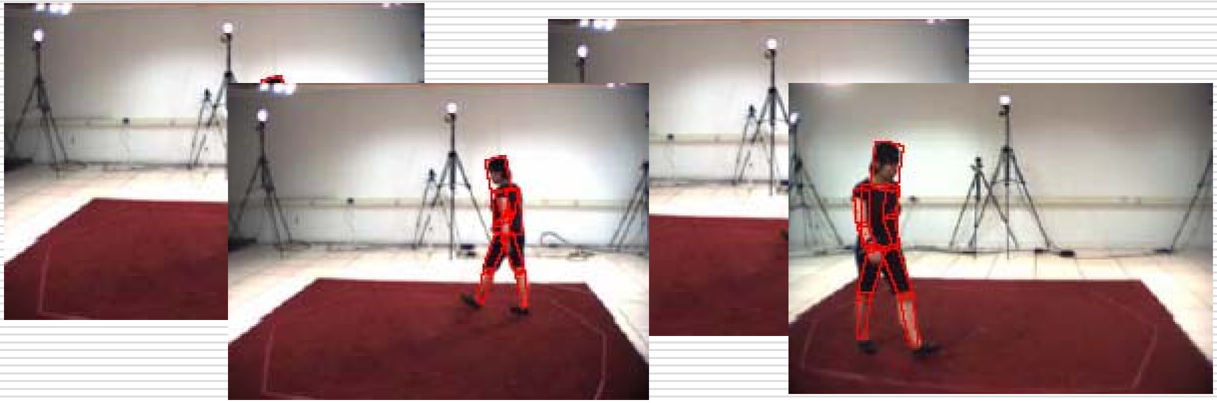
Part I: HumanEva-I dataset and evaluation metrics



Leonid Sigal **Michael J. Black**

Department of Computer Science

Brown University

http://www.cs.brown.edu/people/ls/
http://vision.cs.brown.edu/humaneva/

Motivation

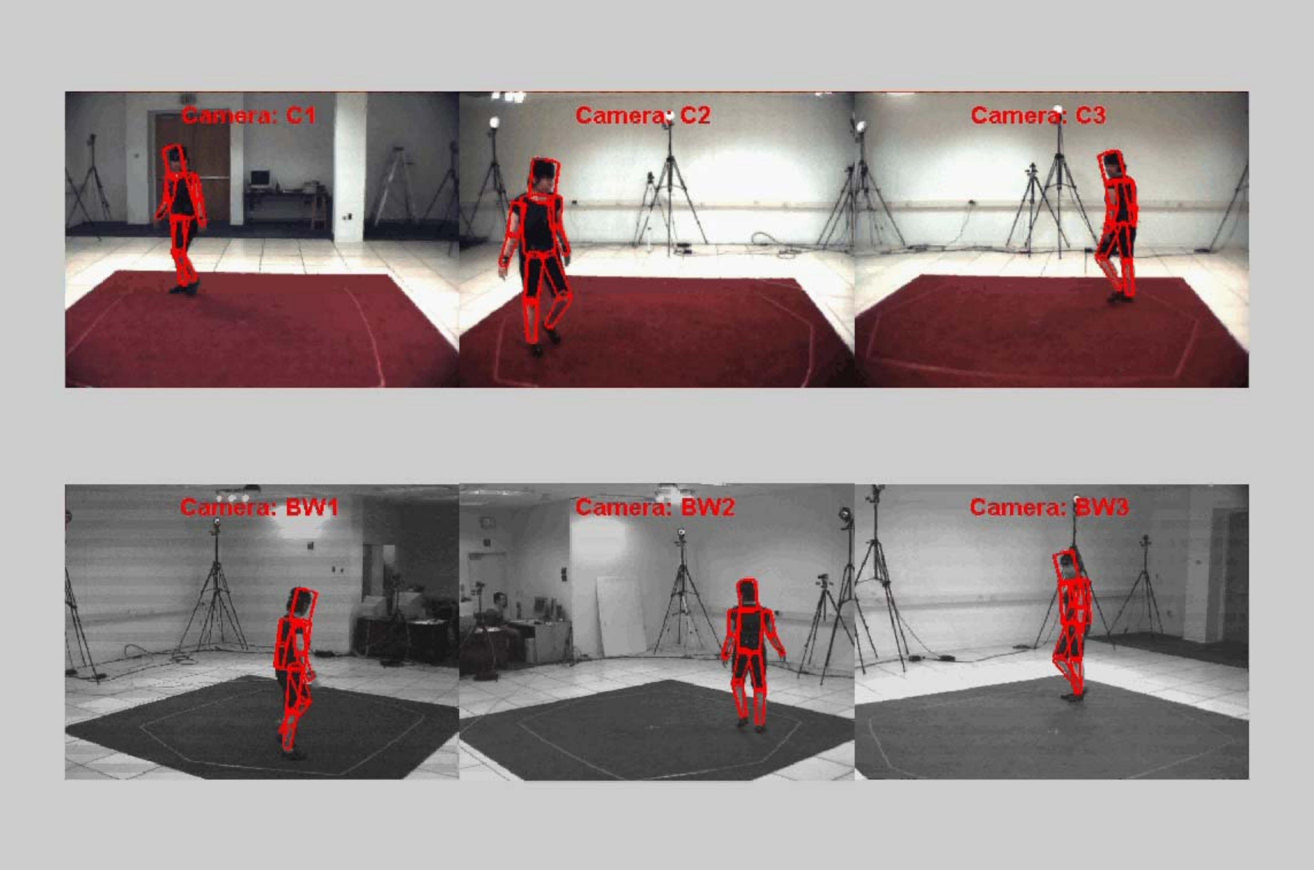
- **381+ papers in the past ~20 years** [***D.A. Forsyth***]

  - **Models**
    - 2D, 2.5D, 3D
    - Number body parts
    - Degrees of freedom per joint
    - ...
  - **Representation**
    - Kinematic (skeleton) tree
    - Part-based models
    - Graphical model
    - ...
  - **Shape**
    - Cylinders
    - Conic cross-section
    - Voxels
    - ...
  - **Likelihood**
    - Silhouette
    - Edges (1st derivative filters)
    - Ridges (2nd derivative filters)
    - Optical flow
    - ...
  - **Priors**
    - Action specific articulation priors
    - Temporal priors
    - ...
  - **Inference Methods**
    - Direct optimization
    - Stochastic optimization
    - Particle filters
    - Hidden Markov Models
    - Belief Propagation
    - ...

- **Real need for a common dataset with ground truth**

Motivation

- 381+ papers in the past ~20 years [*D.A. Forsyth*]



- **Real need for a common dataset with ground truth**

Motivation

That will help to address the following questions:

- What is the state-of-the art in human motion and pose estimation?
- What design choices are important and to what extent?
- What are the strengths and weaknesses of different methods?
- What are the main unsolved problems?

Similar datasets in other fields

- **Face detection** (FERET Dataset)

  P.J. Phillips, H. Moon, S.A. Rizvi and P.J. Rauss. “The FERET evaluation methodology for face-recognition algorithms”. *PAMI*, 2000.

- **Human gait identification** (HumanID Dataset)

  S. Sarkar, P. J. Phillips, Z. Liu, I.Robledo, P.Grother and K. W. Bowyer. “The Human ID Gait Challenge Problem: Data Sets, Performance, and Analysis. *PAMI*, 2005.

- **Dense stereo vision**

  D. Scharstein and R. Szeliski. “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms”. *IJCV*, 2002.

- **Activity Recognition** (CAVIAR Dataset)

  EC Funded CAVIAR project/IST 2001 37540.

- **Pedestrian Classification** (DaimlerChrysler Benchmark Dataset)

  S. Munder and D. M. Gavrila. “An Experimental Study on Pedestrian Classification”. *PAMI*, 2006.

HumanEva-I Hardware Setup

- **Motion Capture: Vicon (6 M1 cameras)**
  - Frame rate of 120 fps

- **Video Capture 1: Spica Tech**
  - 4 Pulnix TM6710 cameras
  - Synchronized capture to disk
  - Monochrome, 644 x 448 pixel, progressive scan.
  - Frame rate of 60 fps (120 fps max)
  - Hot-mirror filters (to filter out IR from Vicon)

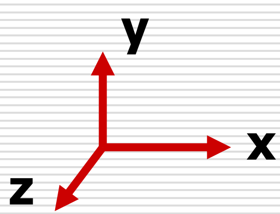
I don't recommend this camera!

- **Video Capture 2: IO Industries**
  - 3 UniQ UC685CL
  - Synchronized capture to disk
  - Color, 10-bit, 659x494 pixel, progressive scan.
  - Frame rate of 60 fps (110 fps max)

This one is much better!

(Thank you to Stan Sclaroff and BU Team)

- **Automated software synchronization. Single world coordinate frame**

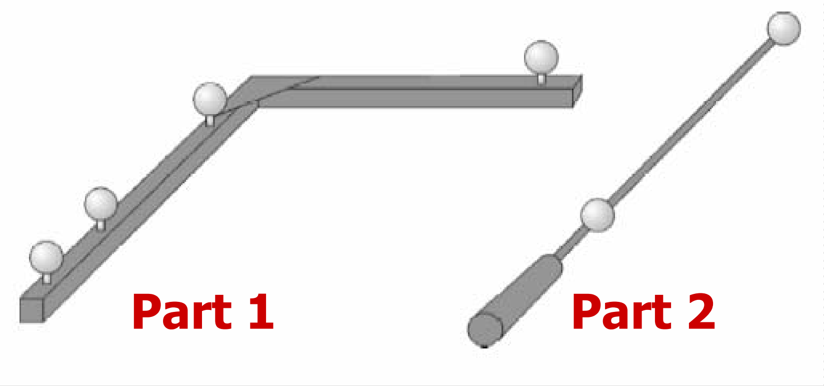


Data collection and processing

- **HumanEva-I data is calibrated and software synchronized**
  - Calibration of Mocap system
  - Intrinsic calibration of video cameras ($\mathbf{F_c}$, $\mathbf{C_c}$, $\mathbf{K_c}$, $\boldsymbol{\alpha_c} = 0$)
  - Extrinsic calibration of video cameras ($\mathbf{R_c}$, $\mathbf{T_c}$)
  - Temporal scaling ($\mathbf{A_c}$)
  - Temporal alignment ($\mathbf{B_{c,s}}$) (*per sequence*)

Data collection and processing

- **HumanEva-I data is calibrated and software synchronized**
  - Calibration of Mocap system

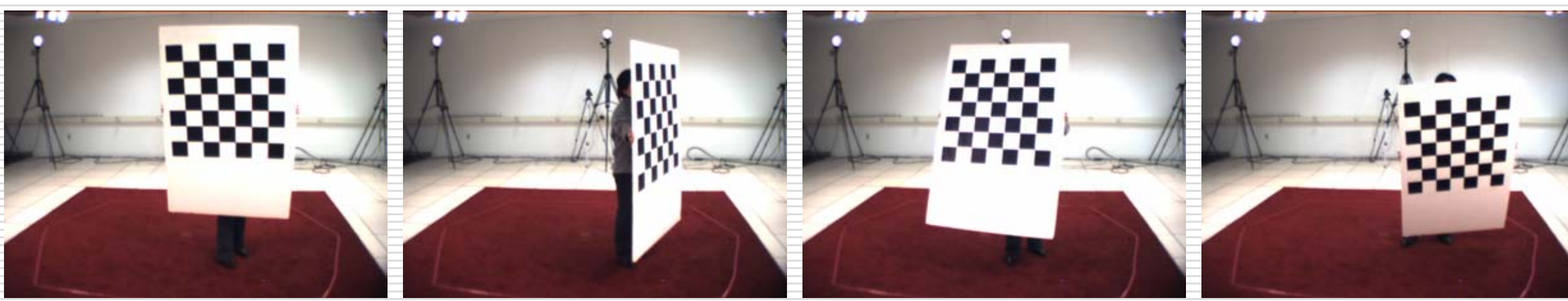


  - Intrinsic calibration of video cameras ($\mathbf{F}_c$, $\mathbf{C}_c$, $\mathbf{K}_c$, $\boldsymbol{\alpha}_c = 0$)
  - Extrinsic calibration of video cameras ($\mathbf{R}_c$, $\mathbf{T}_c$)
  - Temporal scaling ($\mathbf{A}_c$)
  - Temporal alignment ($\mathbf{B}_{c,s}$) (*per sequence*)

Data collection and processing

- **HumanEva-I data is calibrated and software synchronized**
  - Calibration of Mocap system
  - Intrinsic calibration of video cameras (**F_c**, **C_c**, **K_c**, **α_c** = 0)
    - Focal point **– $F_c \in R^2$**
    - Principle point **– $C_c \in R^2$**
    - Radial distortion **- $K_c \in R^5$**
    - Skew (we assume squared pixels) **- α_c** = 0

Based on Caltech Calibration Toolbox for Matlab

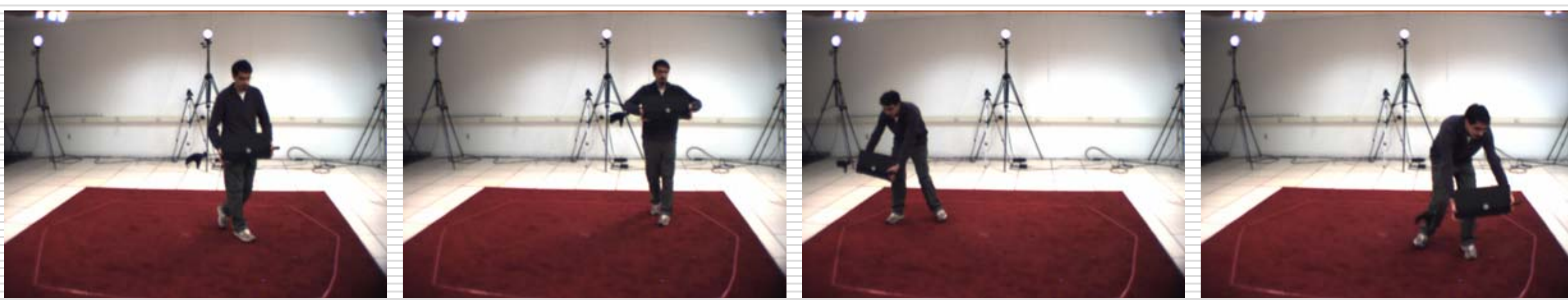


  - Extrinsic calibration of video cameras (**R_c**, **T_c**)
  - Temporal scaling (**A_c**)
  - Temporal alignment (**$B_{c,s}$**) (*per sequence*)

Data collection and processing

- **HumanEva-I data is calibrated and software synchronized**
  - Calibration of Mocap system
  - Intrinsic calibration of video cameras ($\mathbf{F_c}$, $\mathbf{C_c}$, $\mathbf{K_c}$, $\boldsymbol{\alpha_c} = 0$)
  - Extrinsic calibration of video cameras ($\mathbf{R_c}$, $\mathbf{T_c}$)
    - Global rotation – $\mathbf{R_c \in SO(3)}$
    - Global translation – $\mathbf{T_c \in R^3}$
  - Temporal scaling ($\mathbf{A_c}$)

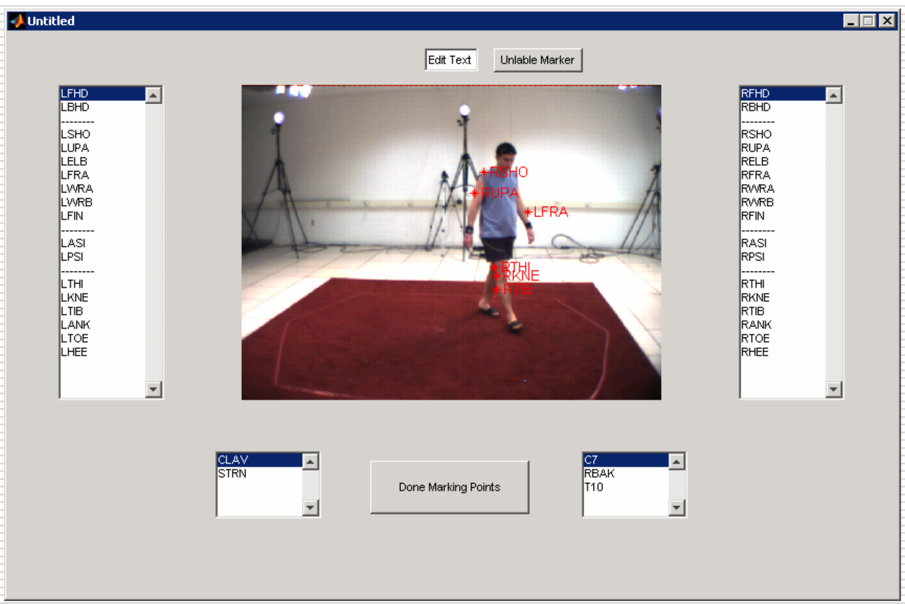
Based on Caltech Calibration Toolbox for Matlab



- Temporal alignment ($\mathbf{B_{c,s}}$) (*per sequence*)

Data collection and processing

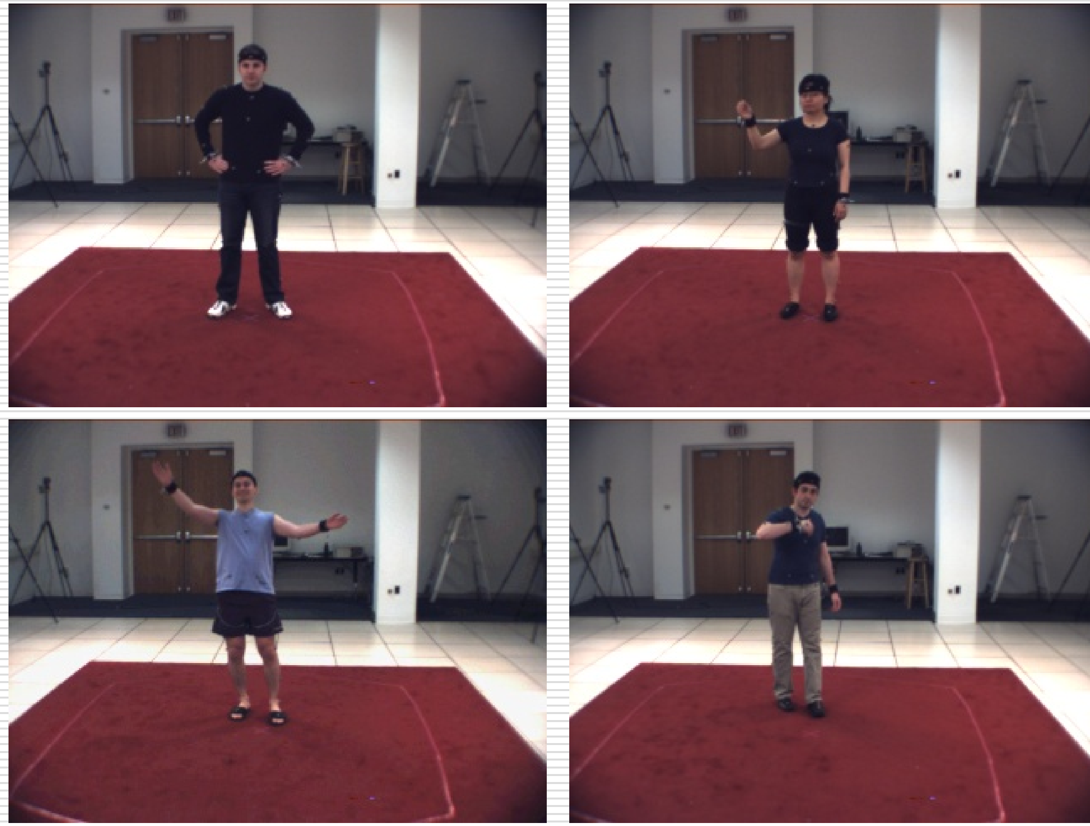
- **HumanEva-I data is calibrated and software synchronized**
  - Calibration of Mocap system
  - Intrinsic calibration of video cameras ($\mathbf{F_c}$, $\mathbf{C_c}$, $\mathbf{K_c}$, $\mathbf{\alpha_c}$ = 0)
  - Extrinsic calibration of video cameras ($\mathbf{R_c}$, $\mathbf{T_c}$)
  - Temporal scaling ($\mathbf{A_c}$)
  - Temporal alignment ($\mathbf{B_{c,s}}$) (*per sequence*)



Manually mark some visible markers in a few frames
+
use direct optimization

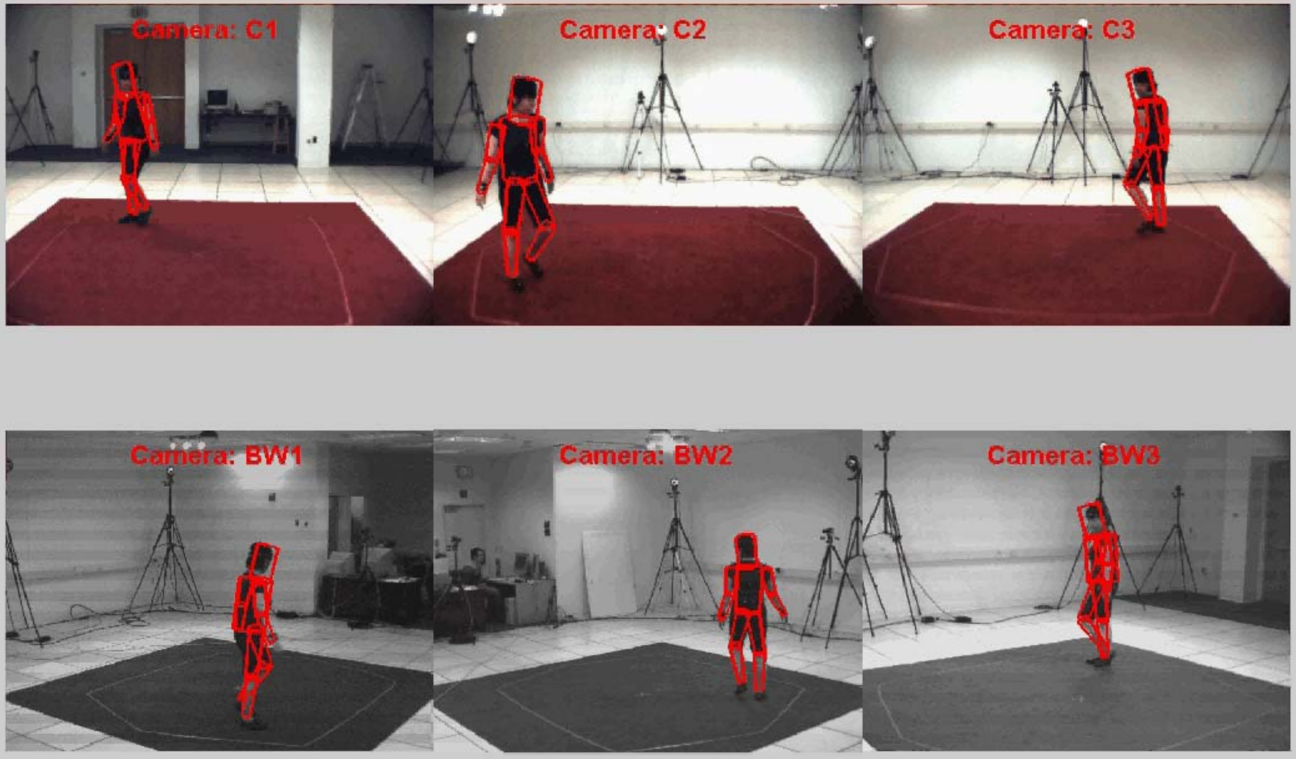
HumanEva-I Dataset

- 7 video cameras
  - 4 grayscale
  - 3 color
- 4 subjects
- 6 actions each



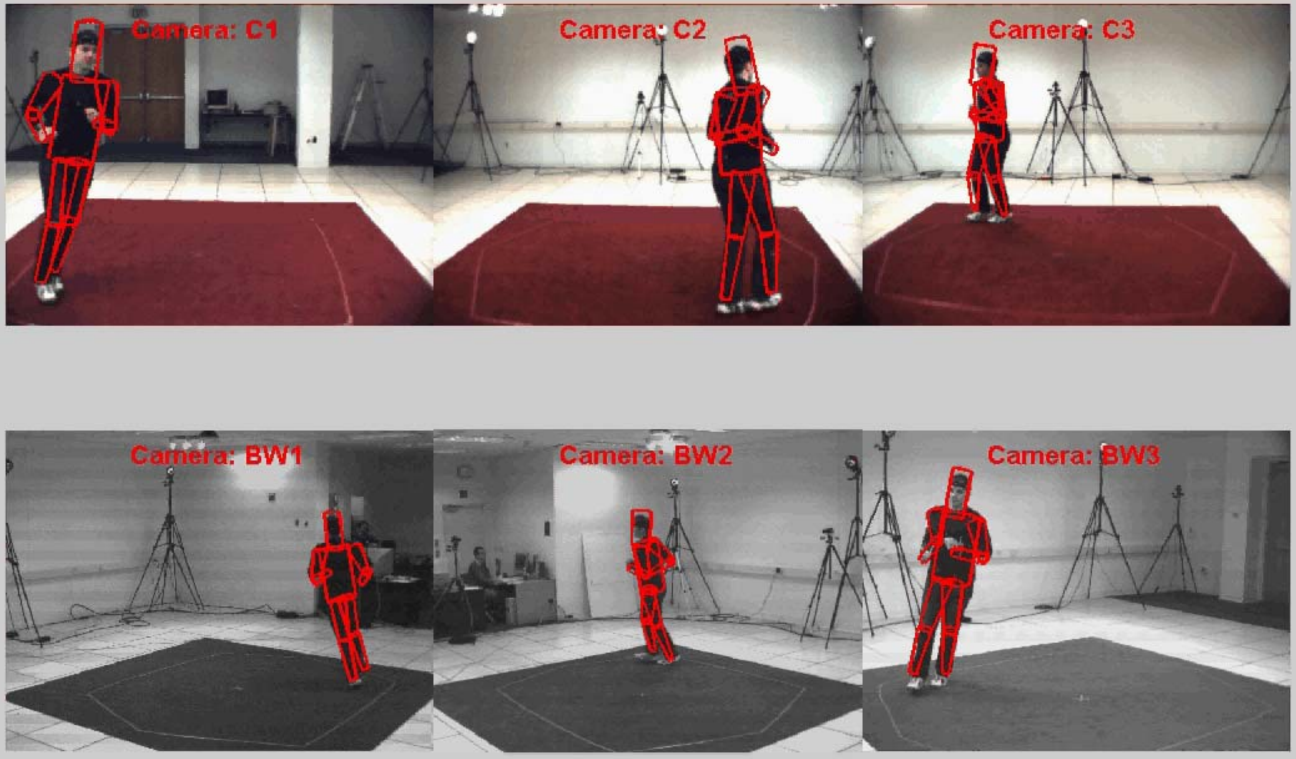
- Each action is repeated 3 times (twice with synchronized MoCap and video and once with MoCap Only)

HumanEva-I Dataset



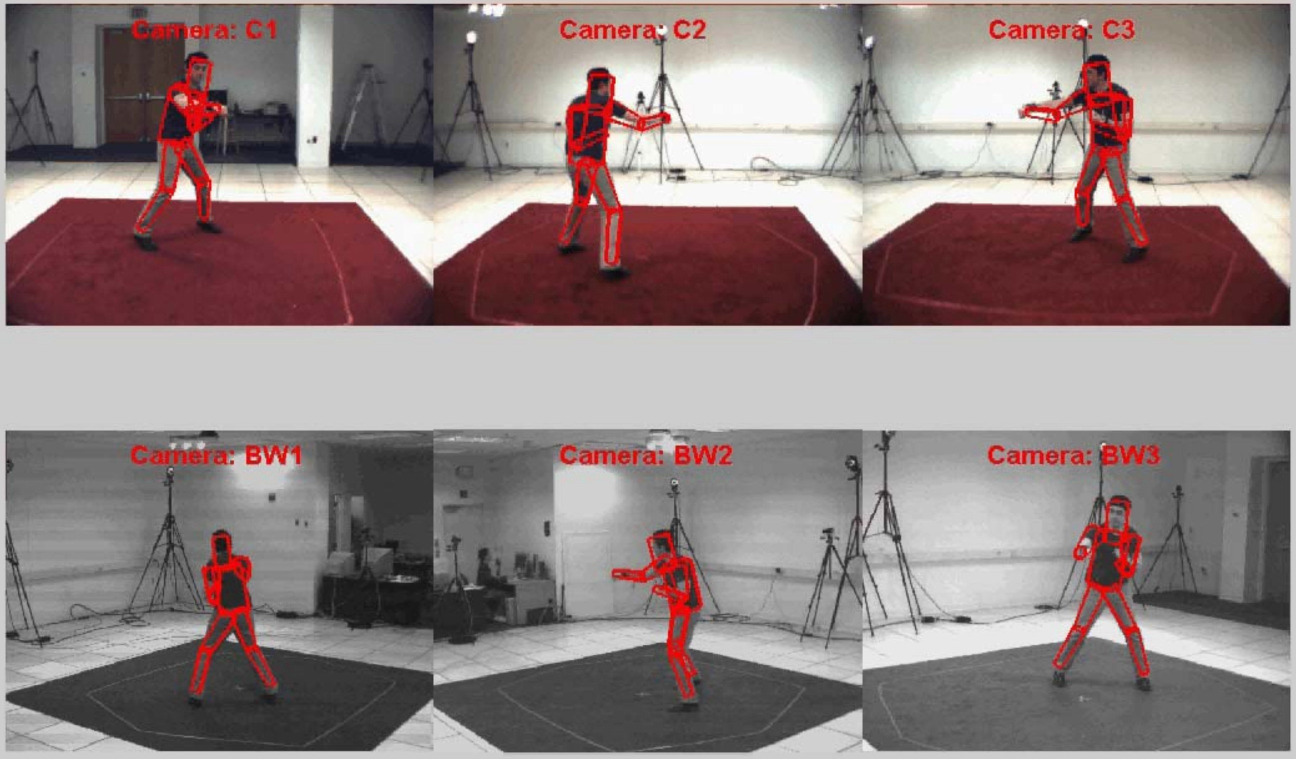
Walking, Subject - S1

HumanEva-I Dataset



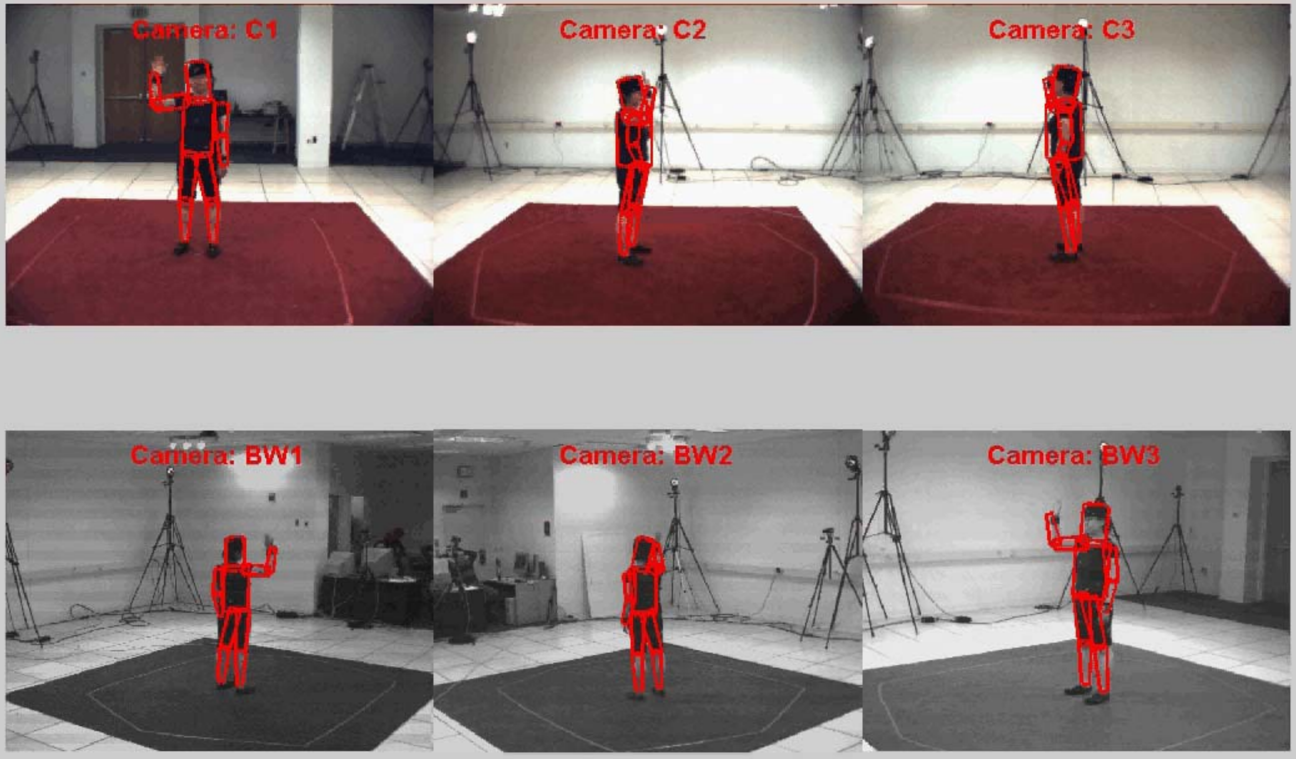
Jogging, Subject - S3

HumanEva-I Dataset



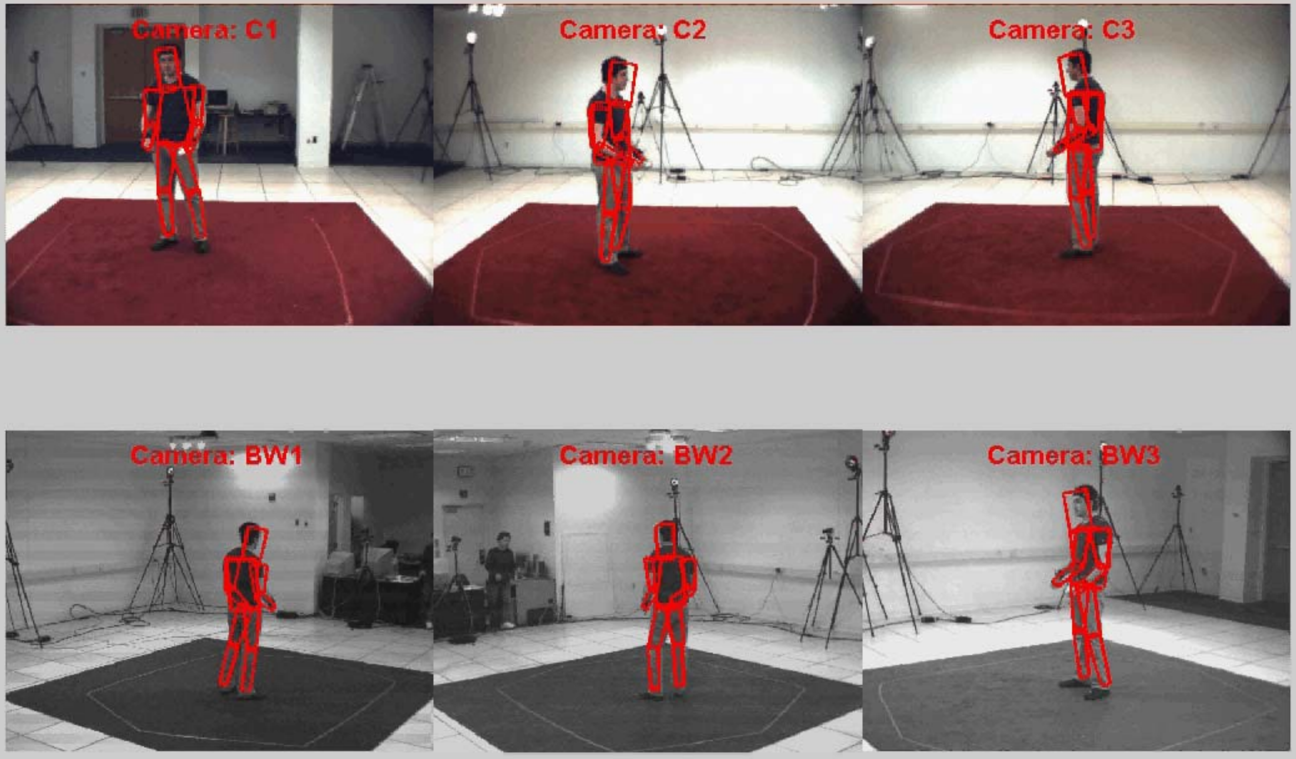
Boxing, Subject – S2

HumanEva-I Dataset



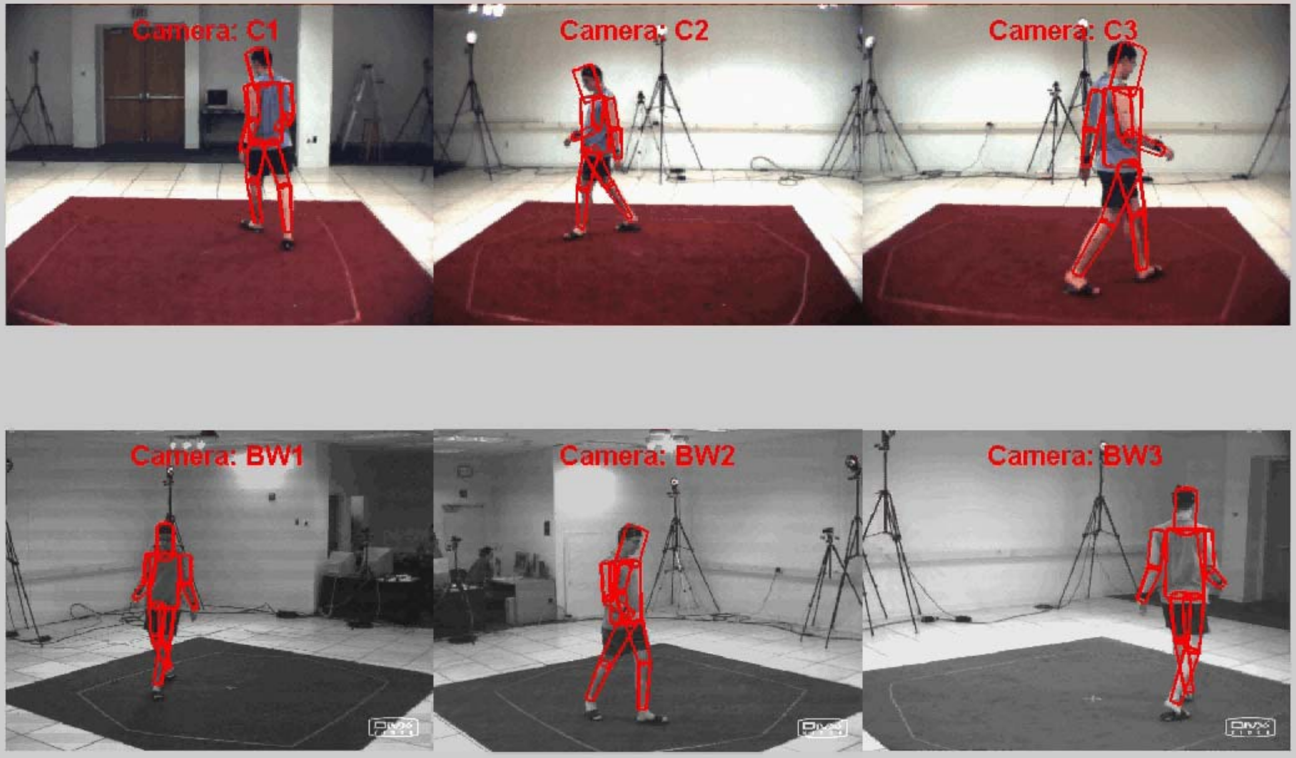
Gestures, Subject – S1

HumanEva-I Dataset



Throw and Catch, Subject – S2

HumanEva-I Dataset



Combo, Subject – S4

HumanEva-I Dataset

- **Training**
  - Mocap (~35,000 frames)
  - Synchronized MoCap and Video (~6,800 frames)
- **Validation**
  - Synchronized MoCap and Video (~6,800 frames)
- **Testing**
  - Video only (~24,000 frames)
  - Synchronized MoCap is withheld
  - On-line evaluation (to disallow tweaking of parameters)

Background Subtraction

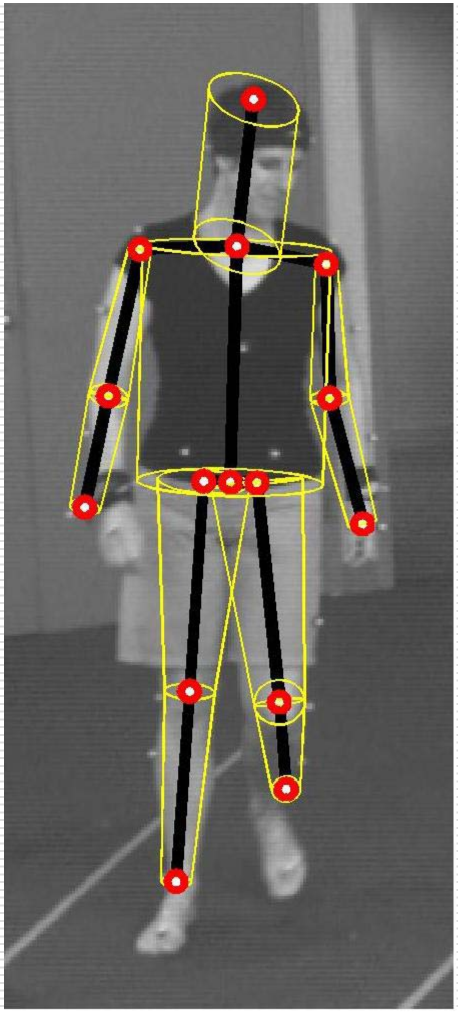
- Background template images are given
- Sample background subtraction support code

Better background subtraction techniques will be presented today

Color Cameras

Grayscale Cameras

Quantitative Evaluation



- Average distance between markers corresponding to joints and limb endpoints

$$D(X, \hat{X}, \hat{\Delta}) = \sum_{m=1}^{M} \frac{\hat{\delta}_m \|x_m - \hat{x}_m\|}{\sum_i^M \hat{\delta}_i}$$

where,

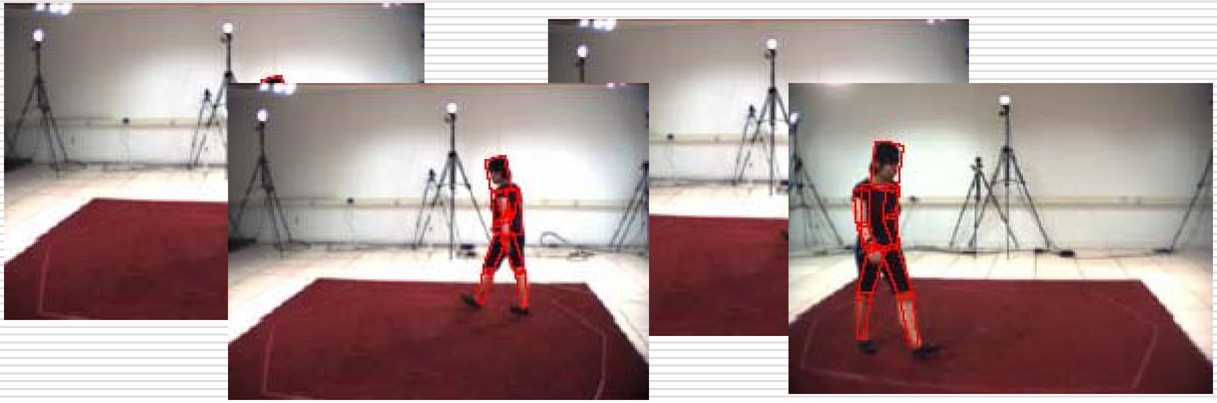
$$X = \{x_1, x_2, \dots, x_M\}, \quad x_i \in \Re^3$$

$$\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_M\}, \quad \hat{x}_i \in \Re^3$$

$$\hat{\Delta} = \{\hat{\delta}_1, \hat{\delta}_2, \dots, \hat{\delta}_M\}, \quad \hat{\delta}_i \in [0,1]$$

- **M=15**

Part II: Performance of APF on HumanEva-I



Alexandru Balan   Leonid Sigal   Michael J. Black

Department of Computer Science
Brown University

http://www.cs.brown.edu/people/alb/
http://vision.cs.brown.edu/humaneva/

Benchmark Reference Algorithm

- **Annealed Particle Filtering** [***Deutscher, Blake & Reid, CVPR'00***]

Alexandru Balan, Leonid Sigal and Michael J. Black. "A Quantitative Evaluation of Video-based 3D Person Tracking". *VS-PETS*, 2005

- **Based on general Bayesian recursive posterior estimation**

Likelihood: **probability that pose generated the image**

Temporal prior

$$p(\mathbf{X}_t \mid \vec{\mathbf{Y}}_{1:t}) \propto p(\vec{\mathbf{Y}}_t \mid \mathbf{X}_t) \int p(\mathbf{X}_t \mid \mathbf{X}_{t-1}) \, p(\mathbf{X}_{t-1} \mid \vec{\mathbf{Y}}_{1:t-1}) \, d\mathbf{X}_{t-1}$$

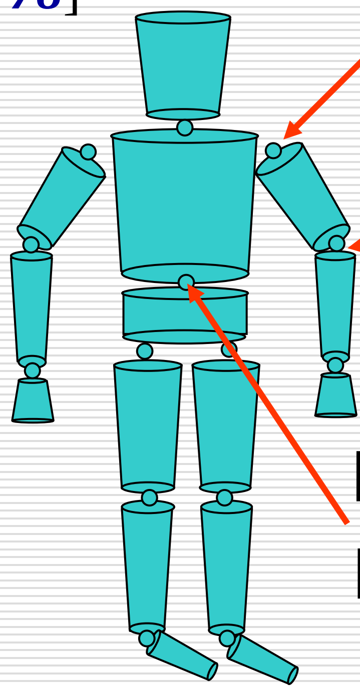
Posterior: **probability of pose given image evidence**

Articulated Body Model

Kinematic tree:

[***Marr&Nishihara ’78***]

40D space



$$[\theta_x^{1,0}, \theta_y^{1,0}, \theta_z^{1,0}]$$

$$[\theta_x^{2,1}]$$

$$[\tau_x^{0,g}, \tau_y^{0,g}, \tau_z^{0,g}]$$

$$[\theta_x^{0,g}, \theta_y^{0,g}, \theta_z^{0,g}]$$

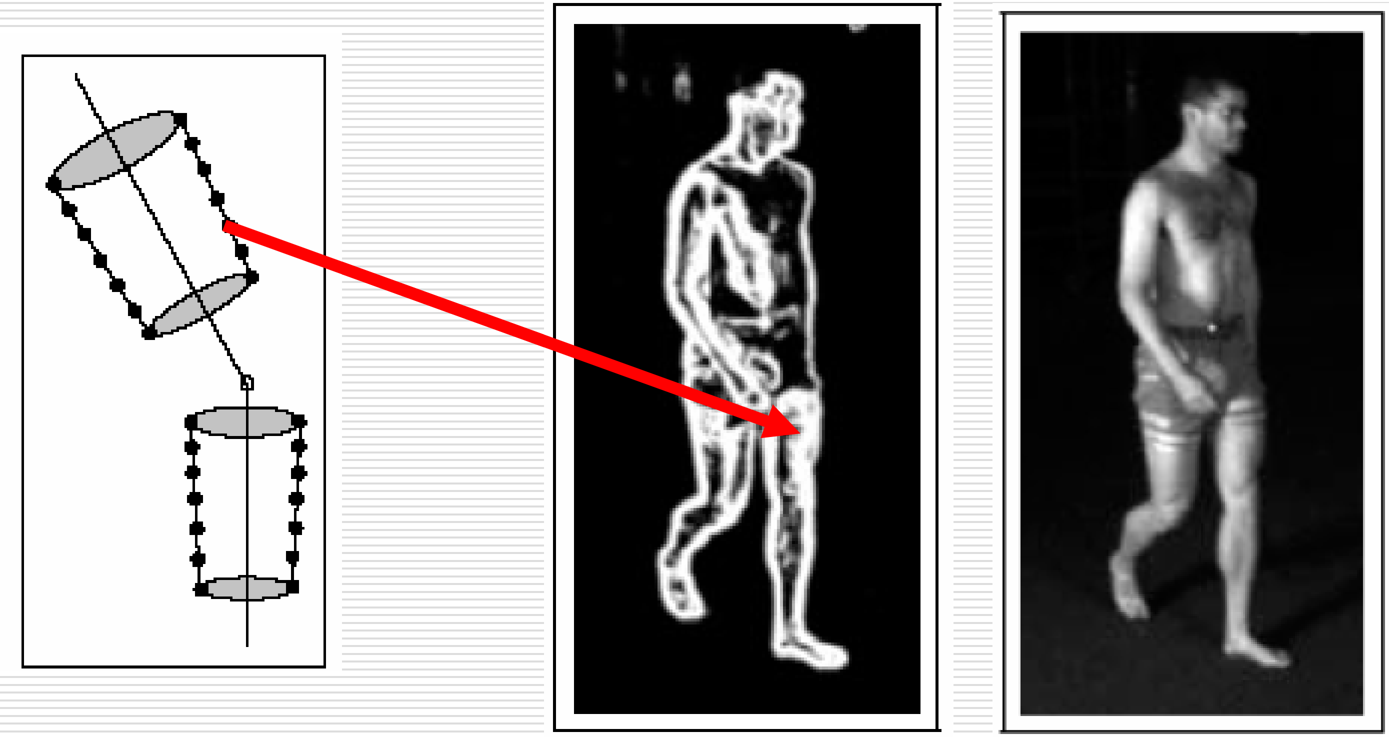
Represent a “pose” at time t by a vector of **all parameters**: $\mathbf{X}_t$

Likelihood $p(\mathbf{Y}_t \,|\, \mathbf{X}_t)$

p(bg pixel | limb location and orientation)

[*Deutscher, Blake & Reid, CVPR'00*]

Likelihood $p(\mathbf{Y}_t \mid \mathbf{X}_t)$



p(edge filter response | limb edge location and orientation)

[*Deutscher, Blake & Reid, CVPR'00*]

Temporal Prior $p(\mathbf{X}_t \mid \mathbf{X}_{t-1})$

- **Prior can be very simple** [***Deutscher, Blake & Reid, CVPR'00***]

$$p(\mathbf{X}_t \mid \mathbf{X}_{t-1}) = N(\mathbf{X}_{t-1}, Q)$$

- **Include constraints via a pose prior (using rejection sampler)**
  - Self-intersection constraints
  - Range of motion constraints for individual joints (can be learned from MoCap)
    - Action-specific
    - General

Inference using Particle Filtering

Posterior $p(\mathbf{X}_{t-1} | \vec{\mathbf{Y}}_{t-1})$

↓ **sample**

Temporal dynamics

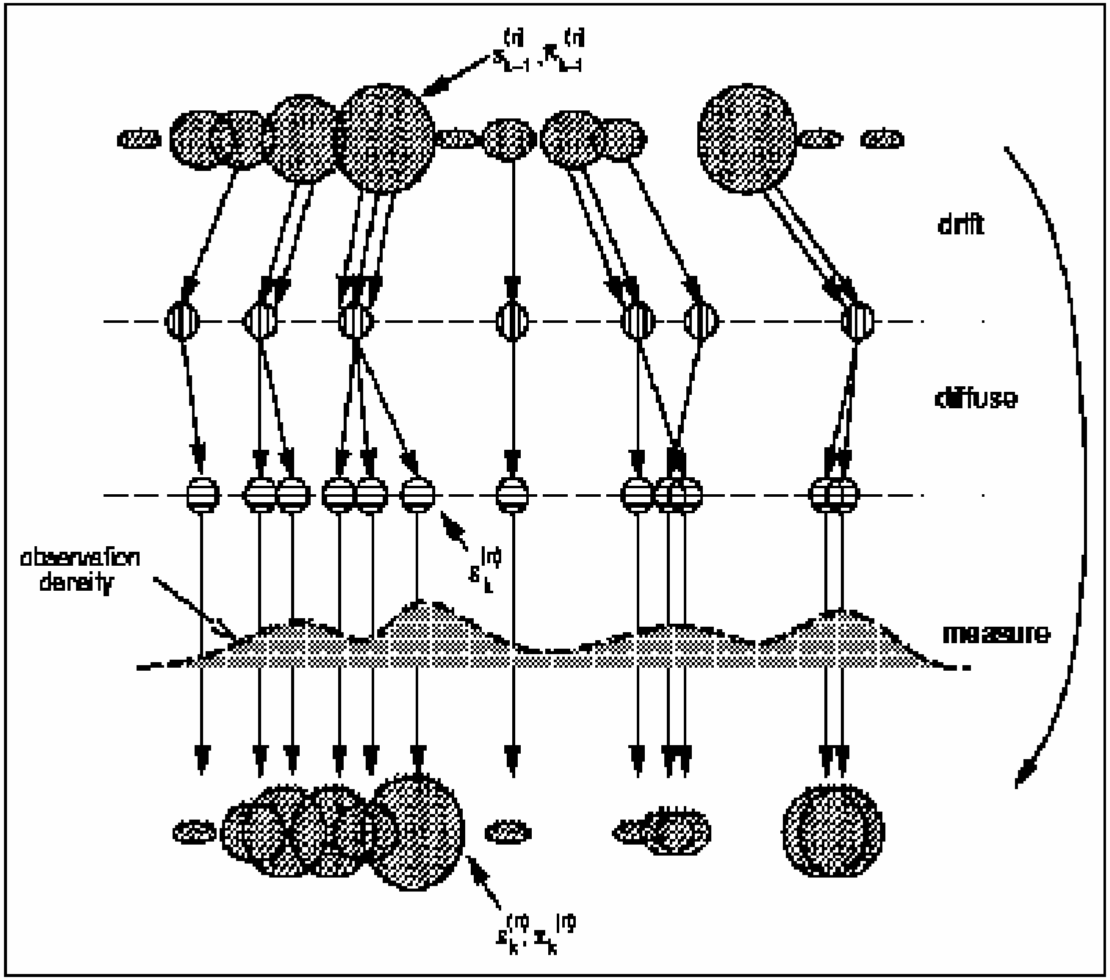
$p(\mathbf{X}_t | \mathbf{X}_{t-1})$

↓ **sample**

Likelihood $p(\mathbf{Y}_t | \mathbf{X}_t)$

↓ **normalize**

Posterior $p(\mathbf{X}_t | \vec{\mathbf{Y}}_t)$



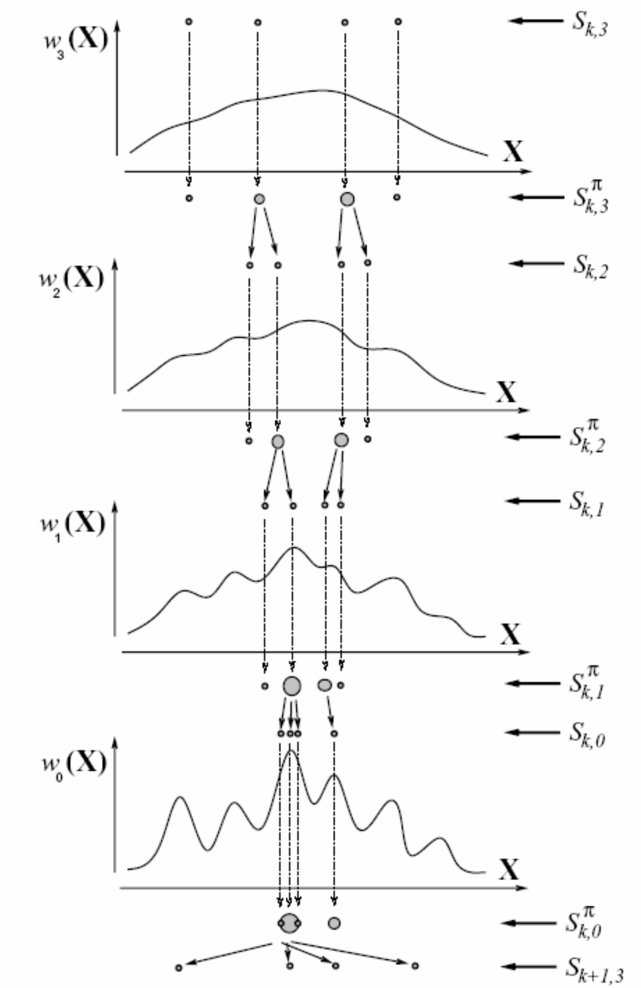
[*Isard & Blake '96*]

Annealed Particle Filter

Smooth the likelihood

$$p(\mathbf{Y}_t \mid \mathbf{X}_t)^{\beta_m}$$

Annealing parameter



Conclusions from VS-PETS 2005

Q: How does performance scale with the number of views?

A: Works poorly with < 3 views, does not gain significant benefit from more then 3 views

Q: How does performance scale with the number of particles?

A: Exponential [log(N) vs. error = straight line]

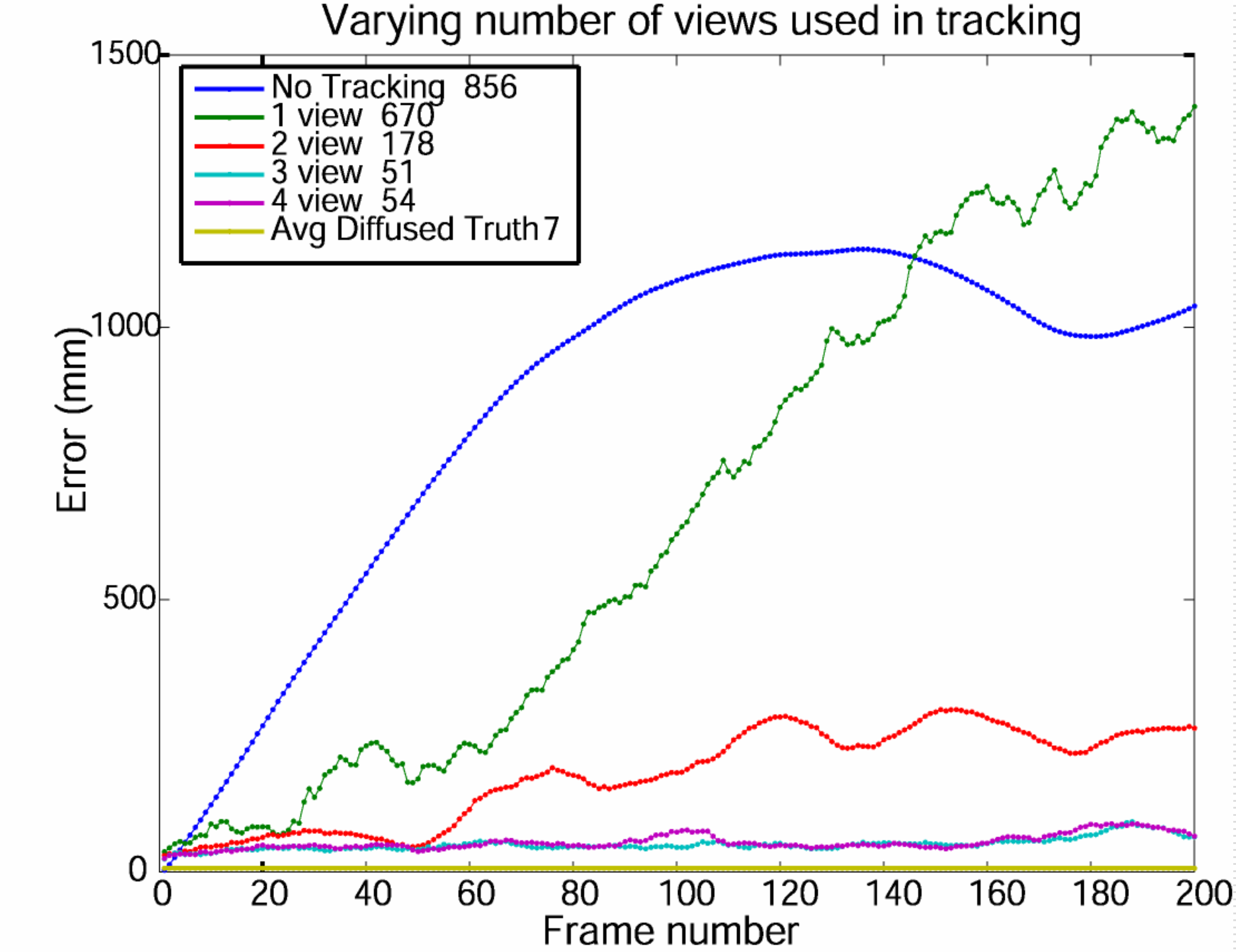
Q: How do different choices of likelihoods effect performance?

A: Silhouettes are most useful, adding edge features helps with internal edges

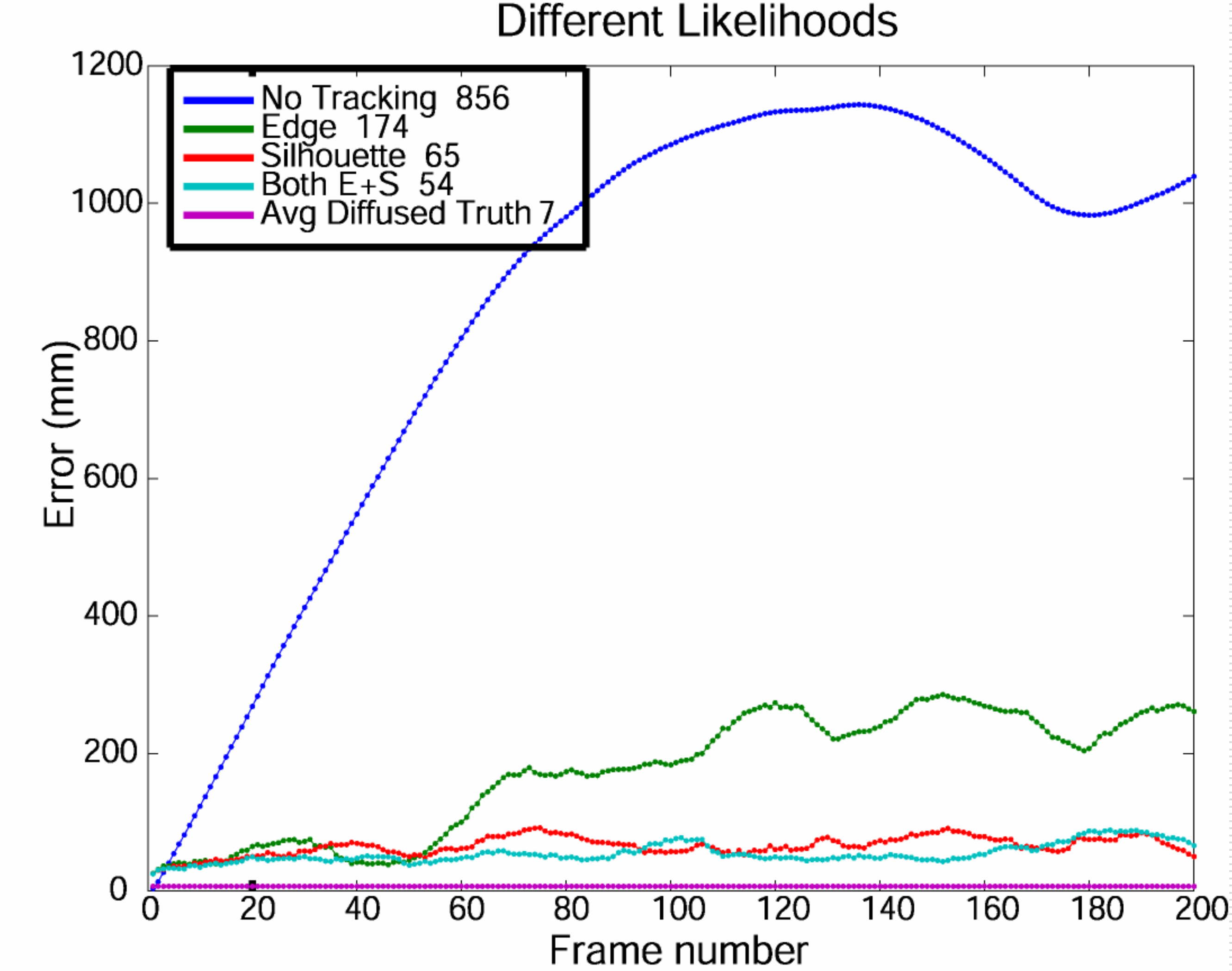
Q: Does annealing help?

A: Not as much as we initially thought

Results from VS-PETS 2005



Results from VS-PETS 2005

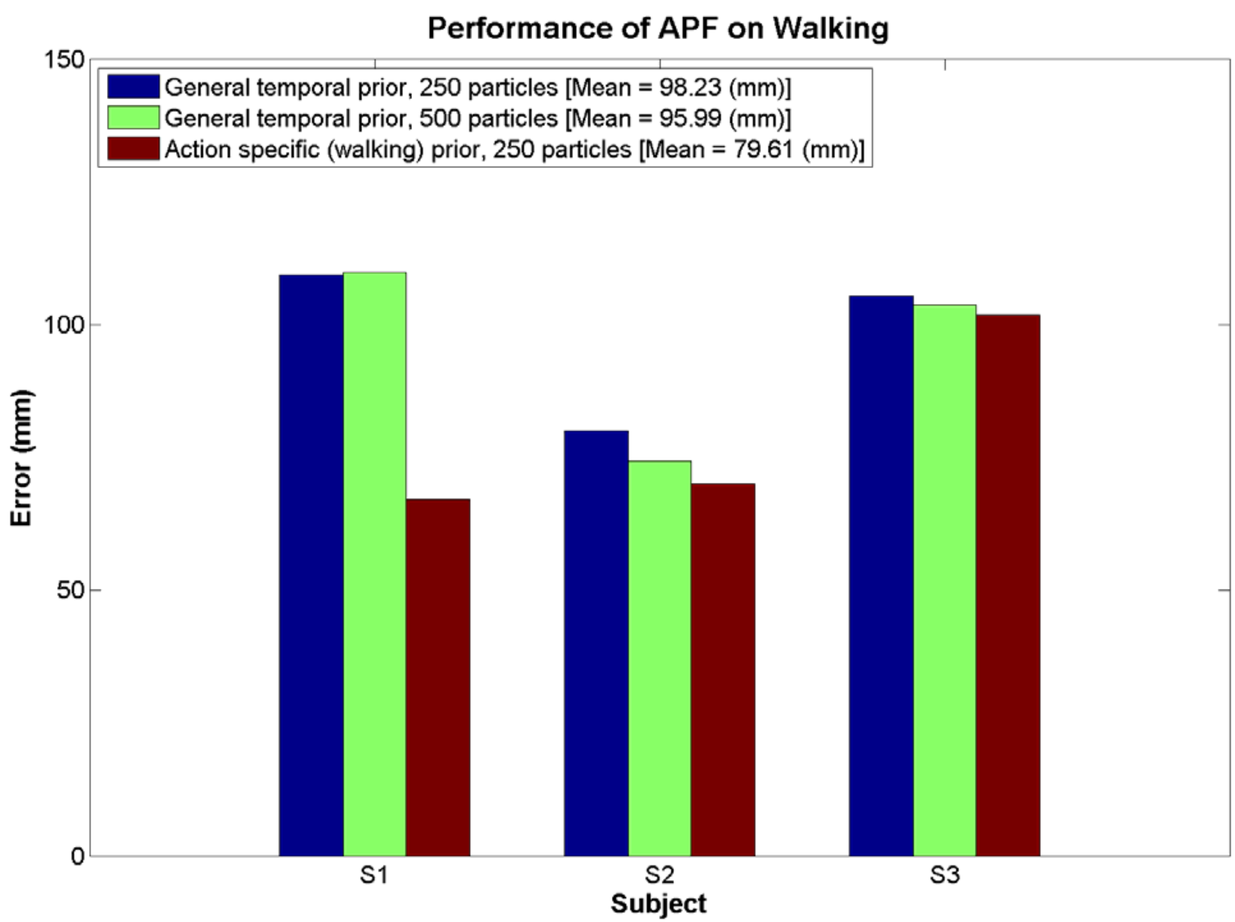


HumanEva-I Experiments

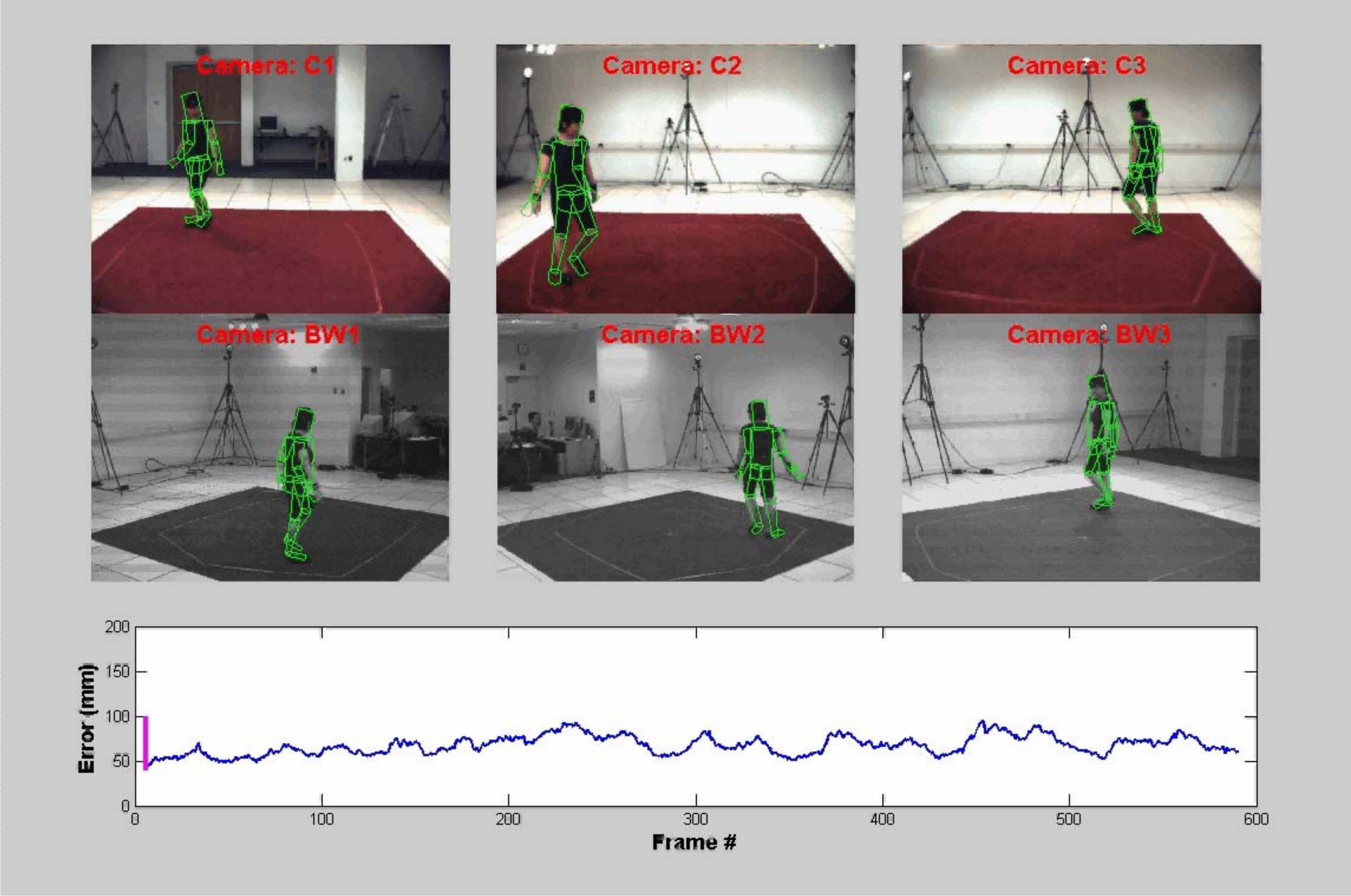
- **We know how to set up APF to produce good tracking performance**
  - Use all 7 views
  - Initialize from ground truth
  - Use 250 particles (more is better)
  - 5 layers of annealing
  - Likelihood (silhouettes + edges)
- **Do observations we have made on VS-PETS data generalize to HumanEva dataset**
- **Do action specific priors help and to what extent?**

Do action specific priors help?

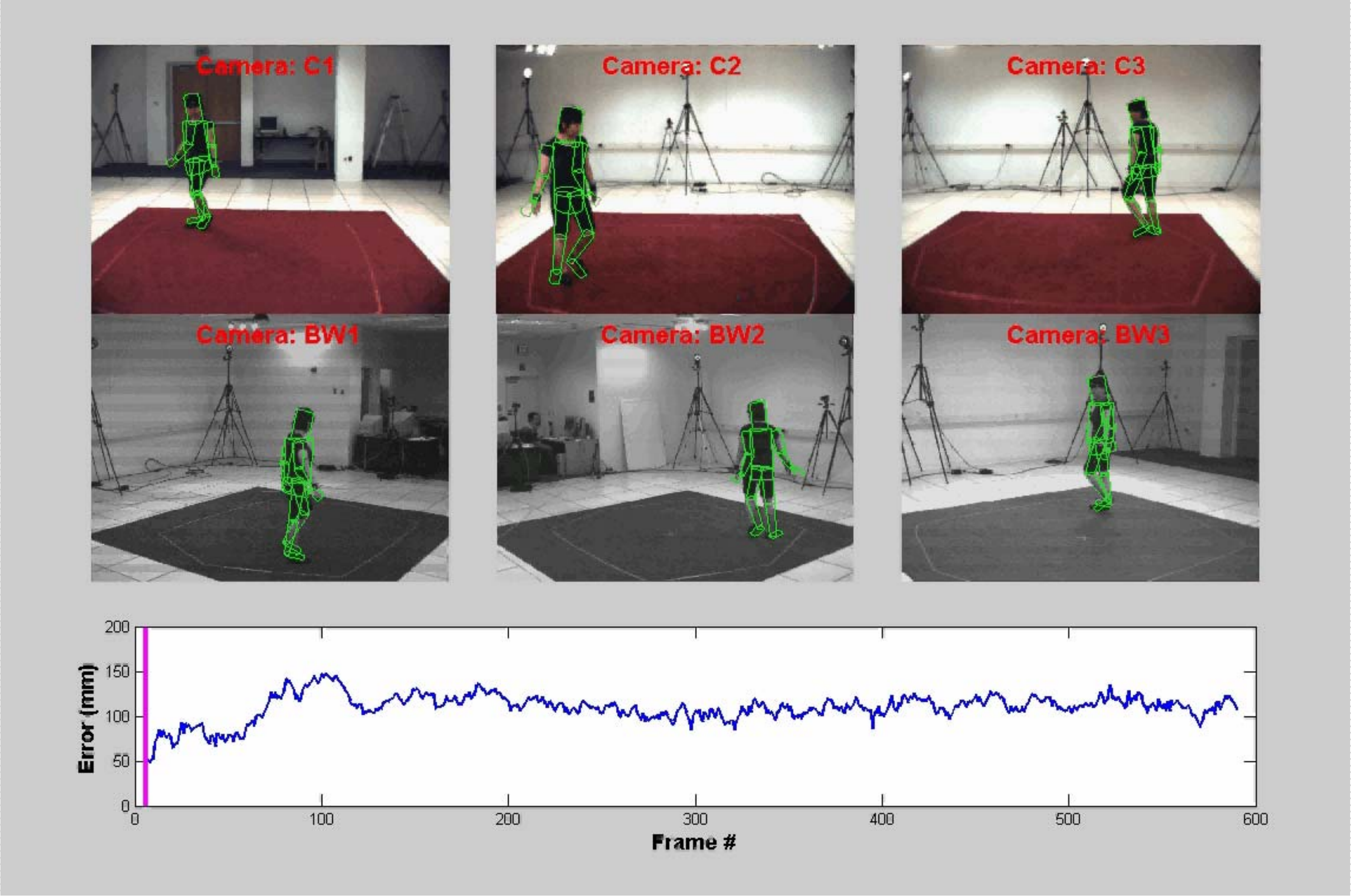
- How much? (**Maybe the benefits of the general prior outweigh the additional error**)



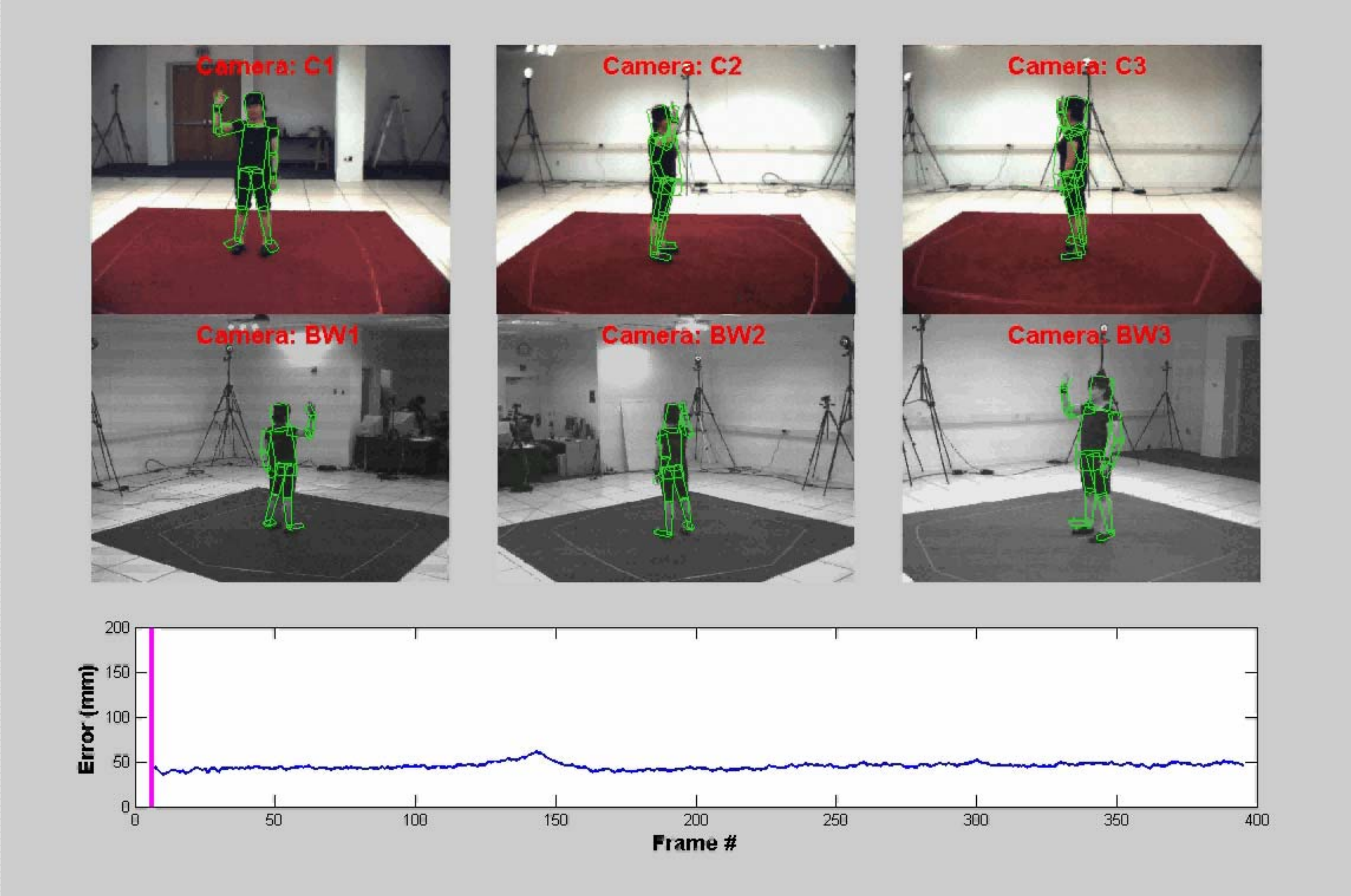
Action-specific prior on walking



General prior on walking



Other actions …



Collaborators

- **HumanEva-I**
  - Alexandru Balan (Brown University)
  - Michael Black (Brown University)
  - Rui Li (Boston University)
  - Payman Yadollahpour (Brown University)
  - Ming-Hsuan Yang (Honda Research Institute)
  - Horst Haussecker (Intel Research)
- **Annealed Particle Filtering**
  - Alexandru Balan (Brown University)
  - Michael Black (Brown University)
- **EHuM Program Committee Members**
- **All contributors and attendees**