Learning Probabilistic Models for Visual Motion

David Ross

Ph.D. Thesis

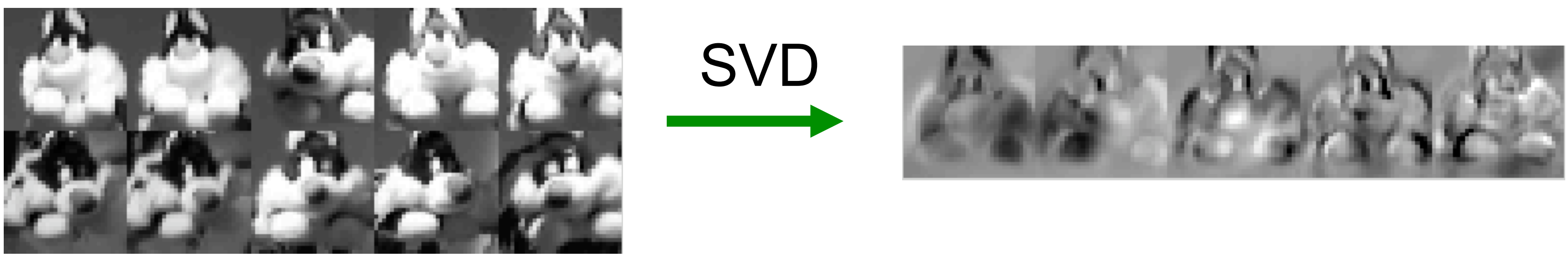
May 23, 2008

Learning for Motion Analysis

- State of the art: human ability is rivaled only in narrow domains, by carefully engineere d systems
  - Tracking a face in video: difficulty with changes in lighting, pose, & occlusions
  - Recovering 3D pose of a human: usually requires detailed kinematic model of human body
- Manual construction limits flexibility & coverage of vision systems
- Machine learning
- Three methods: 2 address training of vis ual trackers, 1 analysis of non-rigid articulated motion

1. Incremental Learning for Visual Tracking

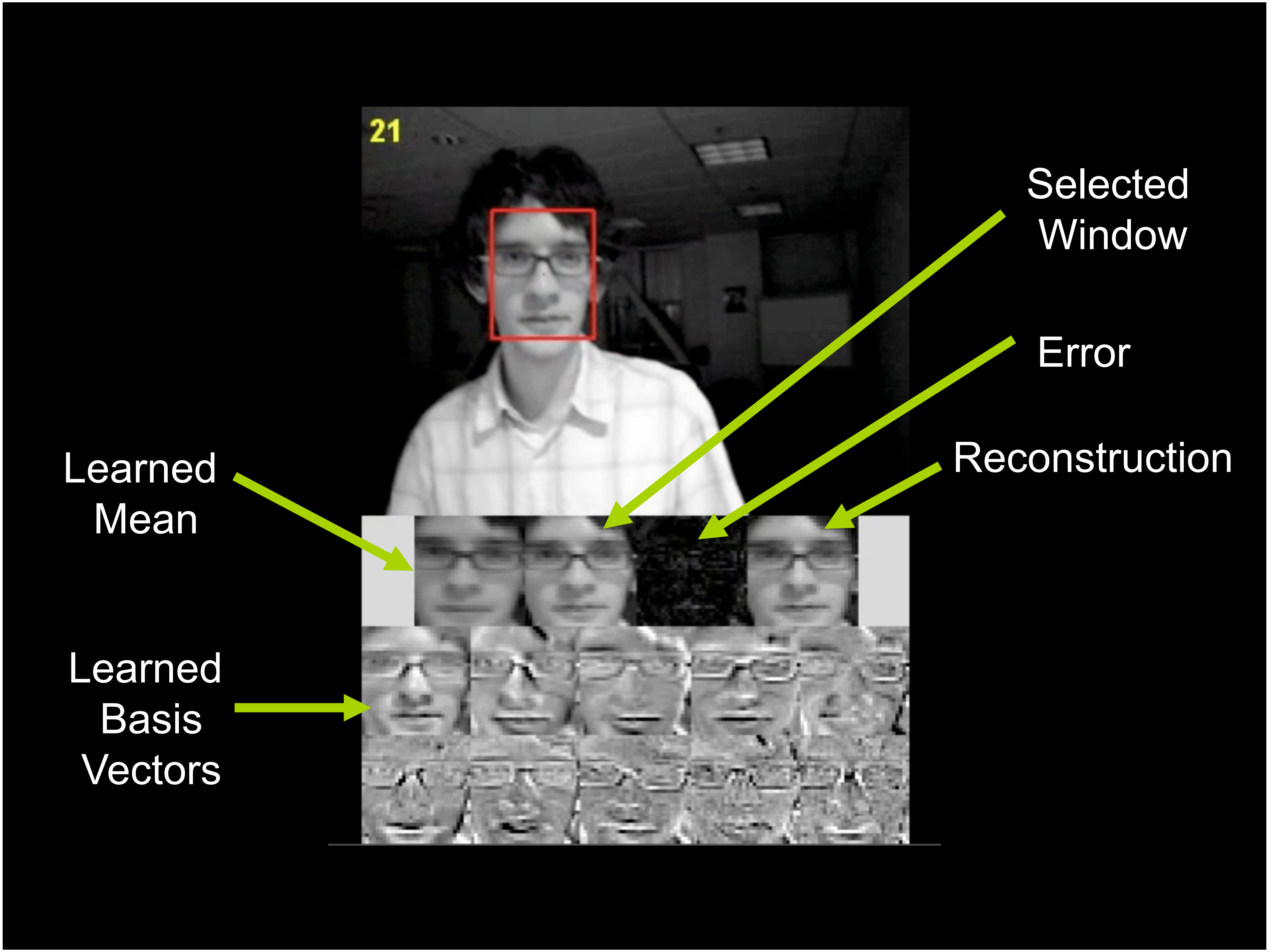
- Tracking requires models of appearance & dynamics
- Principal Components Analysis (PCA) aka Eigentracking

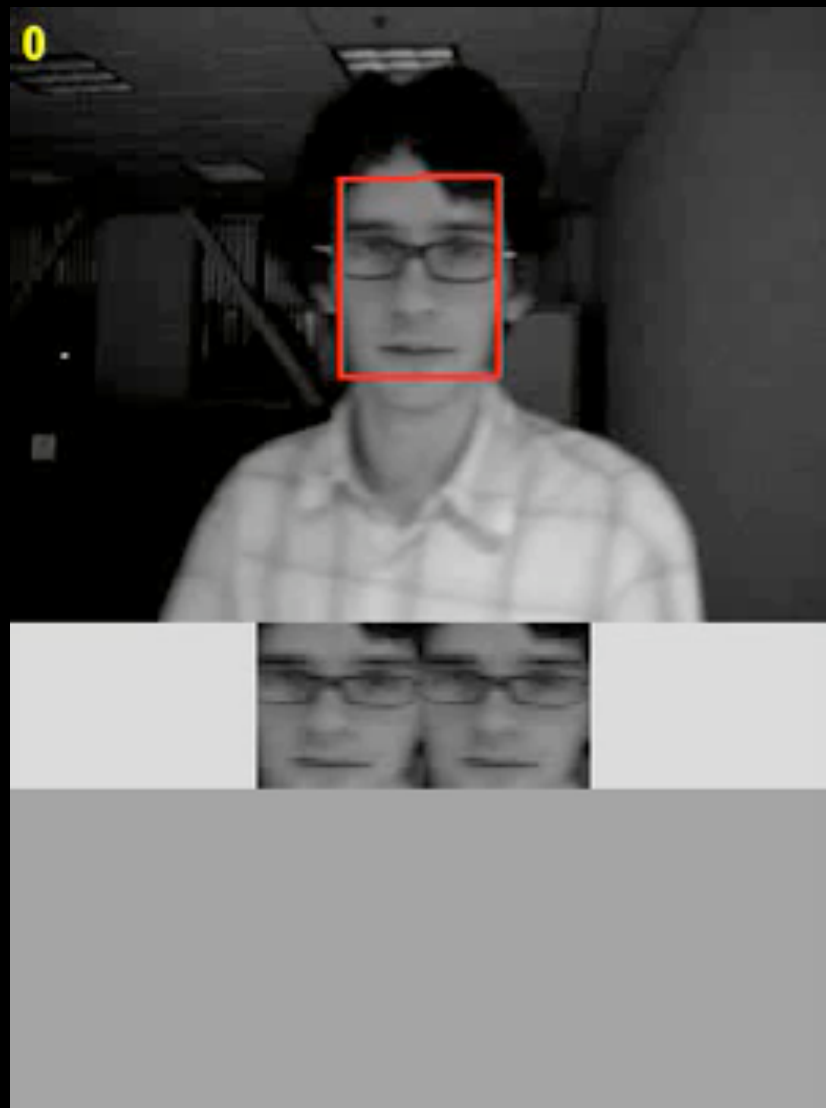


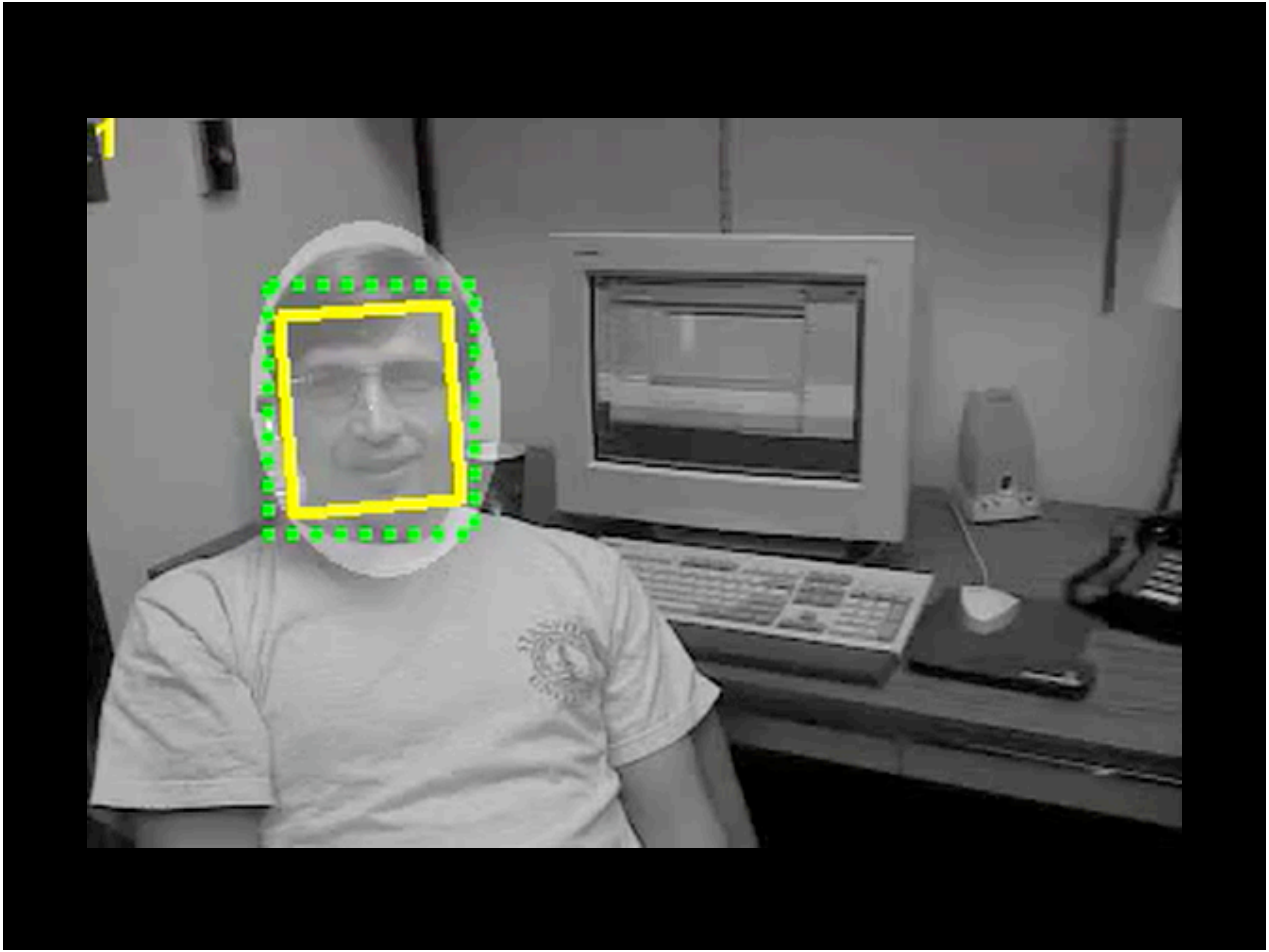
- Drawbacks: requires training data, can't adapt to appearance changes

Incremental PCA

- Incrementally learn PCA model online, from appearances obtained during tracking
- Naïve batch fitting via SVD does not scale
- New algorithm for incremental updates:
  - Fast: constant time updates, constant storage
  - Exact: same result as batch update
  - Correctly updates subspace mean
  - "Forgetting Factor" places more emphasis on recent appearances, improves performance


21
Selected Window
Error
Reconstruction
Learned Mean
Learned Basis Vectors


0


1

2. Combining Discriminative Features

- Previously: Learn PCA
  → not the only appearance model available
- Given a new tracking task, how to select most-appropriate appearance & dynamics models?
- Data driven approach: Learn selection of models + parameters from labeled video sequence.
- Flexible Combination: aggregate tracker more reliable than constituent models

Discriminative Conditional Model

- Model Pr(state|observations) as a (log-linear) combination of dynamics & appearance features
- "Pile on" features
  - Include any features that might be relevant
  - Decide which are useful via learned weights
  - Switch features on and off dynamically

Discriminative Features

- Dynamics features: $f_j(\mathbf{x}_{t-1}, \mathbf{x}_t)$
  – How well do two states match?
  – Li near dynamical models (fly, hold, roll, bounce)
- Observation features: $g_k(\mathbf{x}_t, \mathbf{Y}, t)$
  – Is the target at x_t?
  – Can include information from the entire observation sequence
  – PCA, templates, background subtraction
- Robustify by switching features on and off
  – Hidden switch variables u_{jt} v_{kt}

Features & Switch Potentials

- Weighted distance between state and prediction

$$f_j(\mathbf{x}_{t-1}, \mathbf{x}_t) = -\frac{1}{2}\left(\mathbf{x}_t - \phi_j(\mathbf{x}_{t-1})\right)^T \boldsymbol{\alpha}_j \left(\mathbf{x}_t - \phi_j(\mathbf{x}_{t-1})\right)$$

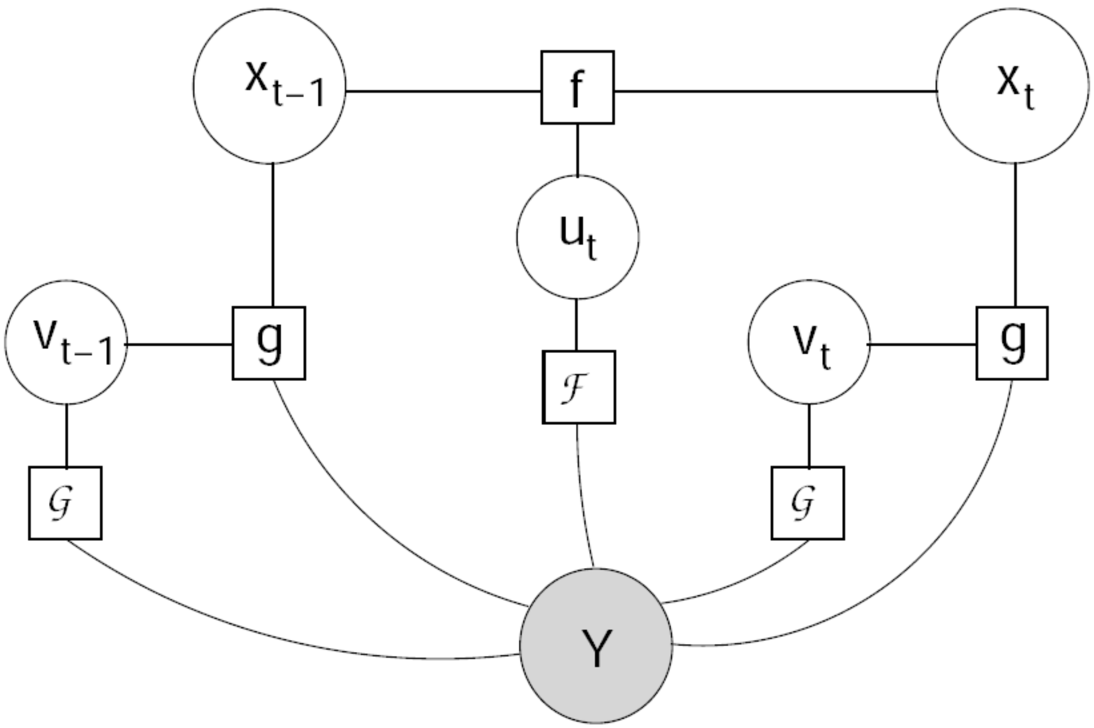
$$g_k(\mathbf{x}_t, \mathbf{Y}) = -\frac{1}{2}\left(\mathbf{x}_t - \gamma_k(\mathbf{Y}, t)\right)^T \boldsymbol{\beta}_k \left(\mathbf{x}_t - \gamma_k(\mathbf{Y}, t)\right)$$

$$\phi_j(\mathbf{x}_{t-1}) = \mathbf{T}_j \mathbf{x}_{t-1} + \mathbf{d}_j$$

- Switch Potentials: extra features help decide if switches should be on or off
- Any classifier (logistic / softmax regression)

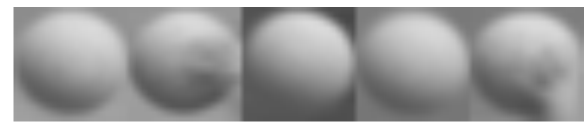
Probability Model

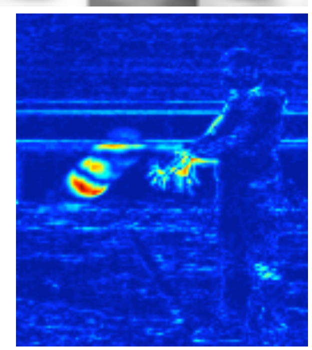
$$\mathrm{P}(\mathbf{X}|\mathbf{Y}) \propto \exp\left(\sum_{t,j} f_j(\mathbf{x}_{t-1}, \mathbf{x}_t) u_{j,t} + \sum_{t,k} g_k(\mathbf{x}_t, \mathbf{Y}) v_{k,t} + \sum_{t,j} \mathcal{F}_j(\mathbf{Y}, t) u_{j,t} + \sum_{t,k} \mathcal{G}_k(\mathbf{Y}, t) v_{k,t}\right)$$



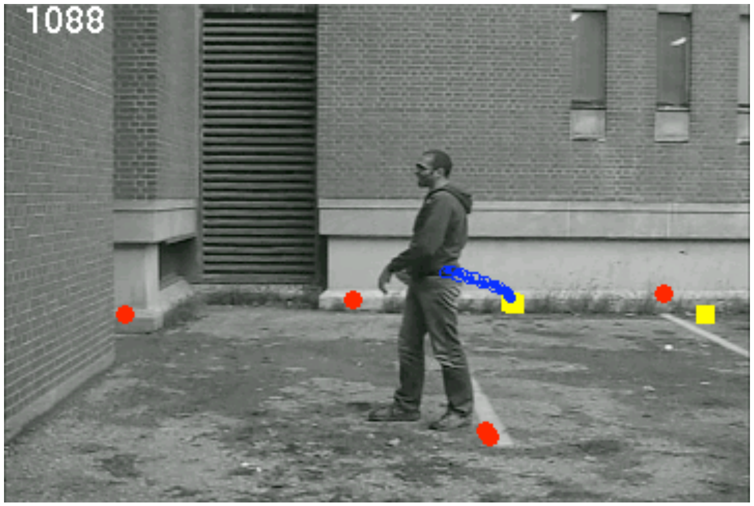
Features Used

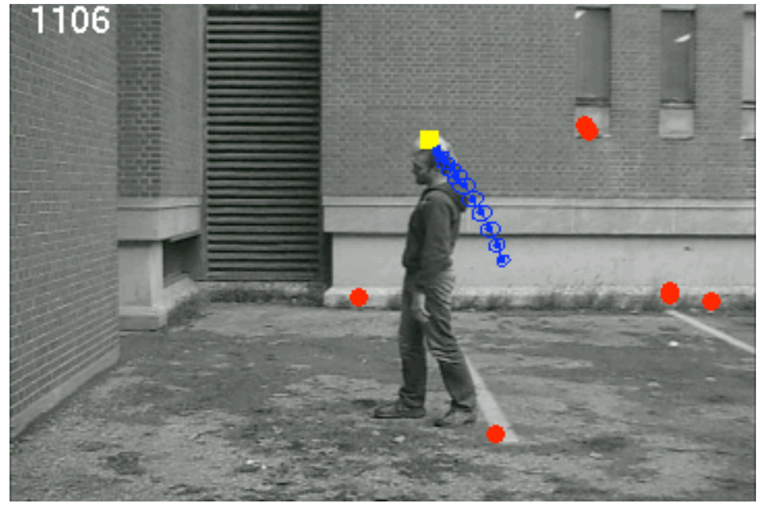
- Template from one frame
- Template from K-Means
- PCA, 3 components
- Local background subtraction
- 4 Linear dynamics (fly, hold, bounce:ground, bounce:wall)

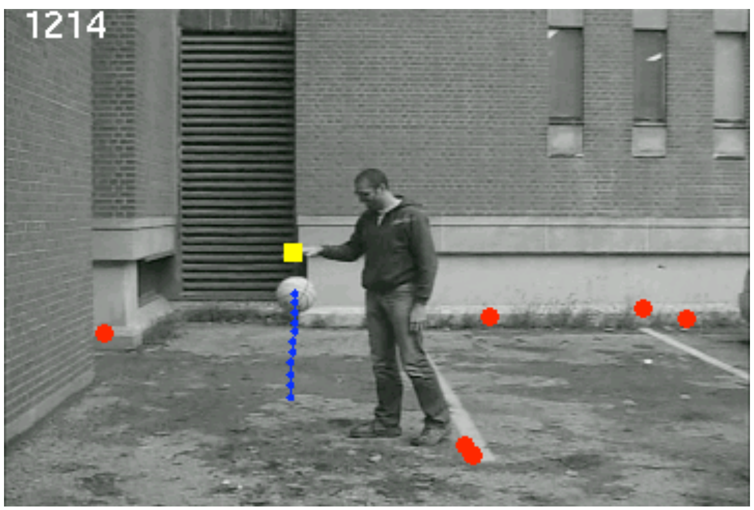




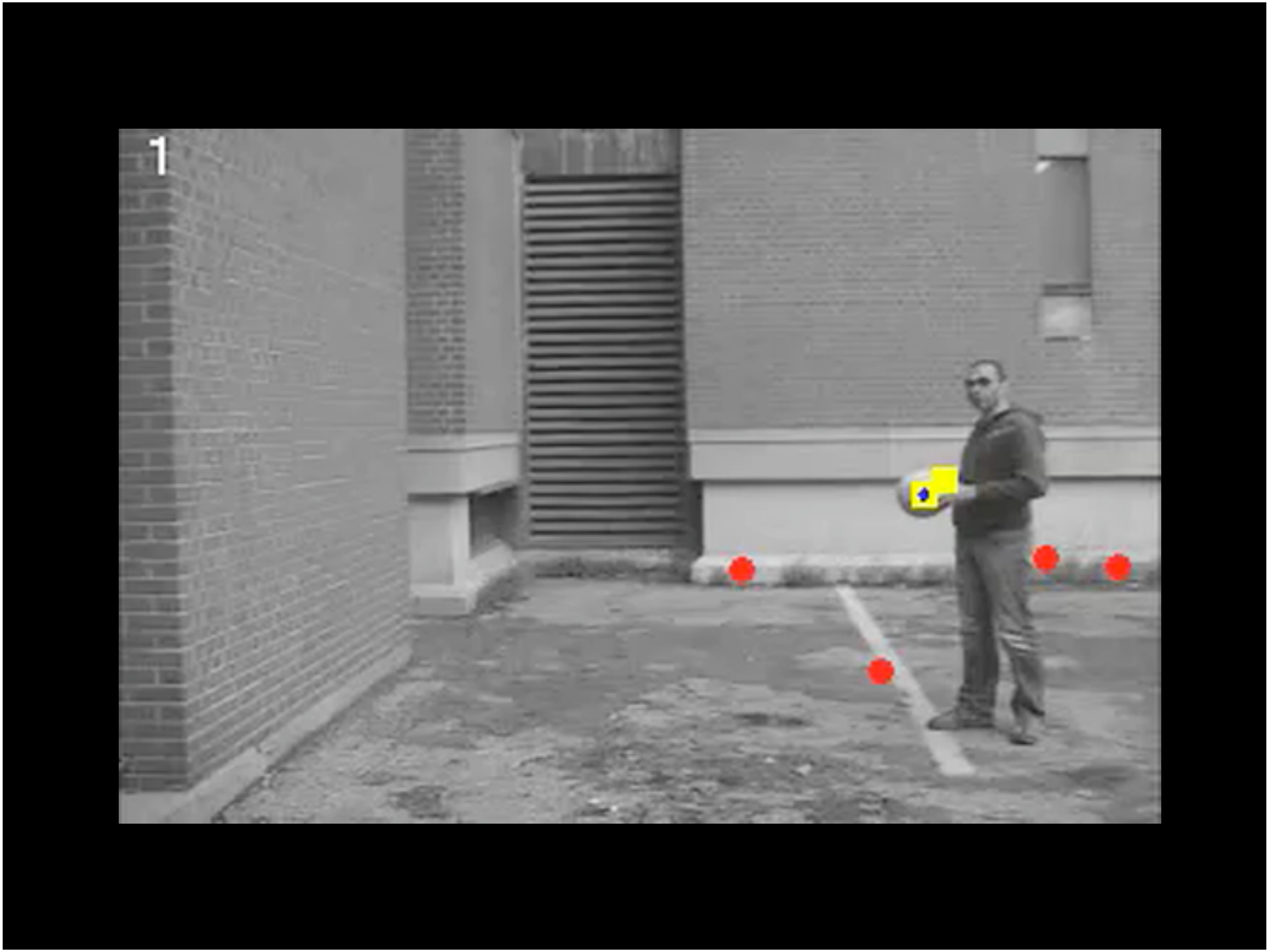
Tracking Performance

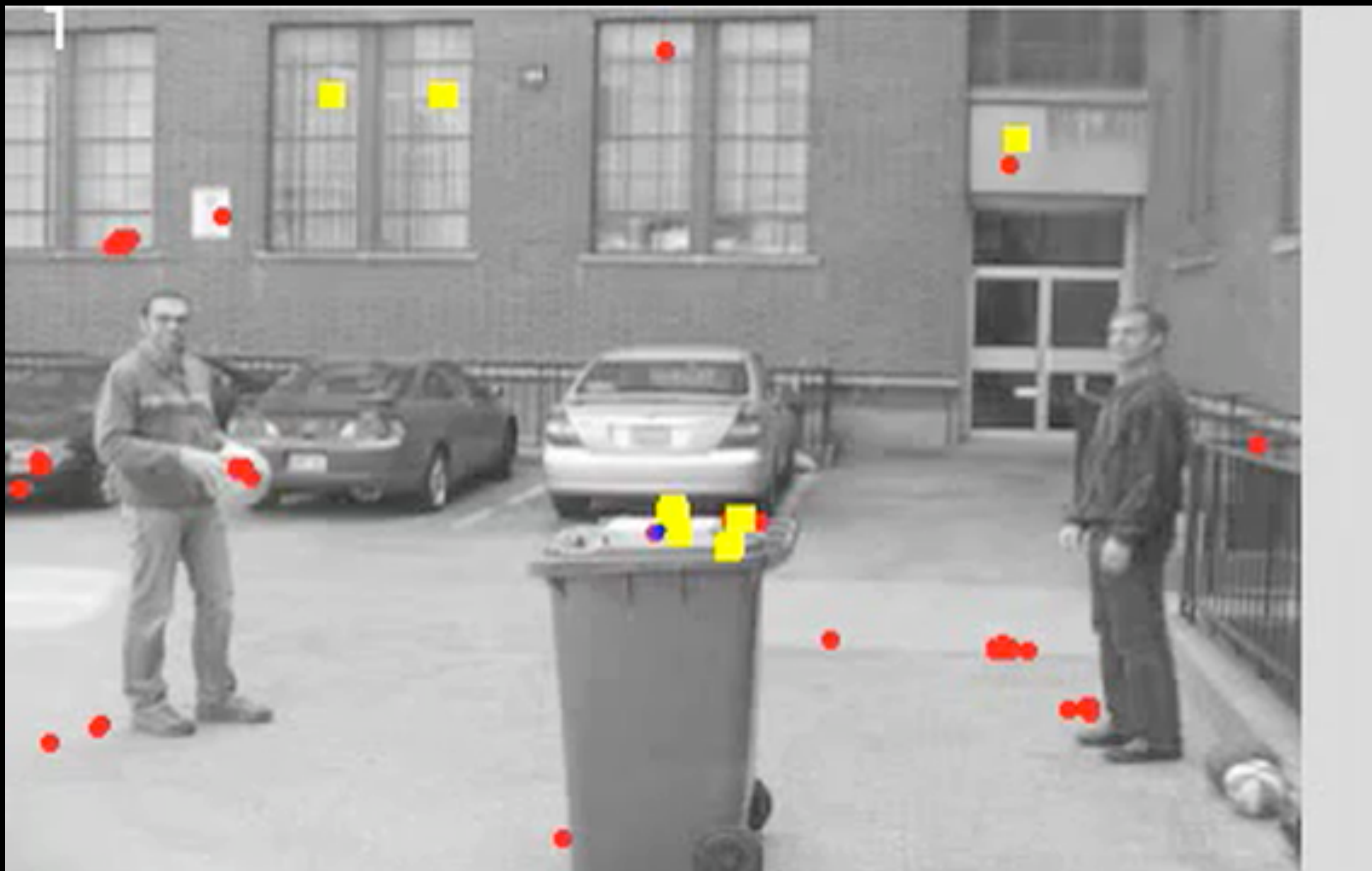






- expected ball position (last 10 frames)
- observation features
  - switched on
  - switched off

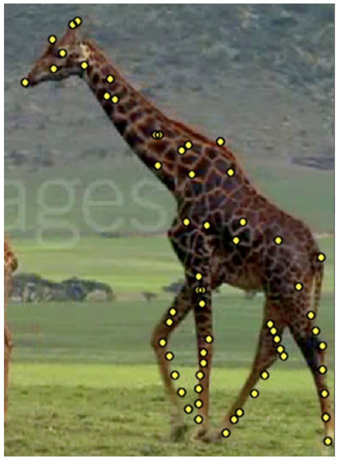

1



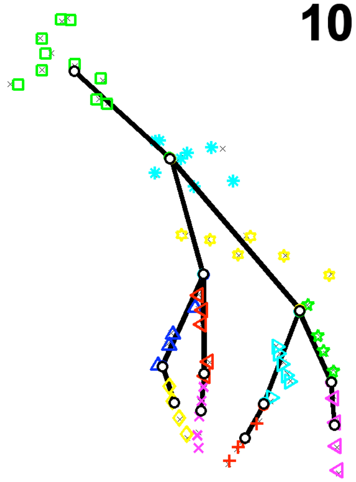
3. Learning Articulated Structure & Motion

- Most interesting objects are non-rigid
- Humans better-described as articulated figures - rigid parts connected by joints
- Advantages of higher-level model
  - Infer locations of occluded body parts
  - "Joint angles" between parts better representation for describing motion
- Challenge: parsing articulated motion typically requires a detailed hand-built physical model

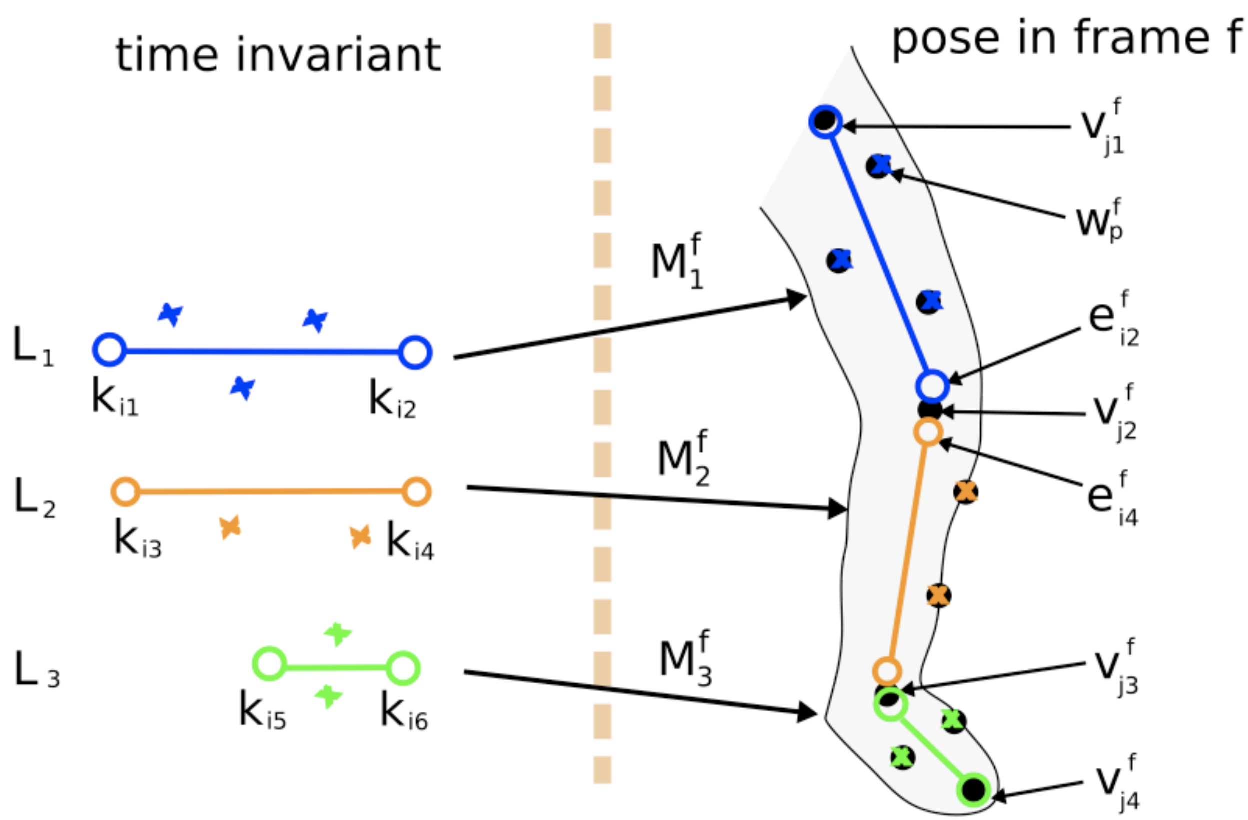
Probabilistic Stick Figures

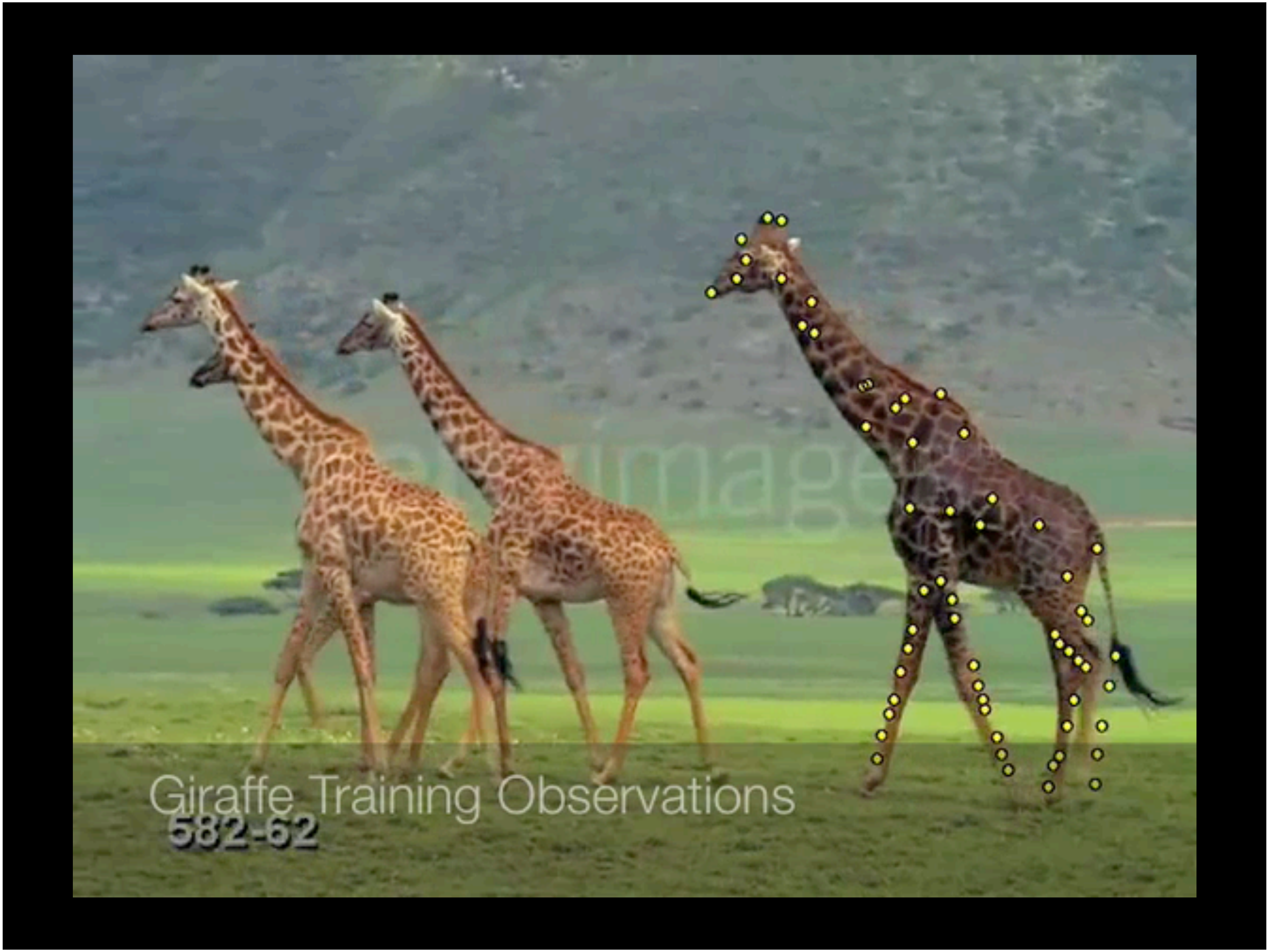


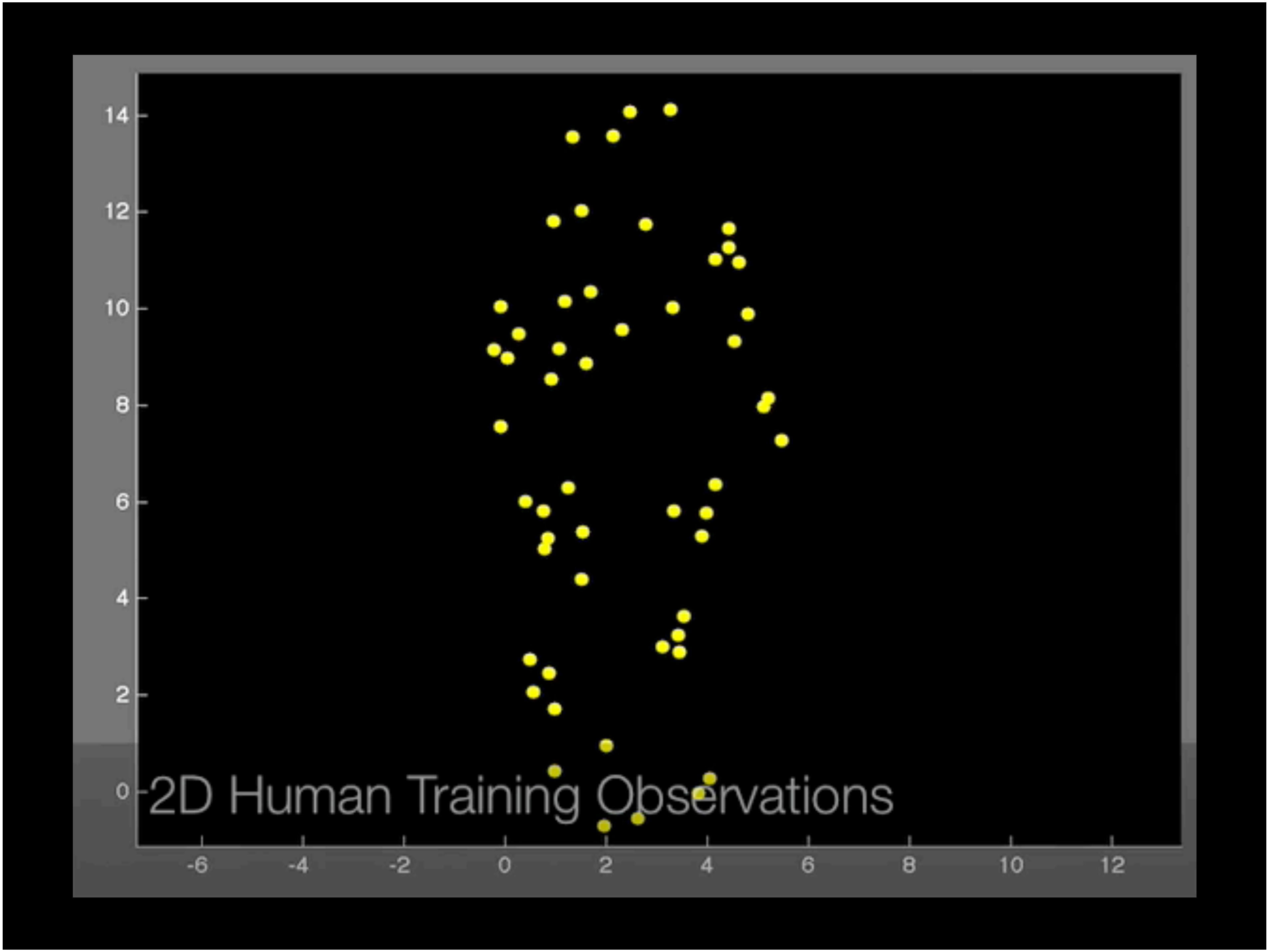
- Entirely-unsupervised recovery of 3D articulated structure and pose from 2D observations
- Probabilistic model for stick figure
- Begin from fully-disconnected structure (SFM)
- Fit parame ters using EM, resample segmentation
- Incrementally merge joints to greedily optimize data log-likelihood
- Model selection by locating maximum in objective function



Probabilistic Model

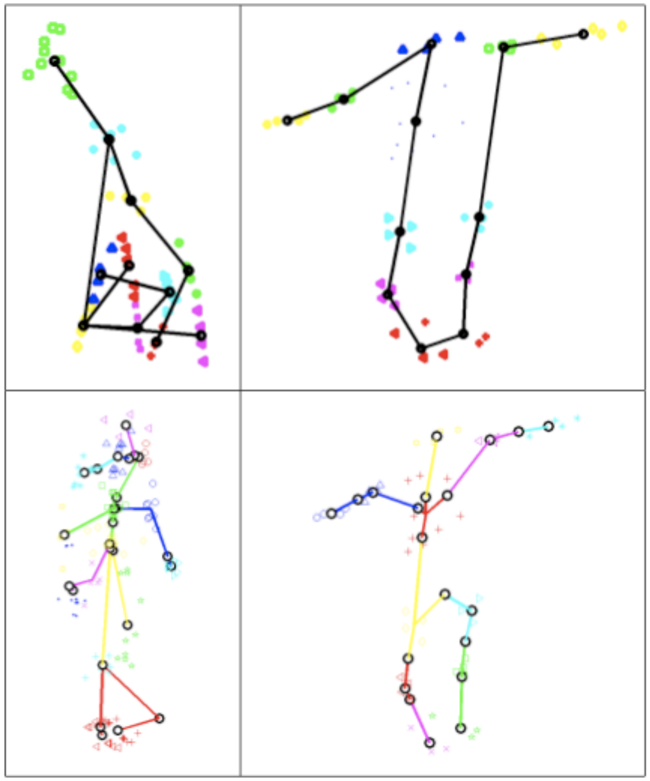


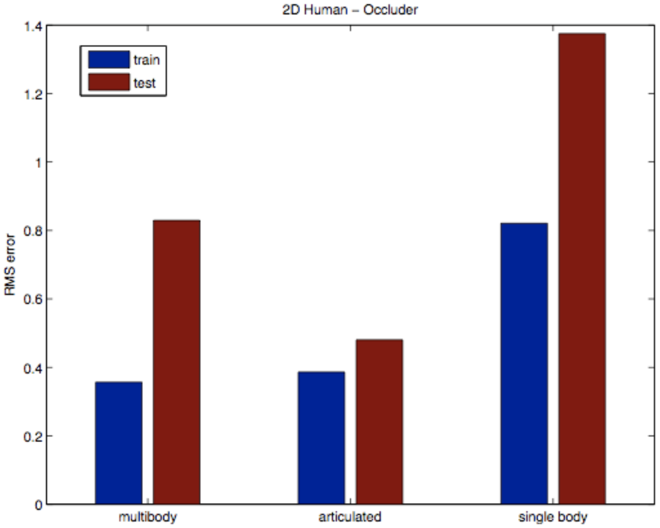

Giraffe Training Observations
582-62


2D Human Training Observations
14
12
10
8
6
4
2
0
-6
-4
-2
0
2
4
6
8
10
12

Contributions

- Robustly recover 3D structure from 2D, handling structured occlusions
- An explicit objective function
- Quantitative evaluation





Future Directions

- Incremental Tracking
  - Robustify to uncertainty in position & loss of track
- Combining Discriminative Features
  - Learn from partially-labeled data
- Articulated Structure
  - Integrate tracking of feature points, knowledge of structure can solve occlusions
- Broader goals
  - Performance that improves with amount of data and # of machines, End-to-end learning

Acknowledgements

Incremental Learning for Visual Tracking:

- Ming-Hsuan Yang, Jongwoo Lim, Ruei-sung Lin

Combining Discriminative Features

- Simon Osindero, Rich Zemel

Articulated Structure & Motion

- Danny Tarlow, Rich Zemel