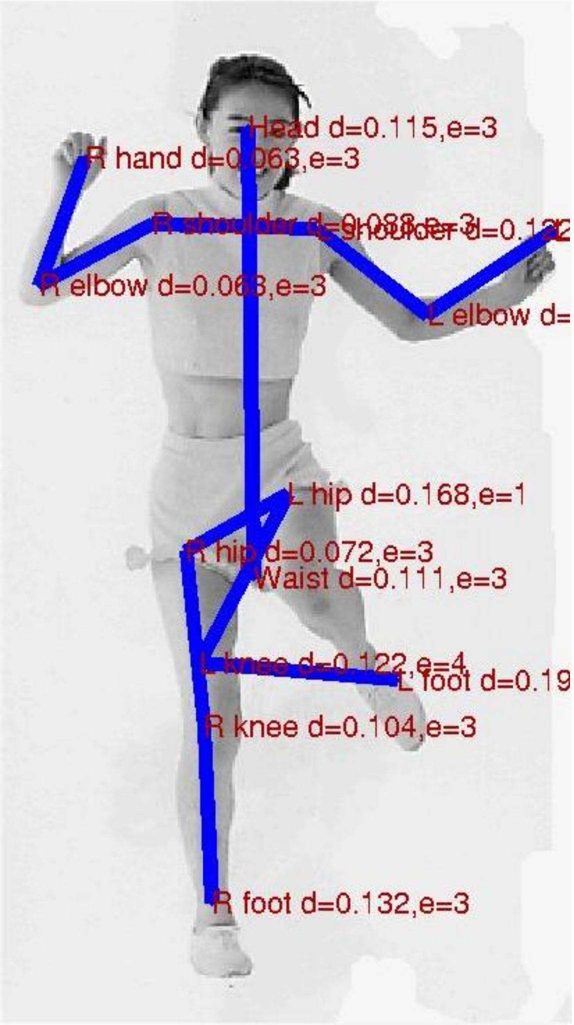


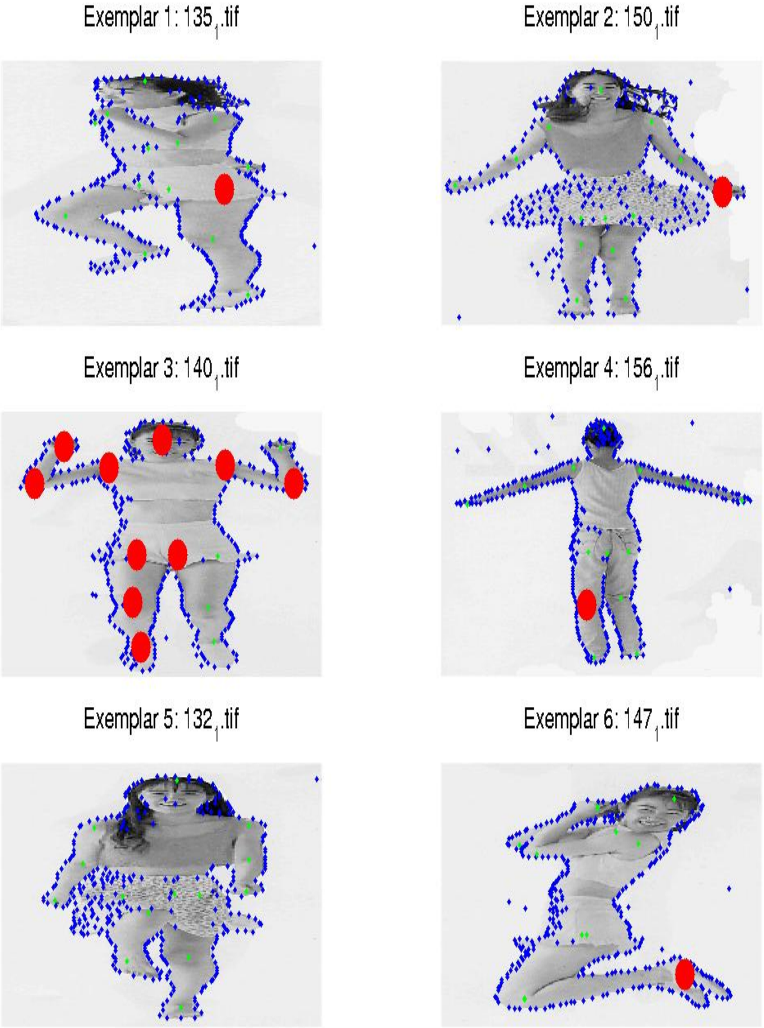
SFU Vision and Media Lab

Multiple Exemplars

- Parts-based approach
  - Use a combination of keypoints or limbs from different exemplars
  - Reduces the number of exemplars needed
- Compute a matching cost for each limb from every exemplar
- Compute pairwise “consistency” costs for neighbouring limbs
- Use dynamic programming to find best K configurations

Combining Exemplars





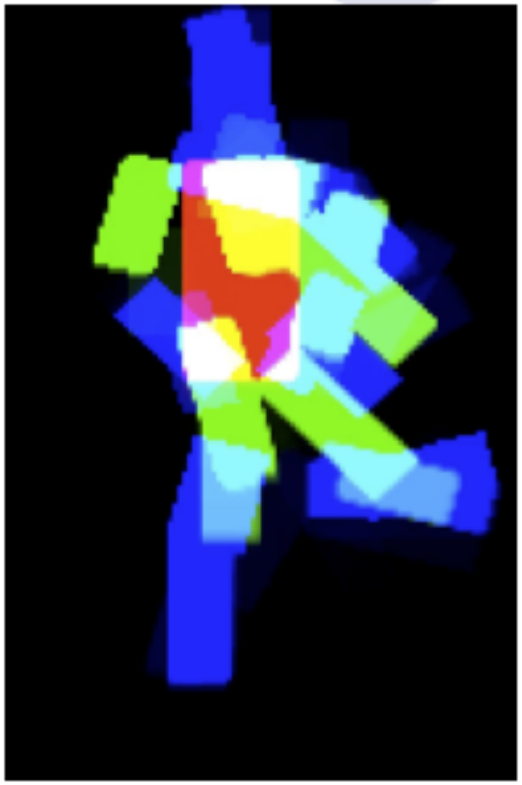
Scaling Up (e.g. Shakhnarovich et al.)

- Methods for automatically generating exemplars
  - Graphics package (e.g. POSER)
- Methods for efficient nearest neighbour search
  - Locality sensitive hashing
  - k-d trees

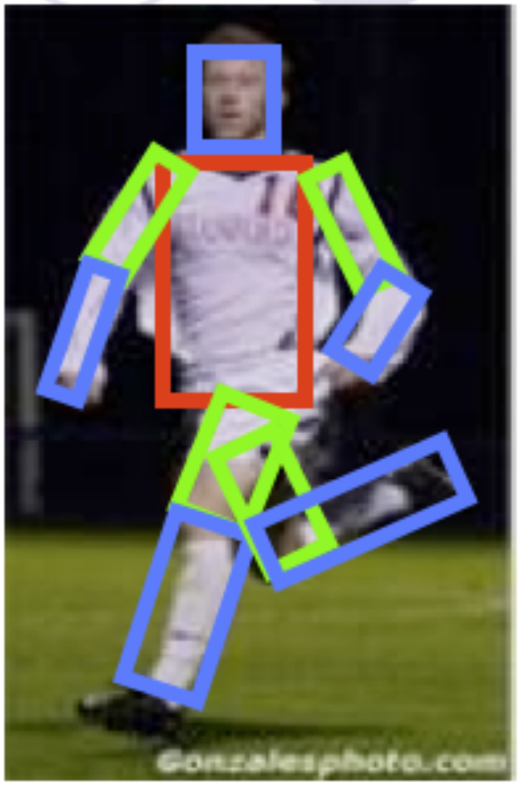
MODEL-BASED METHODS

Problem

Input Image

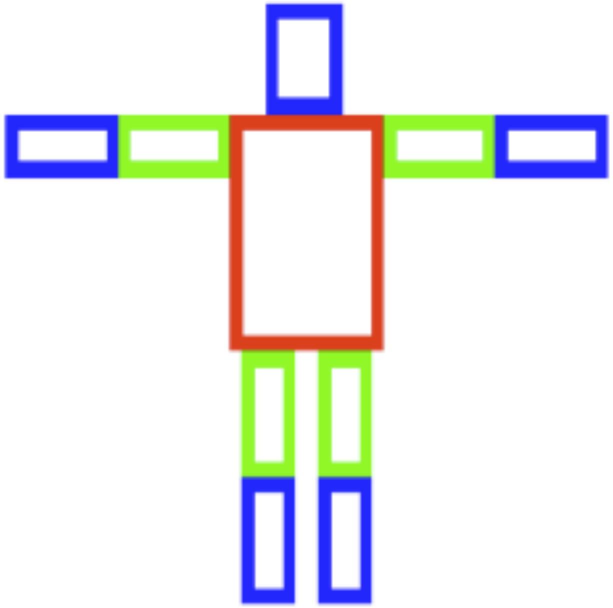


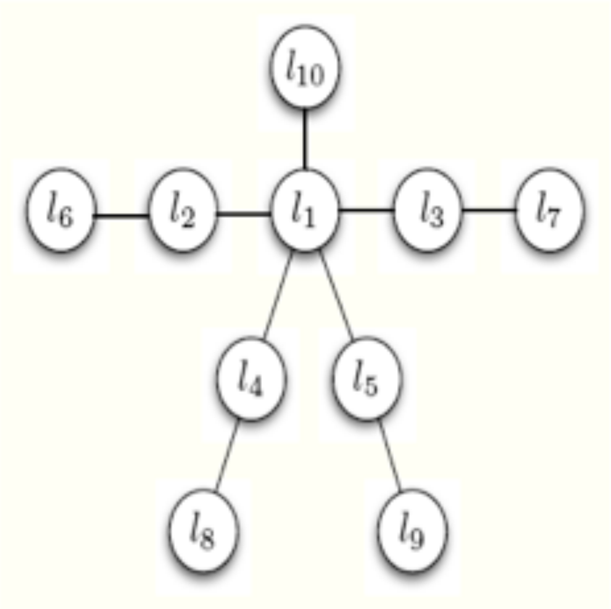
Parts Distribution



Ideal Output

Review of Tree-Structured Deformable Models





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

spatial prior

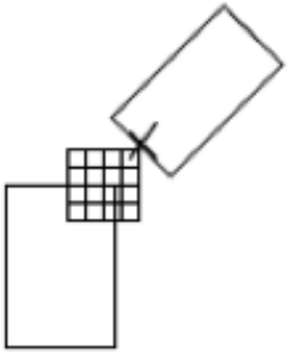
part appearance

Model Parameters (Ramanan,NIPS'06)

- Spatial prior

$$\psi(l_i - l_j) = \alpha_i^T \text{bin}(l_i - l_j)$$

favor certain spatial/angular bins



- Part appearance

$$\phi(l_i) = \beta_i^T f_i(I(l_i))$$

favor certain edge patterns

part-specific binary vector of edges

Learning and Inference

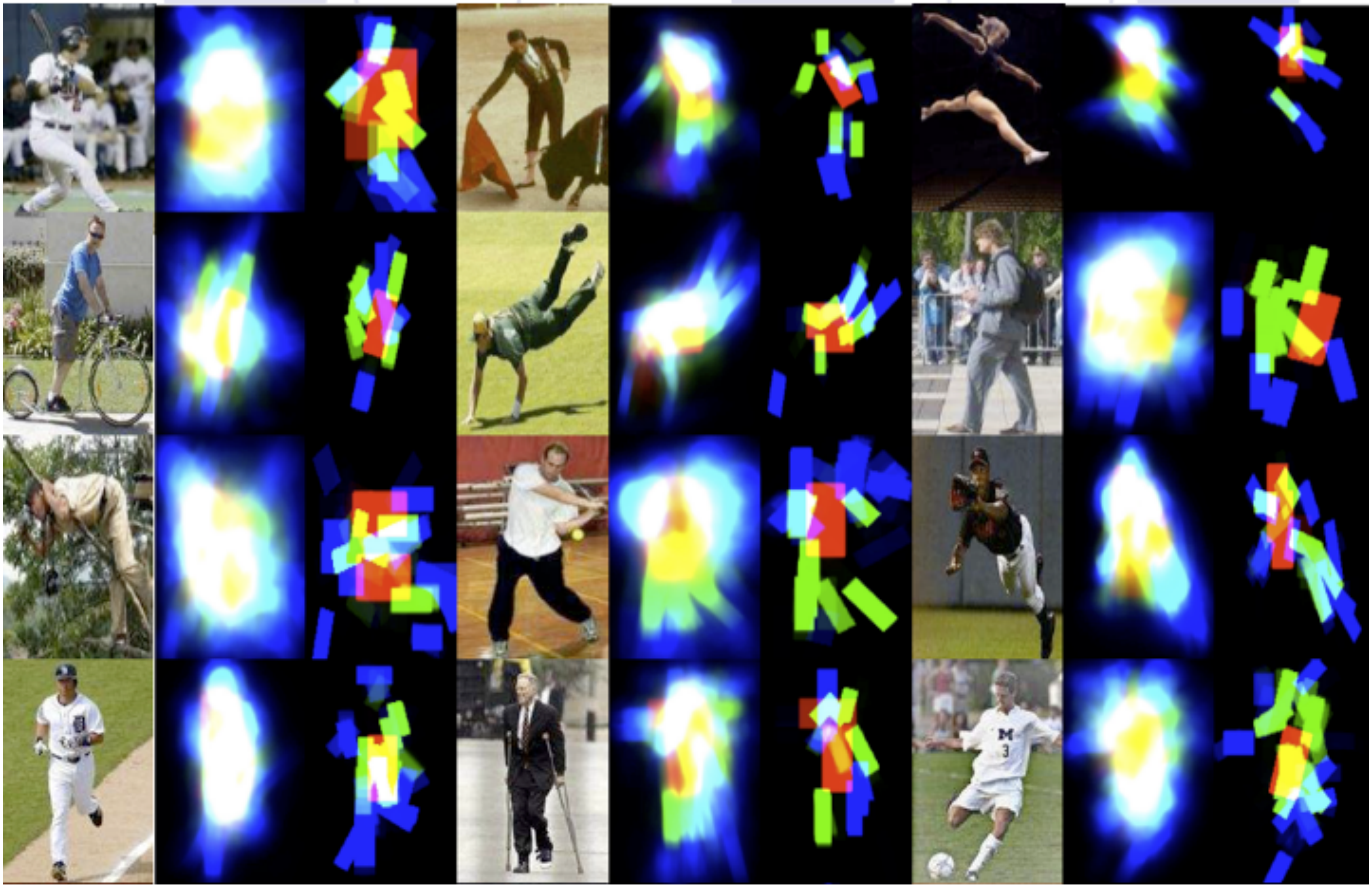
- Inference: message passing with 3D convolution
- Learning Θ_{ML}

$$\Theta_{ML} = \max_{\Theta} \prod_{t} Pr(I^t, L^t | \Theta)$$

- Learning Θ_{CL}

$$\Theta_{CL} = \max_{\Theta} \prod_{t} Pr(L^t | I^t, \Theta)$$

Results

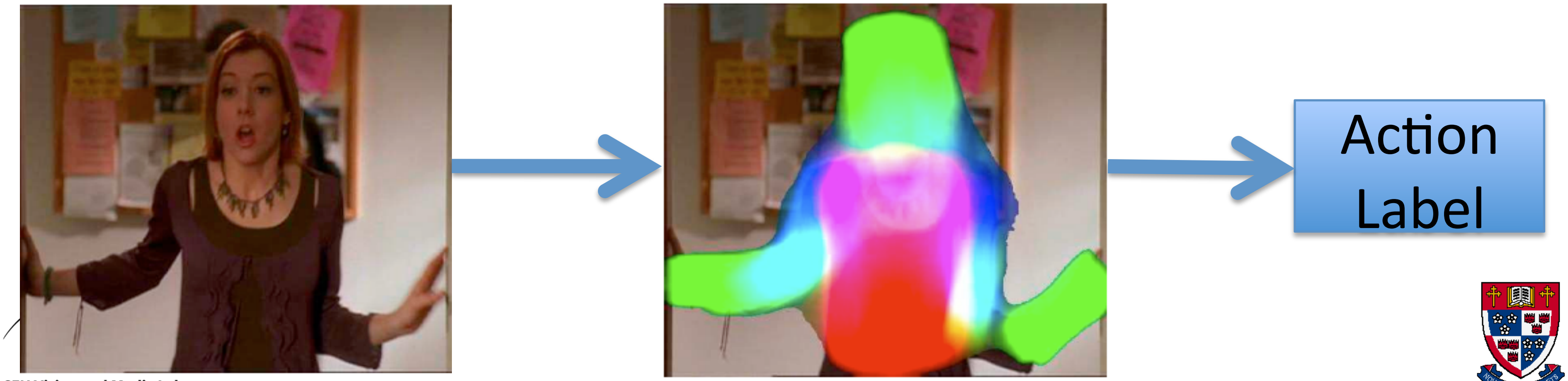


Ferrari, Marin-Jimenez, Zisserman, CVPR 2009

POSE SEARCH

Goal

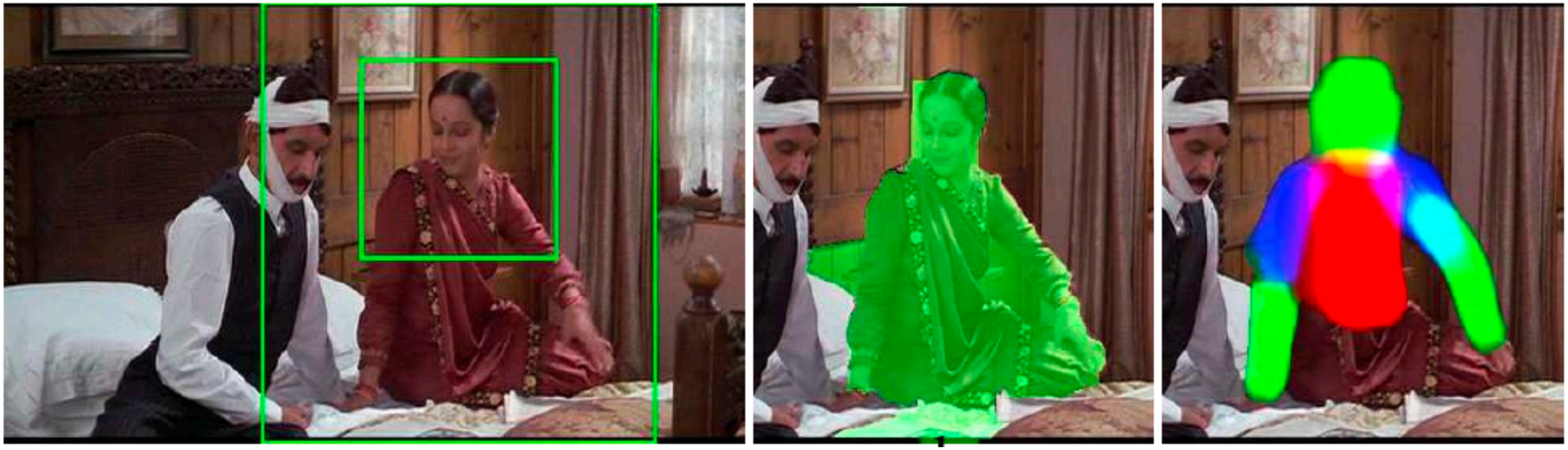
- Video shot retrieval from pose
  - Either *query-by-example* or classification
  - Focus on upper body pose



Action Label

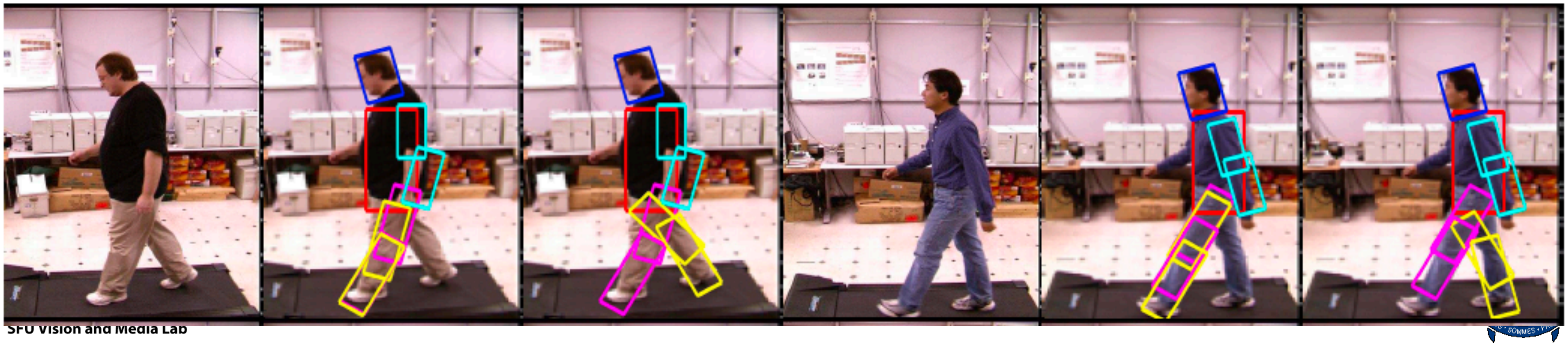
Upper Body Pose Estimation

- Detect upper body (HOG)
- Rough segmentation (GrabCut)
- Pose estimation (Pictorial Structure with Ramanan’s iterative parsing)



Modifications to PS Model

- Prior on pose
  - Uprightness reasonable for TV shows
- Repulsive model
  - Avoid double-counting image evidence



Pose Descriptors

- Pose estimator gives marginals on body parts over time
- Three descriptors are examined:
  - Part positions
    - Discretized absolute part positions/orientations
  - Relative location/orientations
    - Discretized relative part positions/orientations
  - Part segmentations

Pose Comparison

- Bhattacharyya similarity for discrete distributions
- Dot products for segmentations

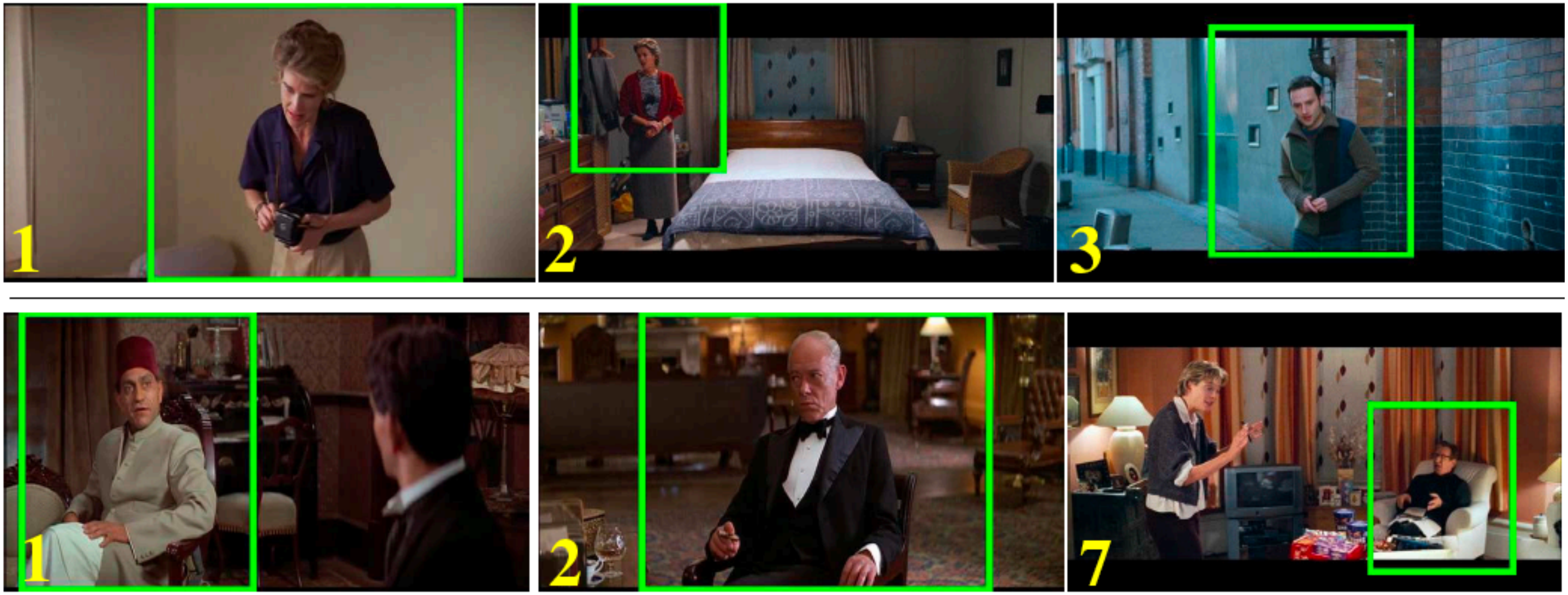
Shot Scores

- How to compare tracks of people?
  - One-to-one
    - Maximum similarity between query pose and track
  - Top-k average
    - As above, but average over best k matches
  - Query interval
    - One-to-one, but allow a max over query sequence too

Classifier Mode

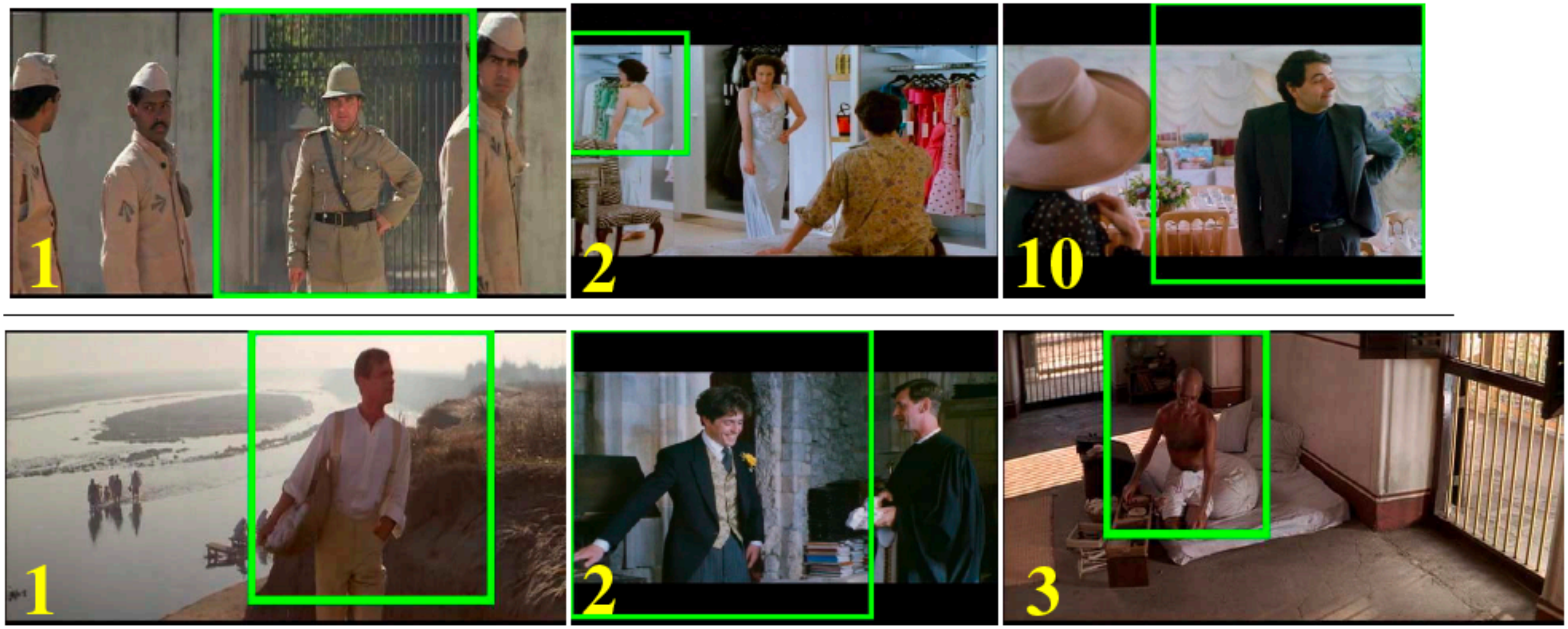
- Train an SVM
  - Useful (standard) tricks about augmenting data

Results



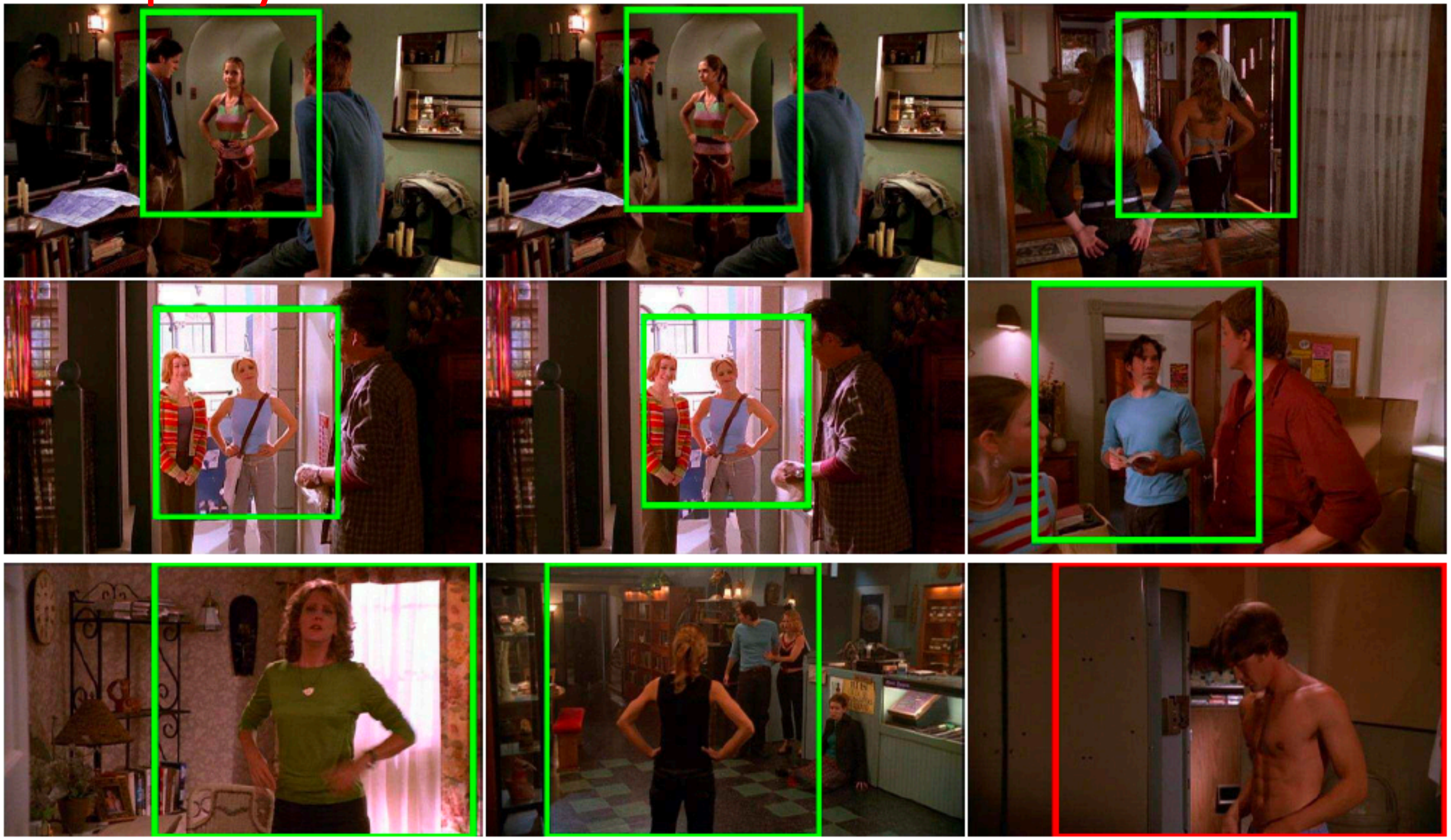
query

Results



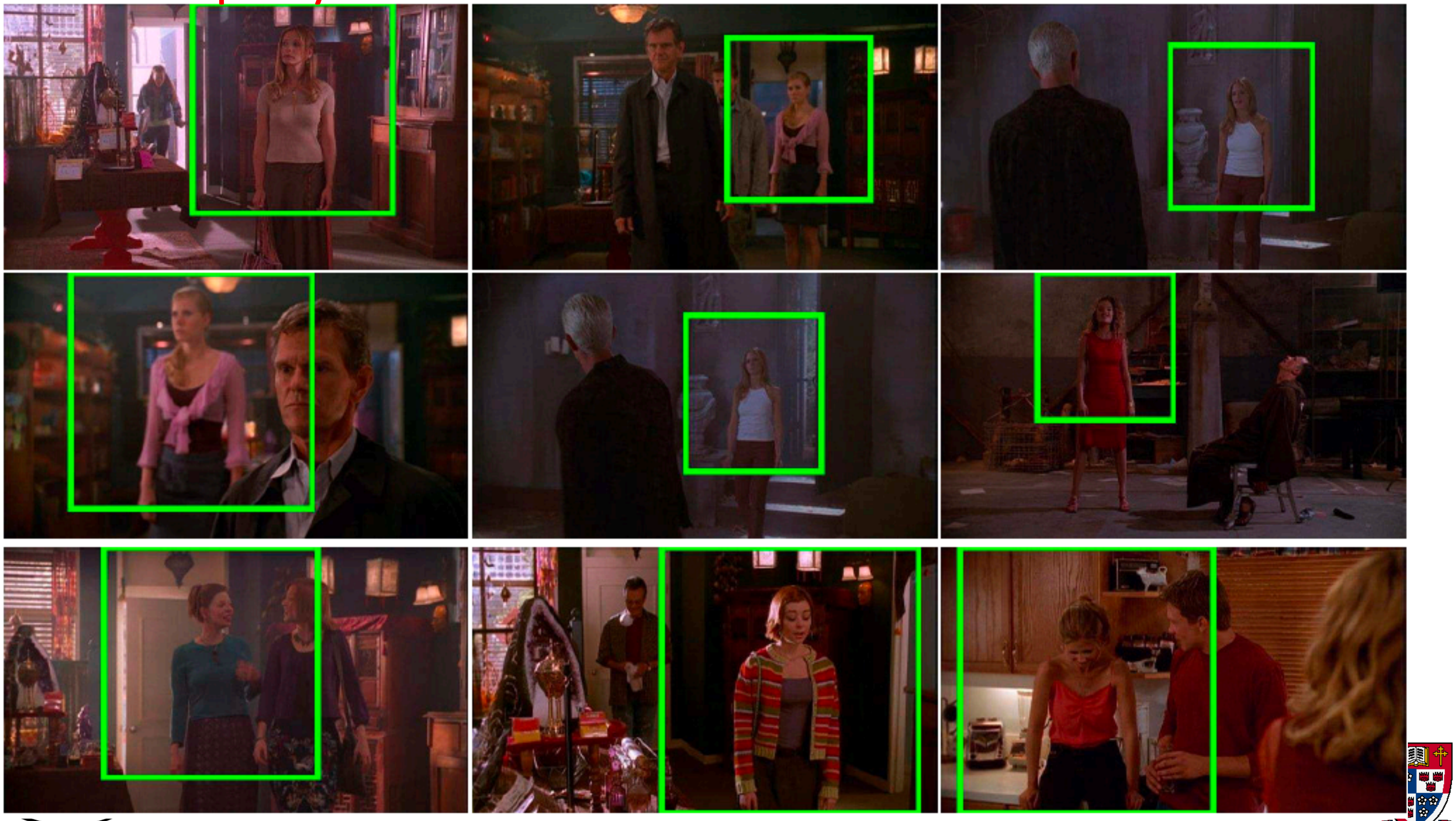
query

Results



query

Results



query

Resources

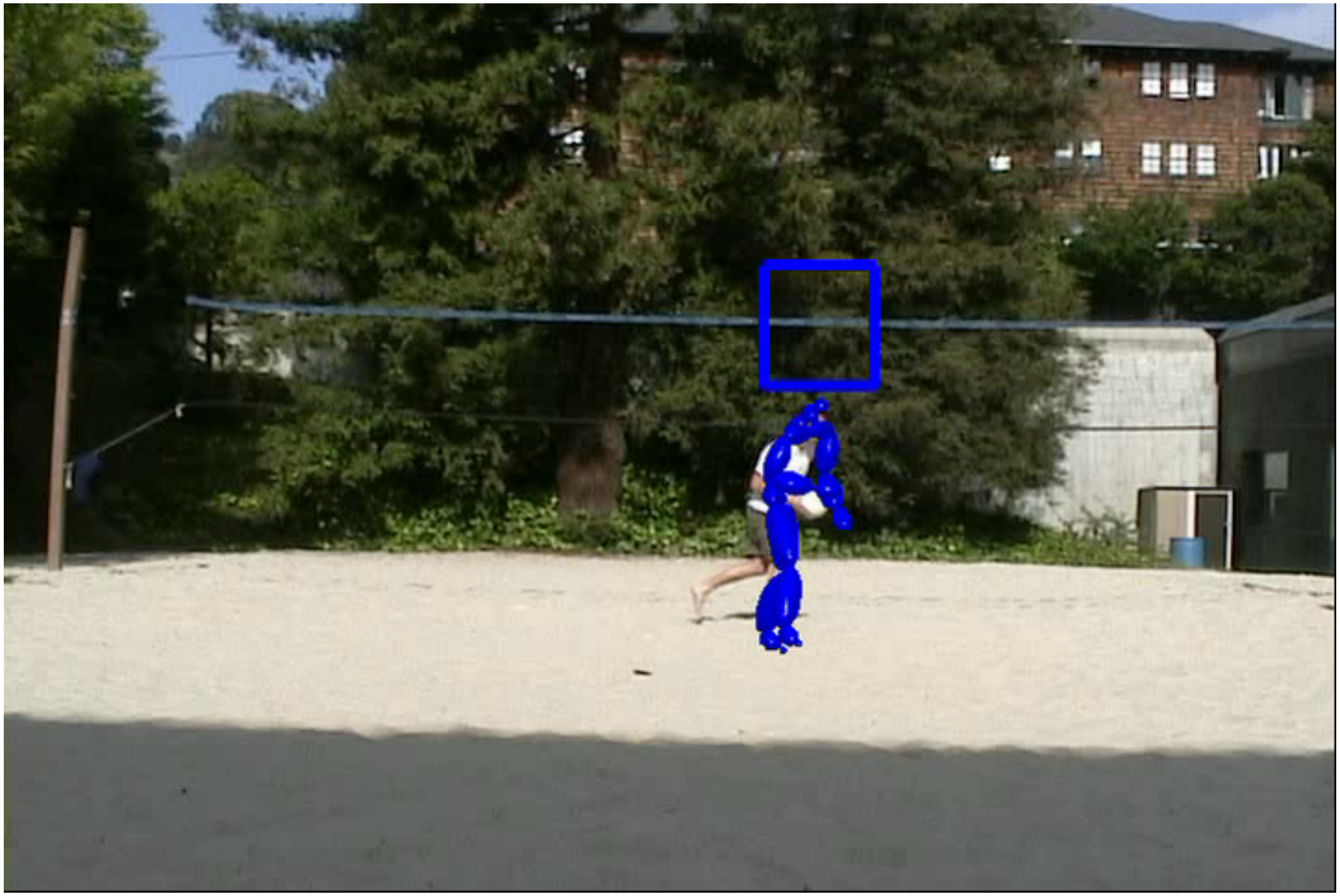
- Code and datasets online

Ramanan and Forsyth NIPS 03

AUTOMATIC ANNOTATION OF EVERYDAY MOVEMENTS

SFU Vision and Media Lab

Goal

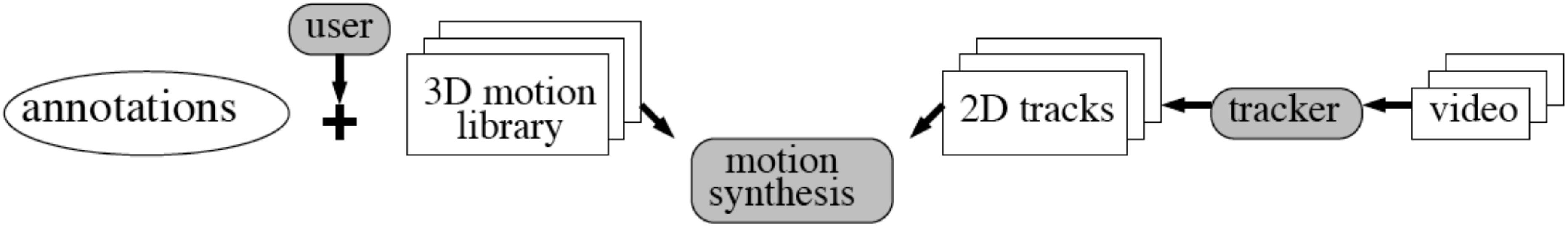


SFU Vision and Media Lab

Representation

- Each frame is labeled with a bit string
  - Each entry denotes presence/absence of an action
  - E.g. run and carry can happen together, both entries would be 1

Approach



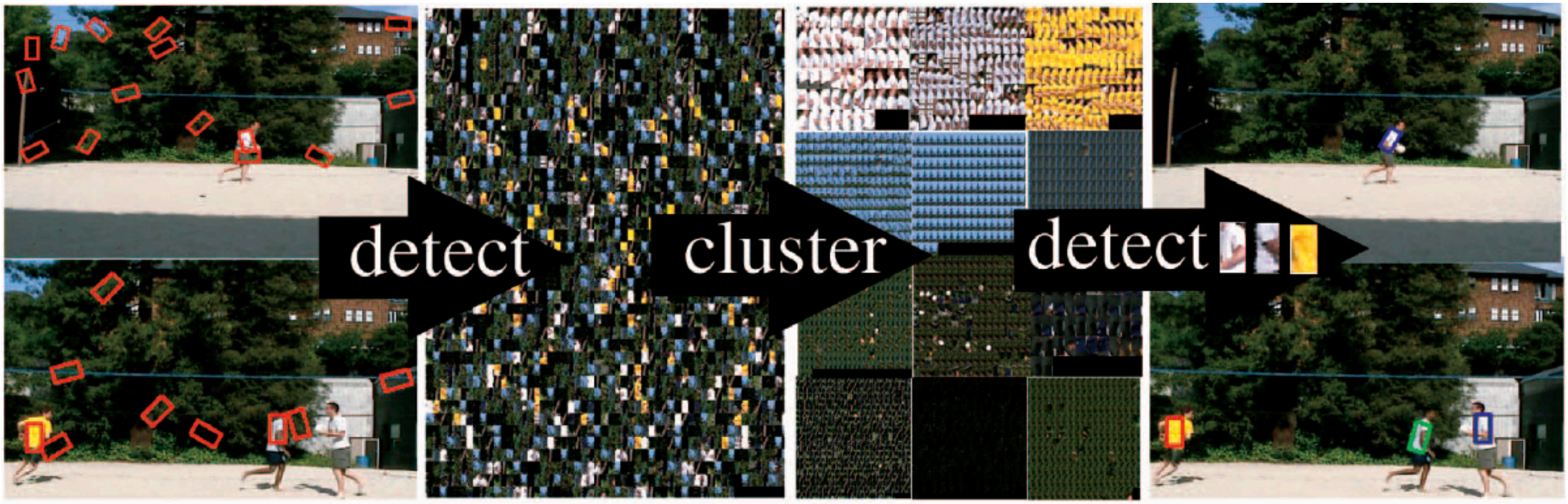
- Start with 3D mocap data
- User annotates data
- Track people in input video
- Compare tracks to mocap data

SFU Vision and Media Lab

Annotations

- 3D mocap data
  - From EA (American) football
- User annotates some frames
- Train SVMs with GRBF kernel on 3D joint positions over 1s as feature
  - One SVM per annotation

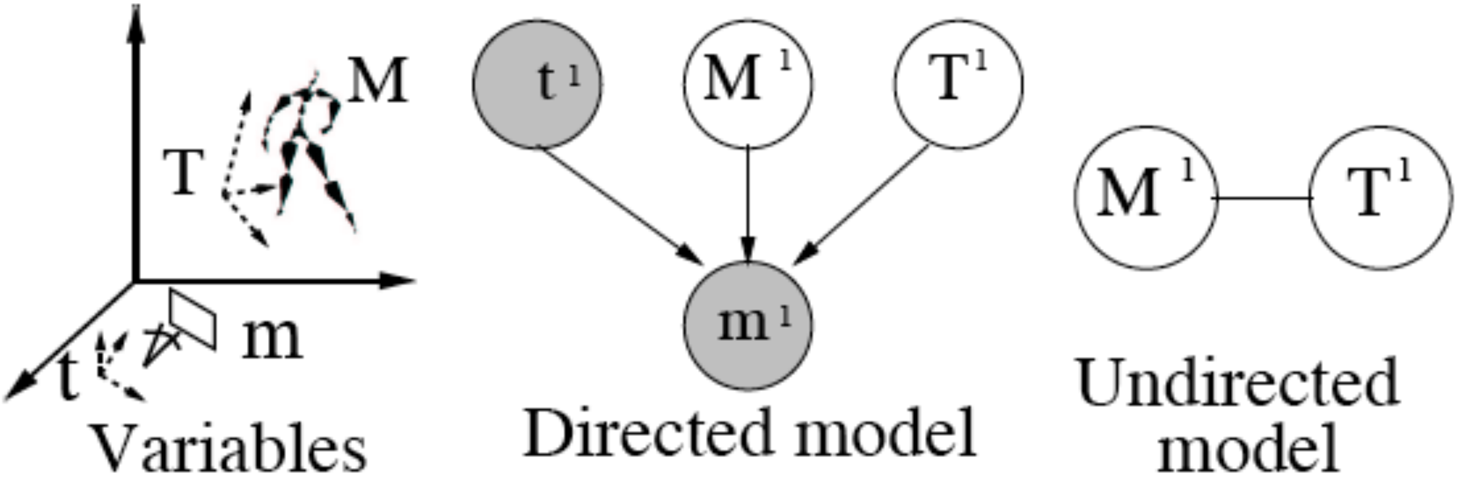
Tracking (CVPR03)



- Detect torsos (rectangles) in video
- Cluster on appearance
- Discard non-moving clusters
- Detect torsos and other parts using pictorial structure model

SFU Vision and Media Lab

Recognition



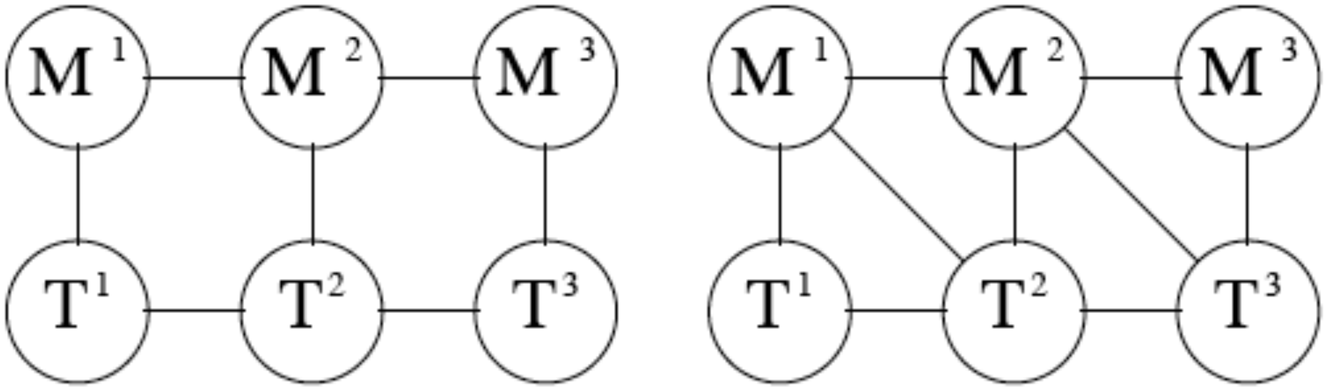
Variables

Directed model

Undirected model

- Discretize 3D poses via k-means clustering (M)
- Assume camera viewing direction parallel to ground plan, torso location known (from tracker)
  - T is simply orientation (direction of torso motion) along ground

Temporal Model I



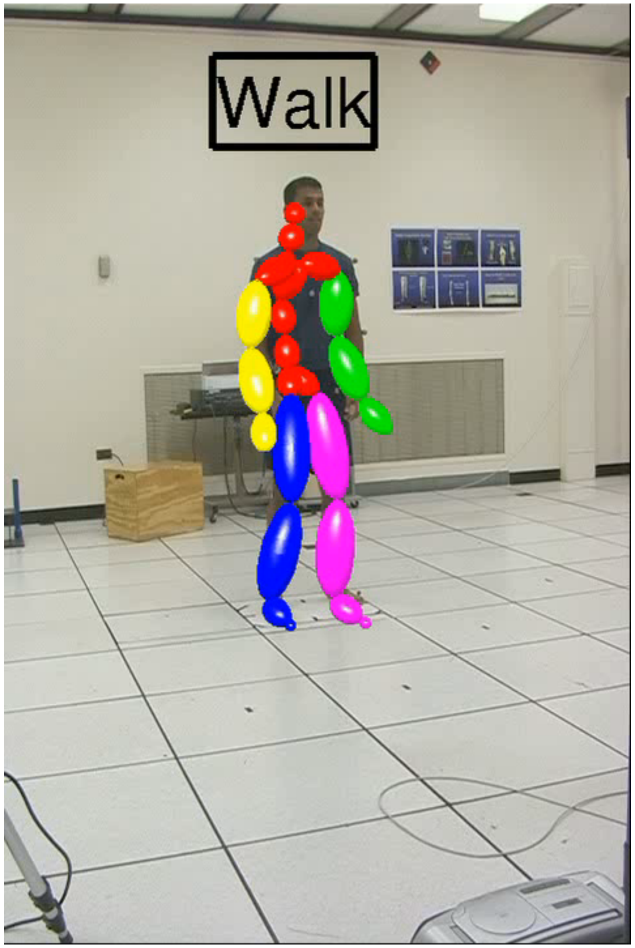
Factorial HMM   Triangulated FHMM

- M-M clique: quantized 3D motion should be smooth
- M-T clique: 3D pose should match 2D pose from tracker
- T-T clique: torso orientation change should be smooth
  - M-T-T: modulate by motion type (some motions can be faster than others

Annotations

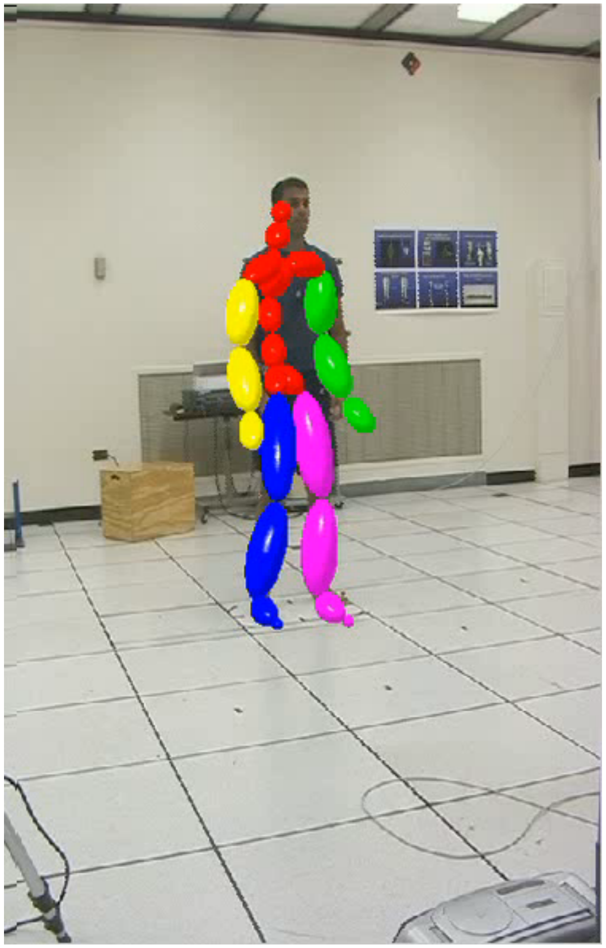
- Use inferred M to give annotation to a frame
  - Various types of hacks possible
    - Medoid (cluster center) annotation
    - Mode of annotations in cluster
    - Annotation of best match in cluster
    - Frequency of annotations (soft annotation)
  - A smoothing approach based on another temporal model (HMM) is used instead

Results

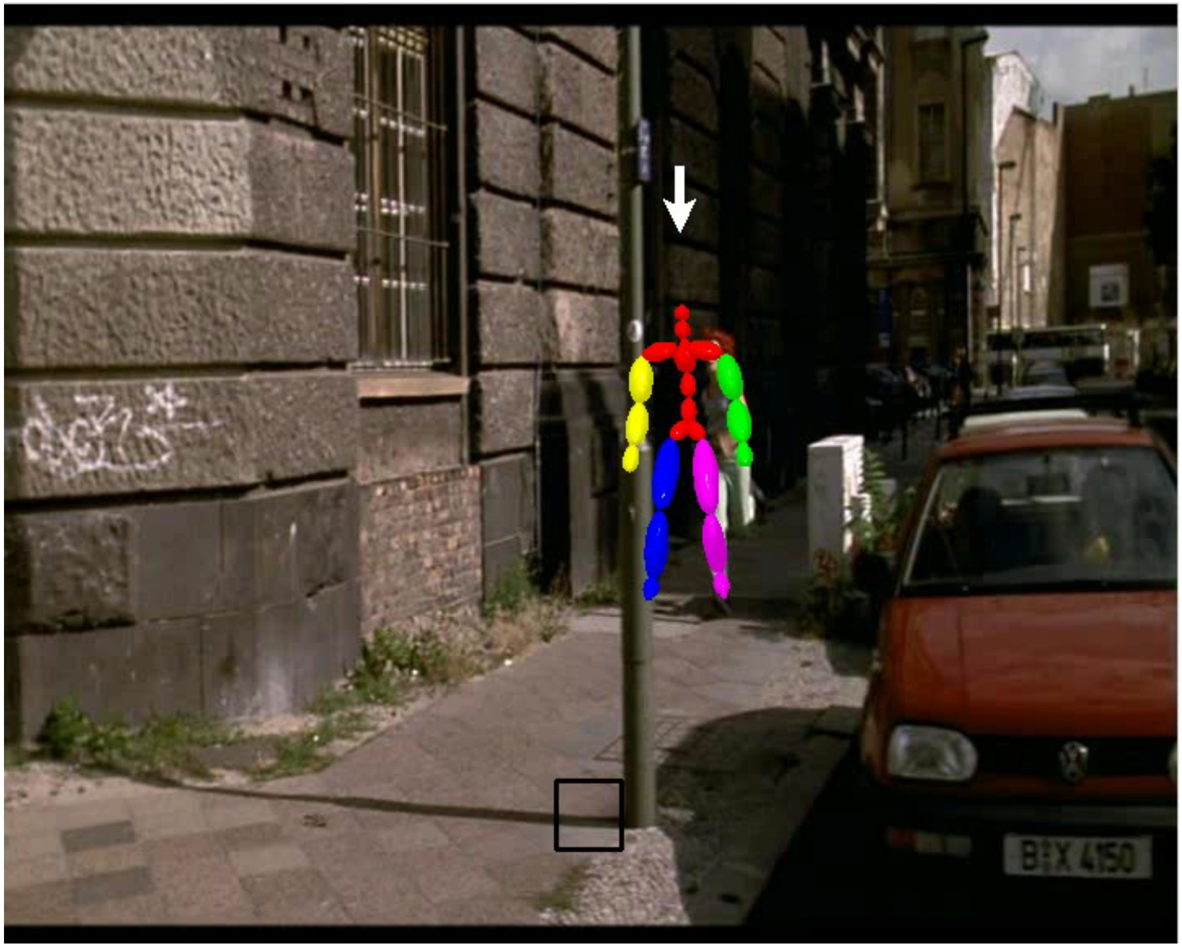


SFU Vision and Media Lab

Results



Results



SFU Vision and Media Lab

Results