Representation and Matching of Articulated Shapes

Jiayong Zhang, Robert Collins and Yanxi Liu

The Robotics Institute

Carnegie Mellon University

Goal: parse static pictures of people

What does "parse" mean?

- find boundary shape
- locate body parts

Why does it matter?

- gait recognition
- tracking
- activity analysis

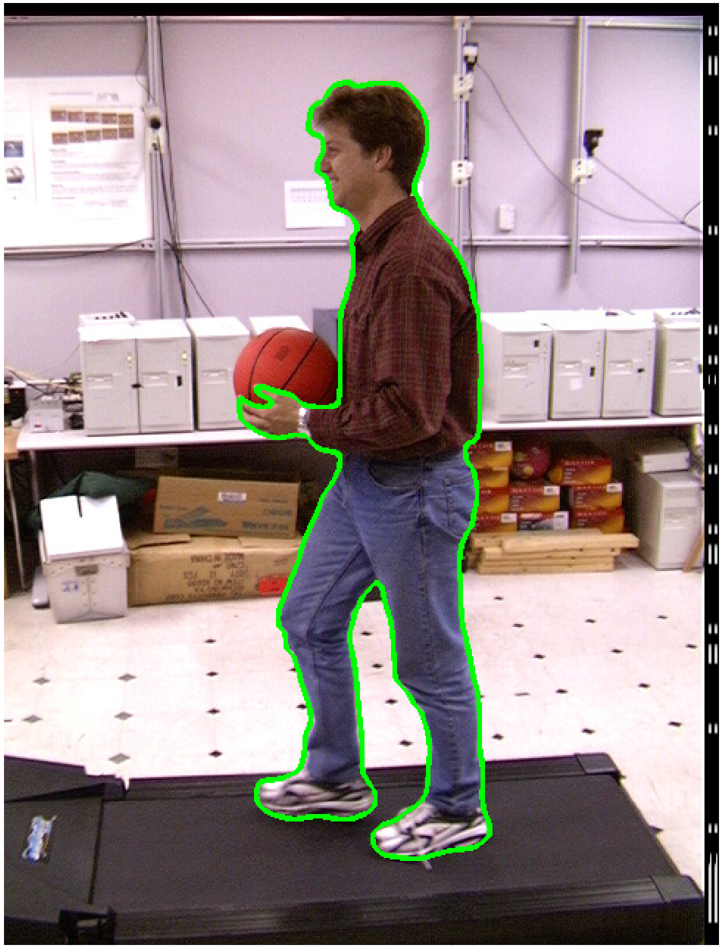
Goal: parse static pictures of people

What does "parse" mean?

- find boundary shape
- locate body parts

Why does it matter?

- gait recognition
- tracking
- activity analysis



Goal: parse static pictures of people

What does “parse” mean?

- find boundary shape
- locate body parts

Why does it matter?

- gait recognition
- tracking
- activity analysis

Goal: parse static pictures of people

What does “parse” mean?

- find boundary shape
- locate body parts

Why does it matter?

- gait recognition
- tracking
- activity analysis

How to parse?

Goal: parse static pictures of people

What does "parse" mean?

- find boundary shape
- locate body parts

Why does it matter?

- gait recognition
- tracking
- activity analysis

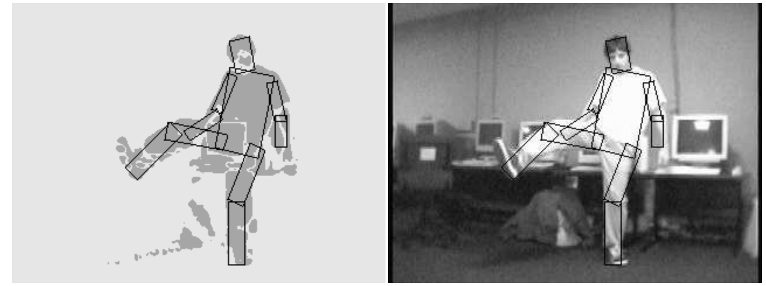
Assumptions

- single person
- side view
- no external occlusion

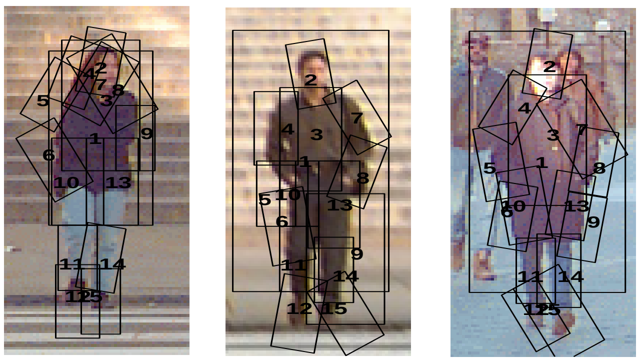
How to parse?

Previous Work

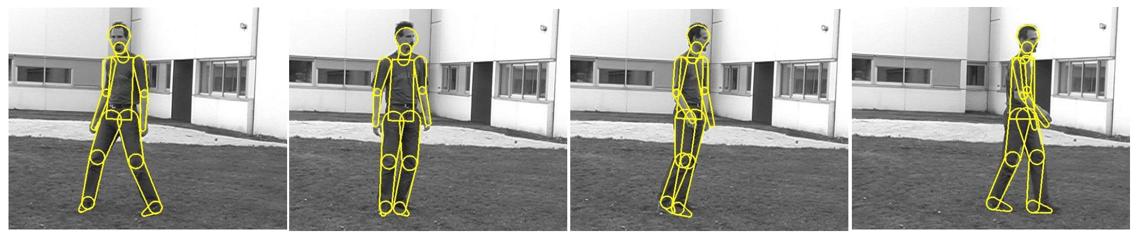
Top-down	vs	Bottom-up
2D	vs	3D
Static	vs	Dynamic



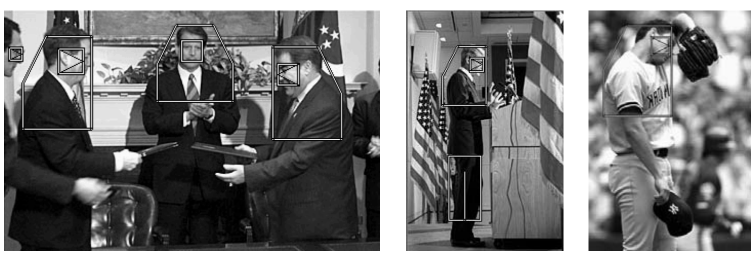
Felzenszwalb & Huttenlocher



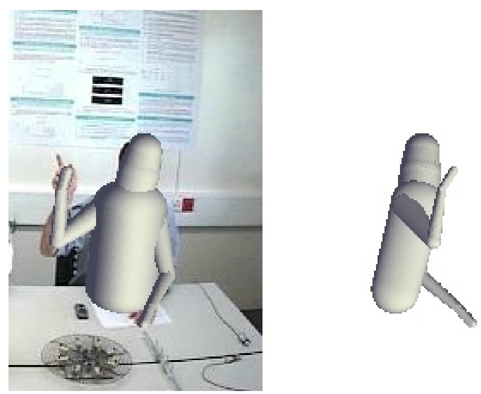
Ronfard, Schmid & Triggs, ECCV02



Agarwal & Triggs, ECCV04



Mikolajczyk, Schmid & Zisserman, ECCV04

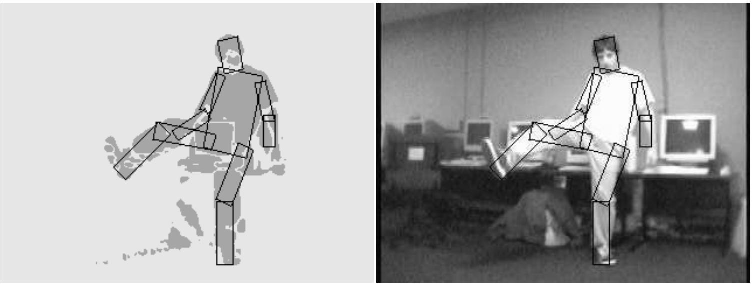


Lee & Cohen ECCV04

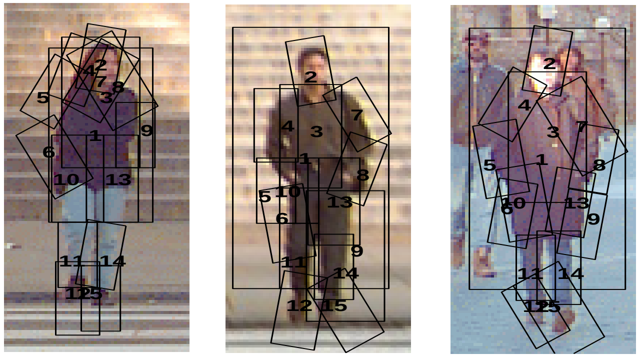
Sminchisescu & Triggs, CVPR03

Previous Work

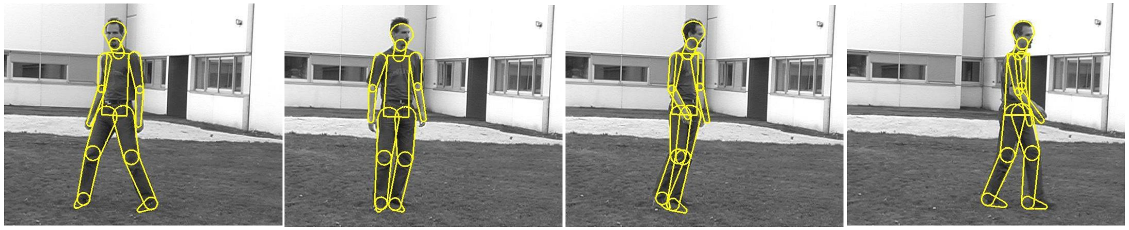
Top-down	vs	Bottom-up
2D	vs	3D
Static	vs	Dynamic



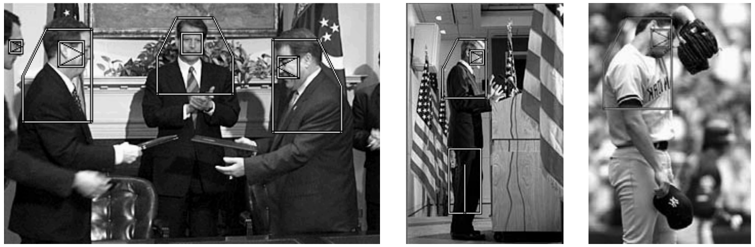
Felzenszwalb & Huttenlocher



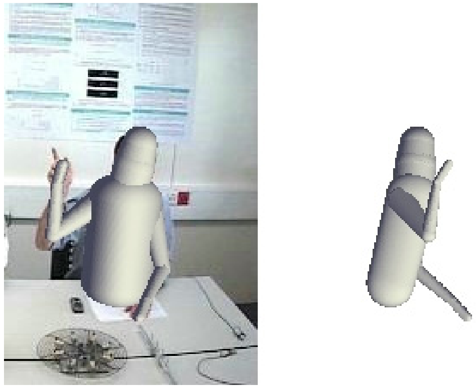
Ronfard, Schmid & Triggs, ECCV02



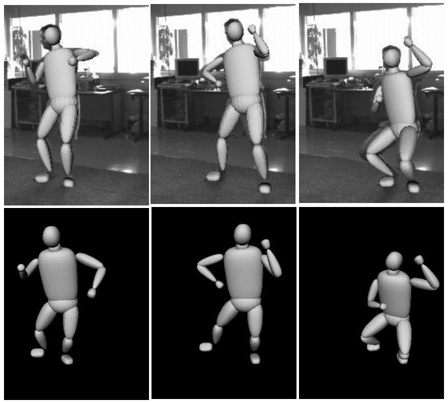
Agarwal & Triggs, ECCV04



Mikolajczyk, Schmid & Zisserman, ECCV04



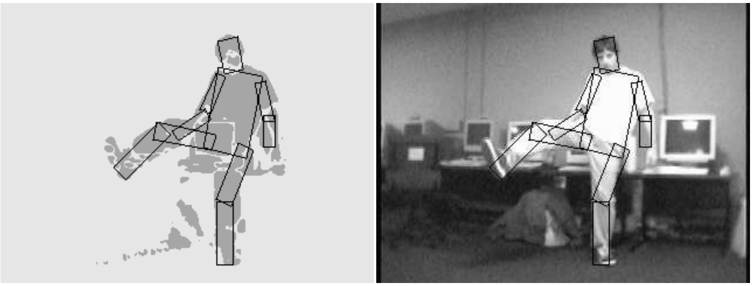
Lee & Cohen ECCV04



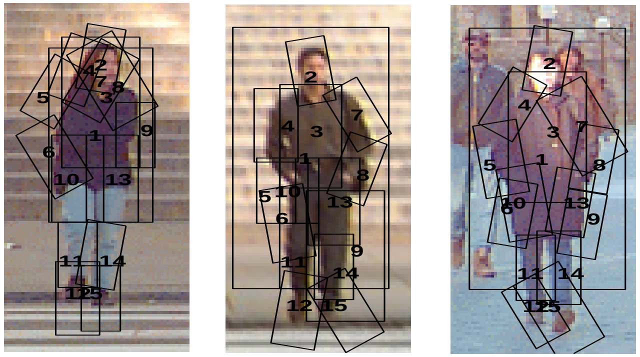
Sminchisescu & Triggs, CVPR03

Previous Work

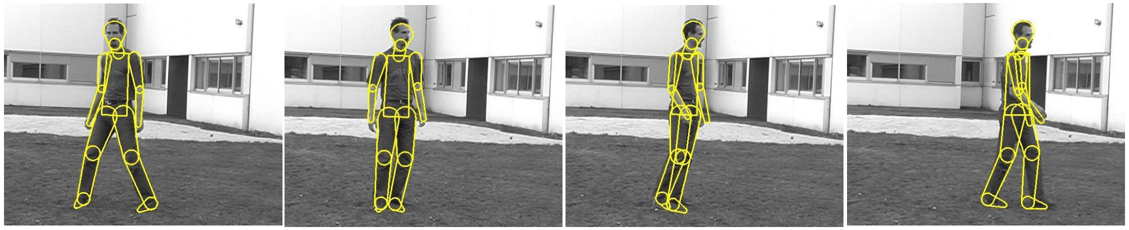
Top-down	vs	Bottom-up
2D	vs	3D
Static	vs	Dynamic



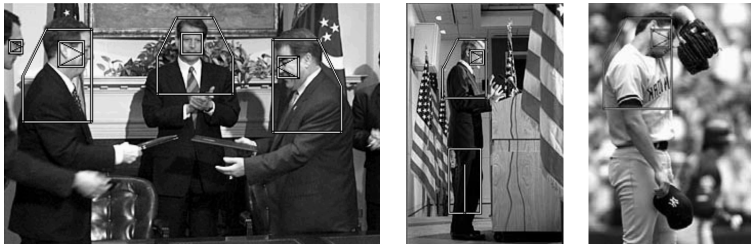
Felzenszwalb & Huttenlocher



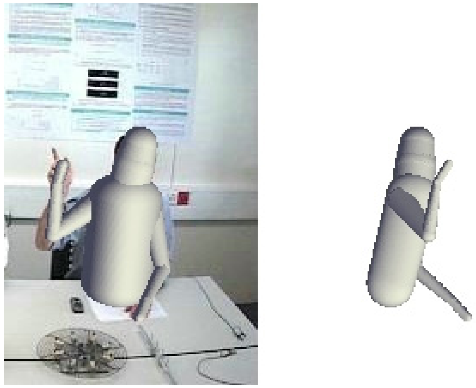
Ronfard, Schmid & Triggs, ECCV02



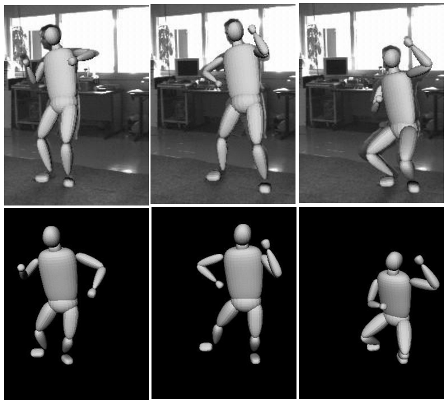
Agarwal & Triggs, ECCV04



Mikolajczyk, Schmid & Zisserman, ECCV04



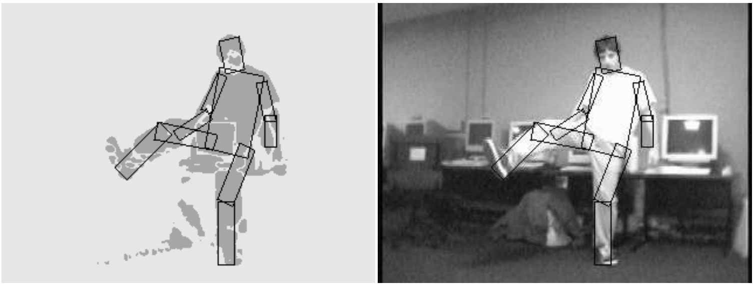
Lee & Cohen ECCV04



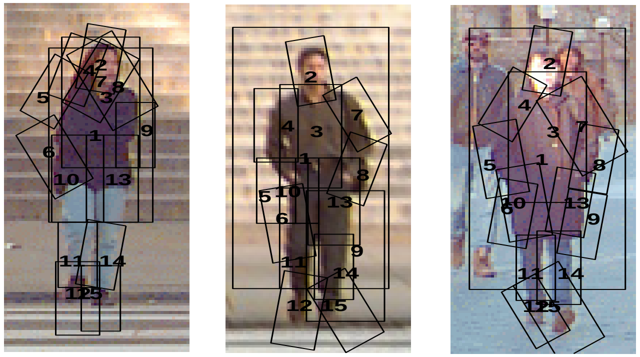
Sminchisescu & Triggs, CVPR03

Previous Work

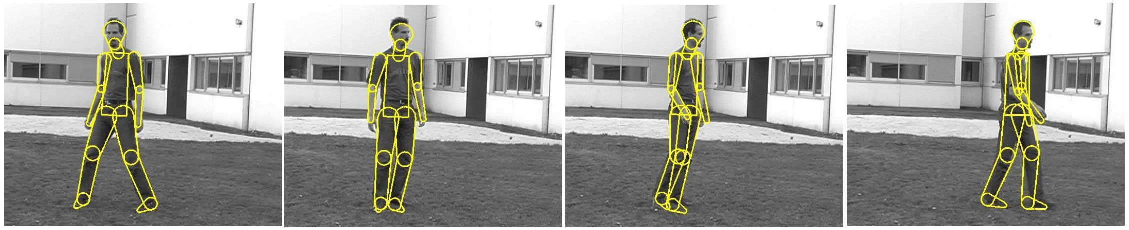
Top-down	vs	Bottom-up
2D	vs	3D
Static	vs	Dynamic



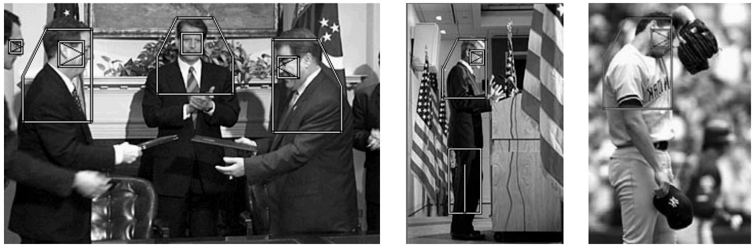
Felzenszwalb & Huttenlocher



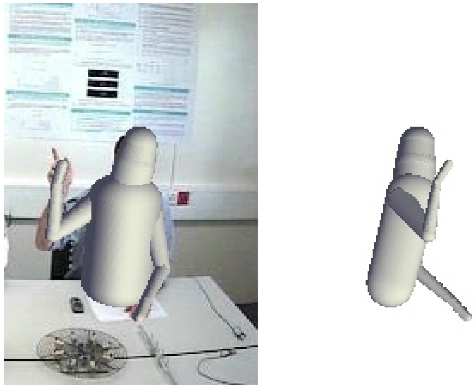
Ronfard, Schmid & Triggs, ECCV02



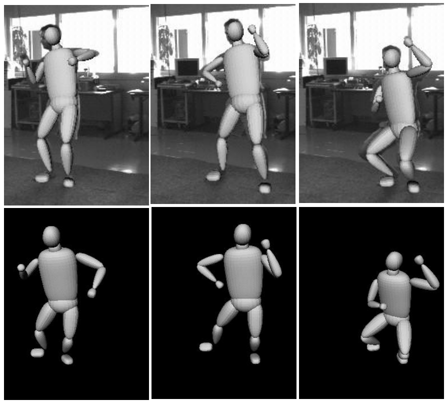
Agarwal & Triggs, ECCV04



Mikolajczyk, Schmid & Zisserman, ECCV04



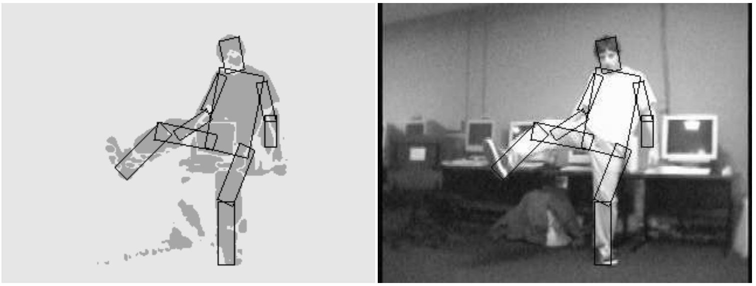
Lee & Cohen ECCV04



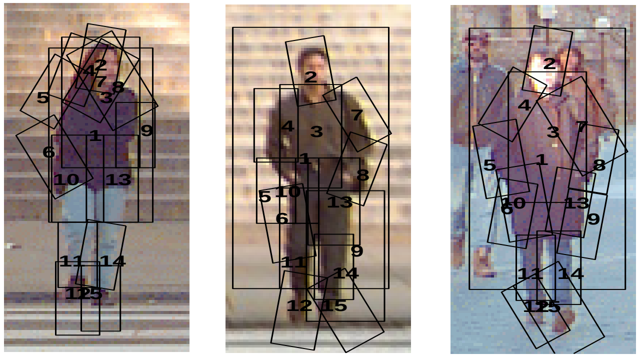
Sminchisescu & Triggs, CVPR03

Previous Work

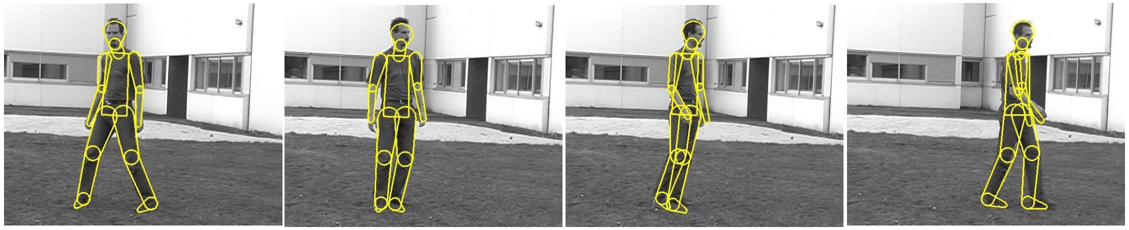
Top-down	vs	Bottom-up
2D	vs	3D
Static	vs	Dynamic



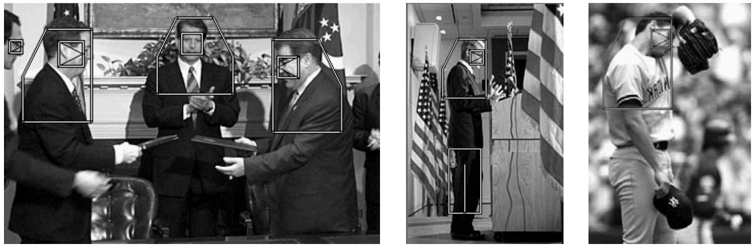
Felzenszwalb & Huttenlocher



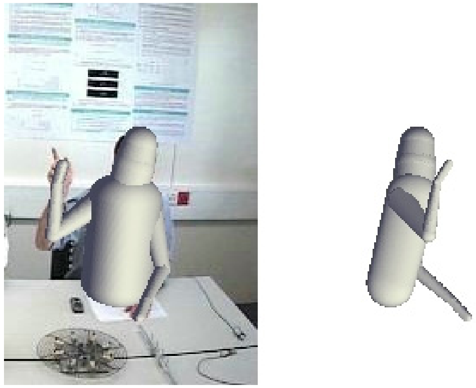
Ronfard, Schmid & Triggs, ECCV02



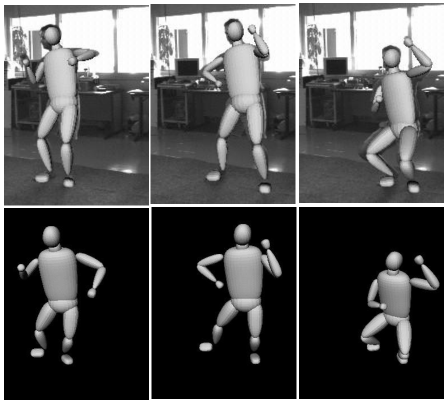
Agarwal & Triggs, ECCV04



Mikolajczyk, Schmid & Zisserman, ECCV04



Lee & Cohen ECCV04



Sminchisescu & Triggs, CVPR03

Outline: A Bayesian Approach

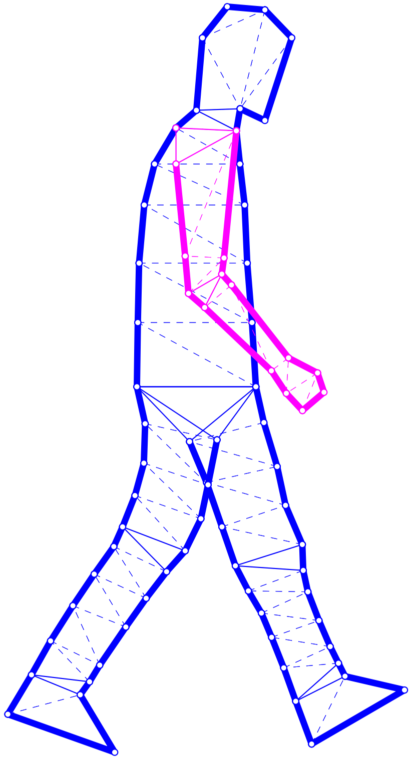
$p(\Omega)$	$+$	$p(\mathcal{I}\|\Omega)$	$\Rightarrow$	$P(\Omega\|\mathcal{I})$
Shape Prior		**Image Likelihood**		**Posterior Distrib.**

Outline: A Bayesian Approach

$p(\Omega)$	$+$	$p(\mathcal{I}\|\Omega)$	$\Rightarrow$	$P(\Omega\|\mathcal{I})$
Shape Prior		**Image Likelihood**		**Posterior Distrib.**

$p(\Omega)$ modeled as Bayes Net

- global & local deformation
- chain-like structure



Outline: A Bayesian Approach

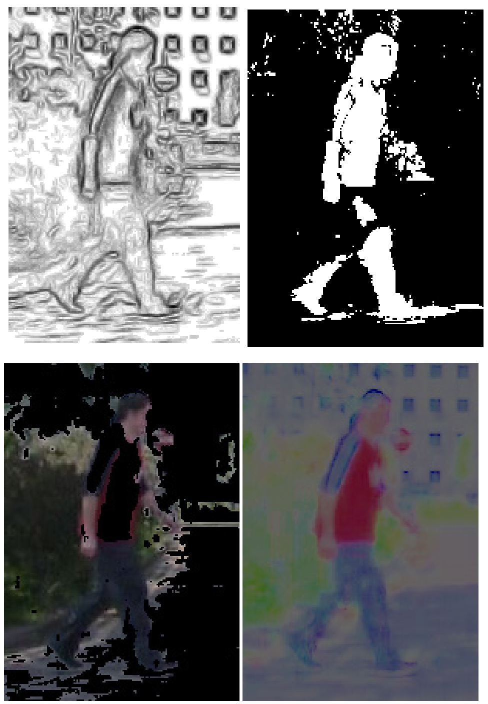
$p(\Omega)$	$+$	$p(\mathcal{I}\|\Omega)$	$\Rightarrow$	$P(\Omega\|\mathcal{I})$
Shape Prior		**Image Likelihood**		**Posterior Distrib.**

$p(\Omega)$ modeled as Bayes Net

- global & local deformation
- chain-like structure

$p(\mathcal{I}|\Omega)$ combines 4 types of cues

- edge, intensity & color based

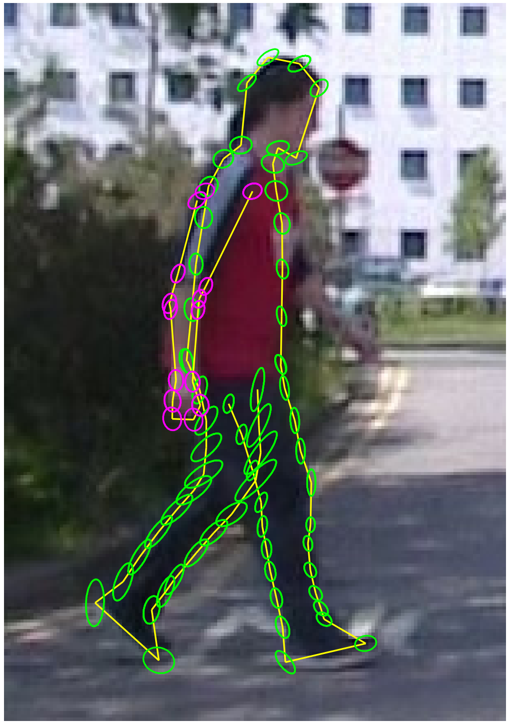


Outline: A Bayesian Approach

$p(\Omega)$	$+$	$p(\mathcal{I}\|\Omega)$	$\Rightarrow$	$P(\Omega\|\mathcal{I})$
Shape Prior		**Image Likelihood**		**Posterior Distrib.**

$p(\Omega)$ modeled as Bayes Net

- global & local deformation
- chain-like structure

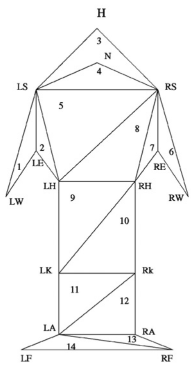


$p(\mathcal{I}|\Omega)$ combines 4 types of cues

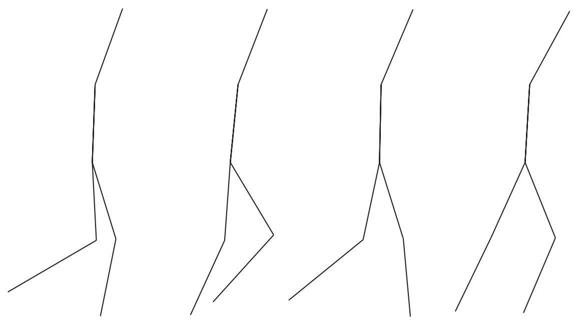
- edge, intensity & color based

$\tilde{\Omega} \sim p(\Omega|\mathcal{I})$ by Sequential MC

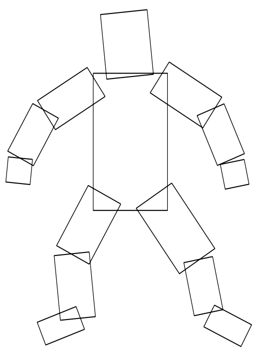
Shape = Joint Config + *Anthropometry*



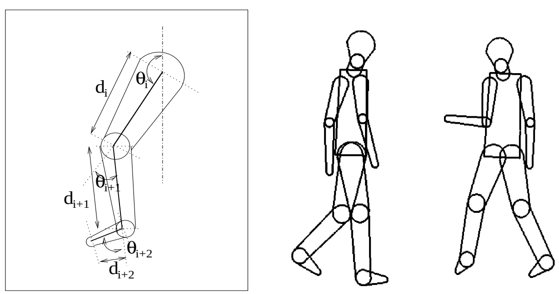
POINT



STICK FIGURE



PICTORIAL STRUCTURE

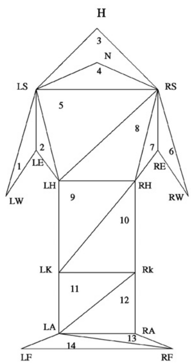


SCALED PRISMATIC MODEL

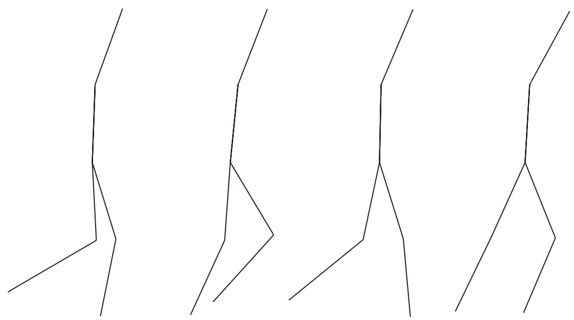
Slim, fat, pear, apple, chest, belly, straight, crooked, sexy ...

Typical part representations.

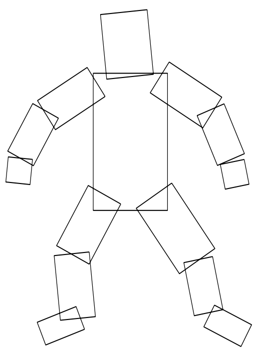
Shape = Joint Config + *Anthropometry*



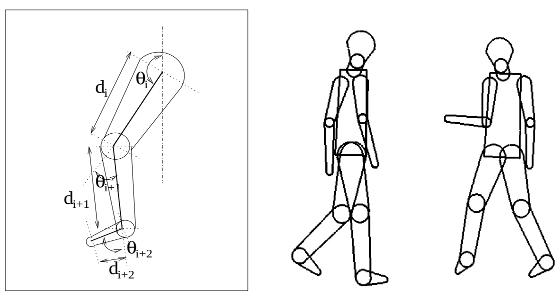
POINT



STICK FIGURE

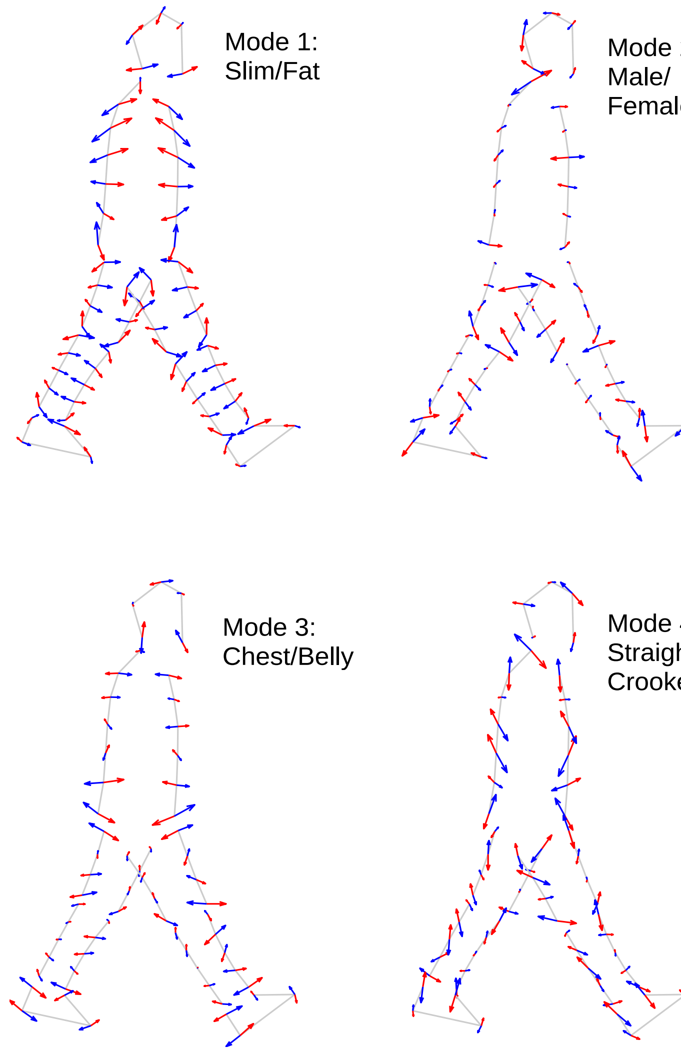


PICTORIAL STRUCTURE



SCALED PRISMATIC MODEL

Typical part representations.



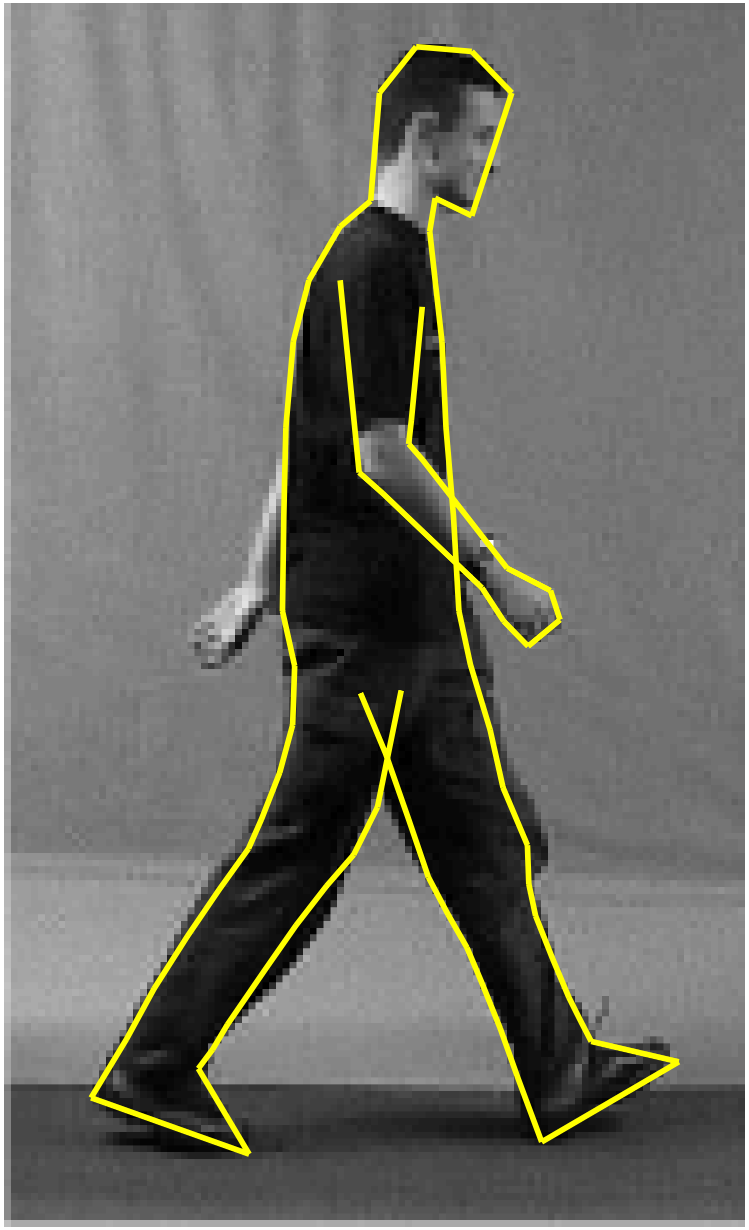
Var. across subject & gait cycle from hand-labeled dataset.

Shape Prior

Shape $\sim K$ landmark points

$$\mathbf{v}_{1:K} = \Omega = \{(x_k, y_k)\}_{k=1}^{K}$$

Prior $\sim p(\Omega), \Omega \in \mathcal{R}^{2K}$

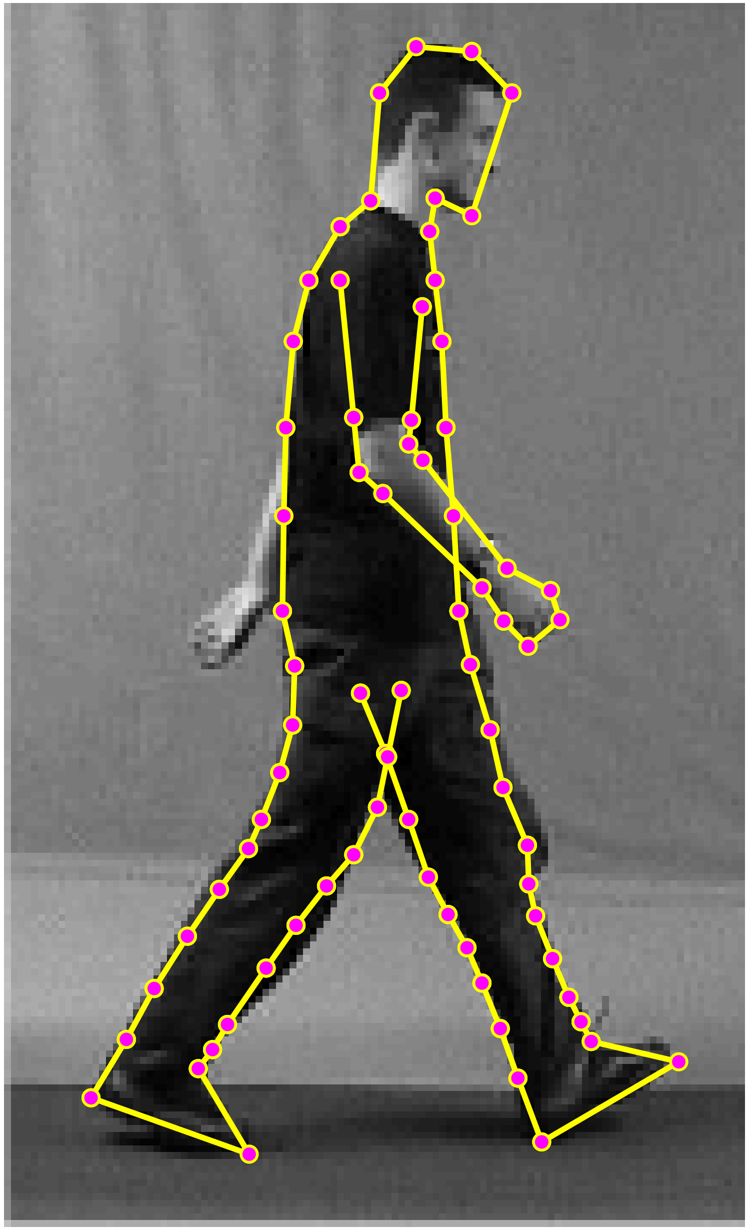


Shape Prior

Shape $\sim K$ landmark points

$$\mathbf{v}_{1:K} = \Omega = \{(x_k, y_k)\}_{k=1}^{K}$$

Prior $\sim p(\Omega), \Omega \in \mathcal{R}^{2K}$

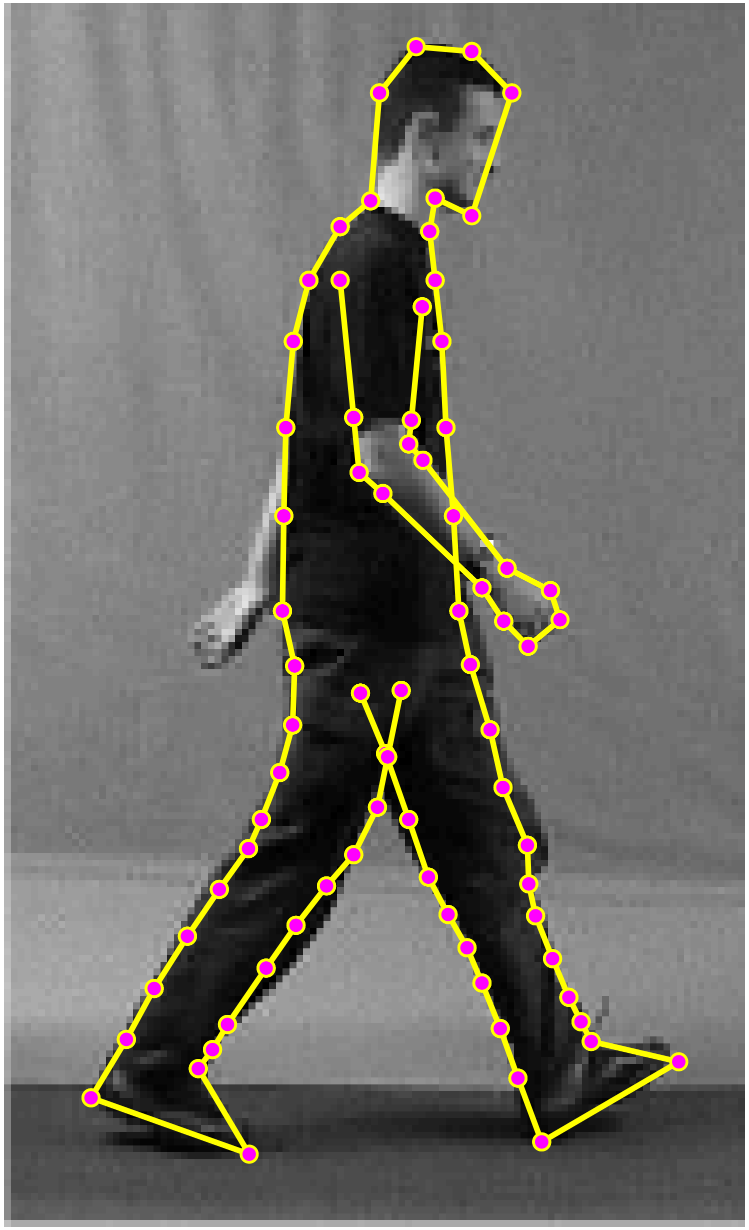


Shape Prior

Shape $\sim K$ landmark points

$$\mathbf{v}_{1:K} = \Omega = \{(x_k, y_k)\}_{k=1}^{K}$$

Prior $\sim p(\Omega), \Omega \in \mathcal{R}^{2K}$



Procrustes analysis on Ω?

- articulated motion
- efficient inference

Shape Prior

Shape $\sim K$ landmark points

$$\mathbf{v}_{1:K} = \Omega = \{(x_k, y_k)\}_{k=1}^{K}$$

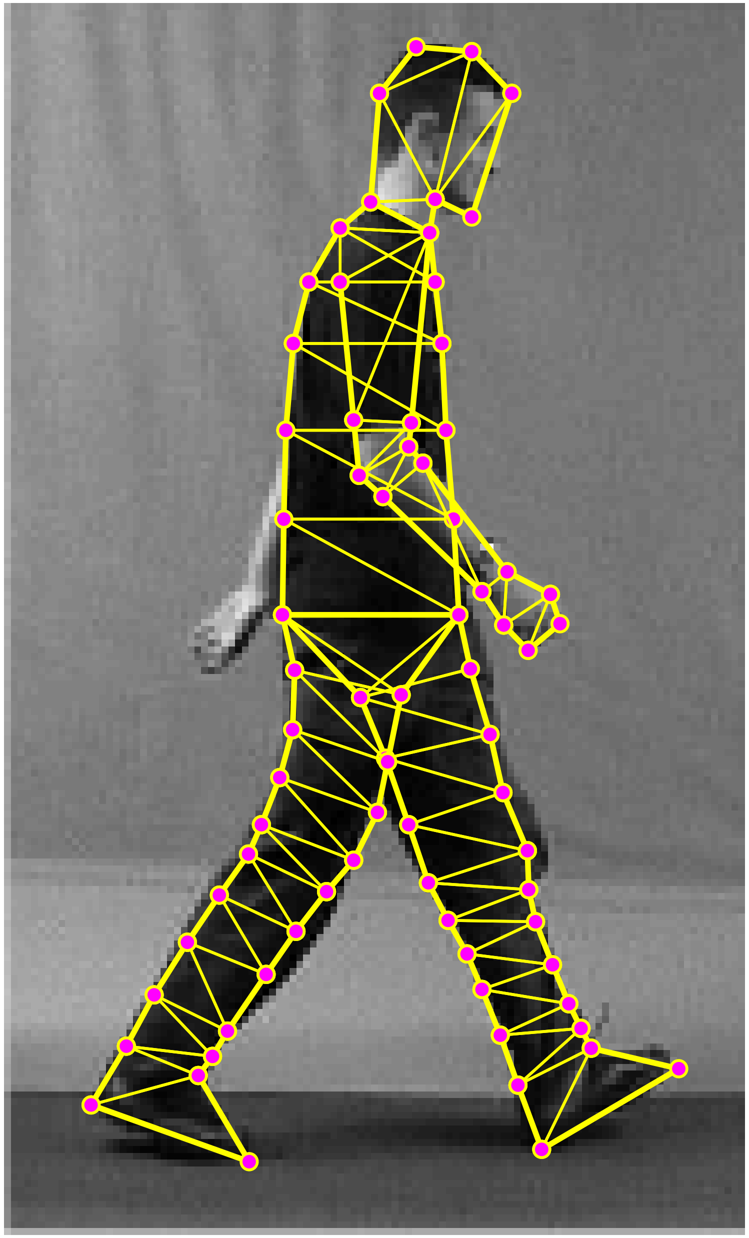
Prior $\sim p(\Omega), \Omega \in \mathcal{R}^{2K}$

Shape as Bayes network by

- triangulation
- cond. independence

$$p(\mathbf{v}_{1:K}) = \prod_k p(\mathbf{v}_k | \mathbf{v}_{1:k-1})$$

$$= \prod_k p(\mathbf{v}_k | \text{ parent of } \mathbf{v}_k)$$



Shape Prior (cont.)

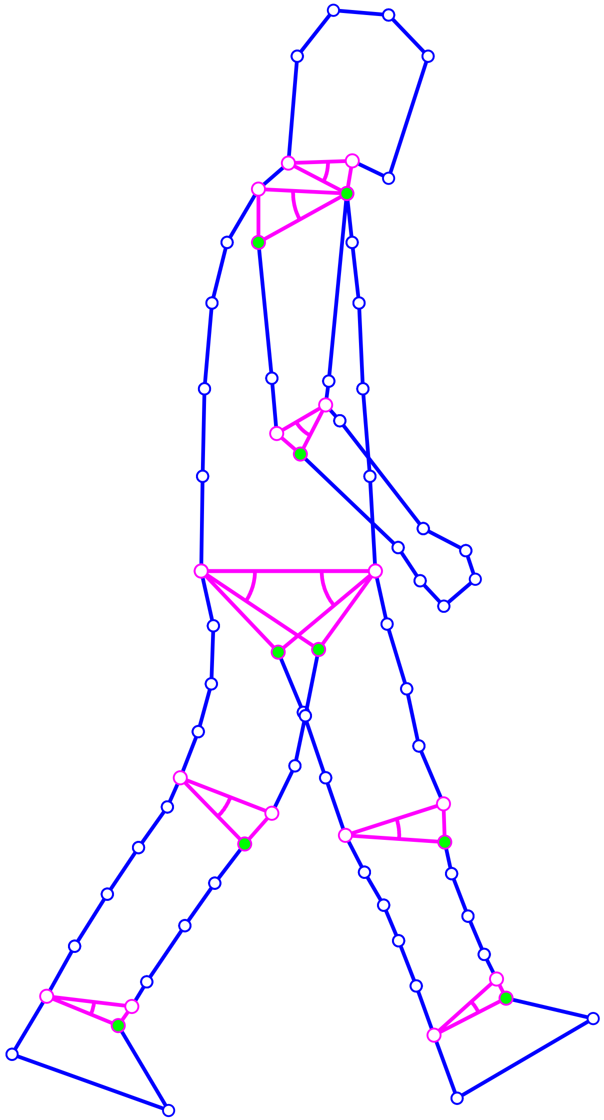
Two deformation mechanisms to specify $p(\mathbf{v}_k|\mathbf{v}_{1:k-1})$.

I. Rotation of Joints

Joint vertex $\mathbf{v}$ and its parent edge $\mathbf{e}$ define joint angle θ.

$$\boxed{\mathbf{v} = \rho \cdot Rot(\theta) \cdot \vec{\mathbf{e}}} + \boxed{p(\theta_{1:9}),\ p(\rho_{1:9})}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{e}_k, \theta\text{'s visited})}$$



Shape Prior (cont.)

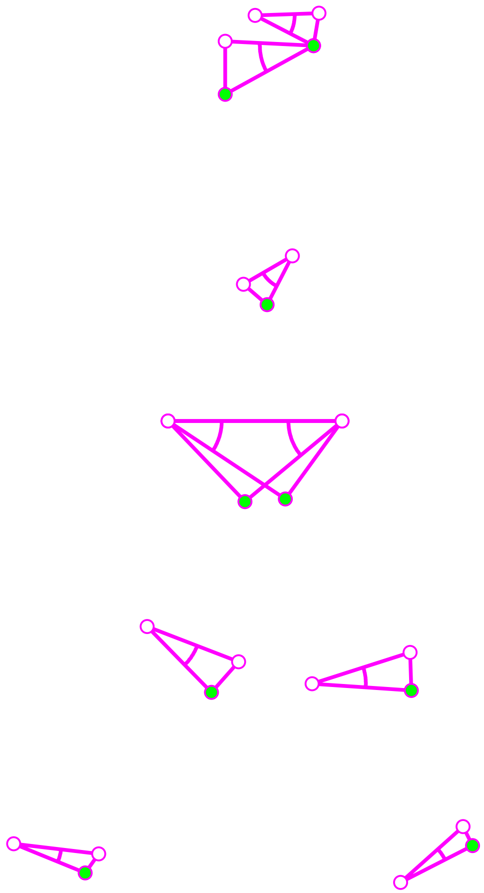
Two deformation mechanisms to specify $p(\mathbf{v}_k|\mathbf{v}_{1:k-1})$.

I. Rotation of Joints

Joint vertex $\mathbf{v}$ and its parent edge $\mathbf{e}$ define joint angle θ.

$$\boxed{\mathbf{v} = \rho \cdot Rot(\theta) \cdot \vec{\mathbf{e}}} + \boxed{p(\theta_{1:9}),\ p(\rho_{1:9})}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{e}_k, \theta\text{'s visited})}$$



Shape Prior (cont.)

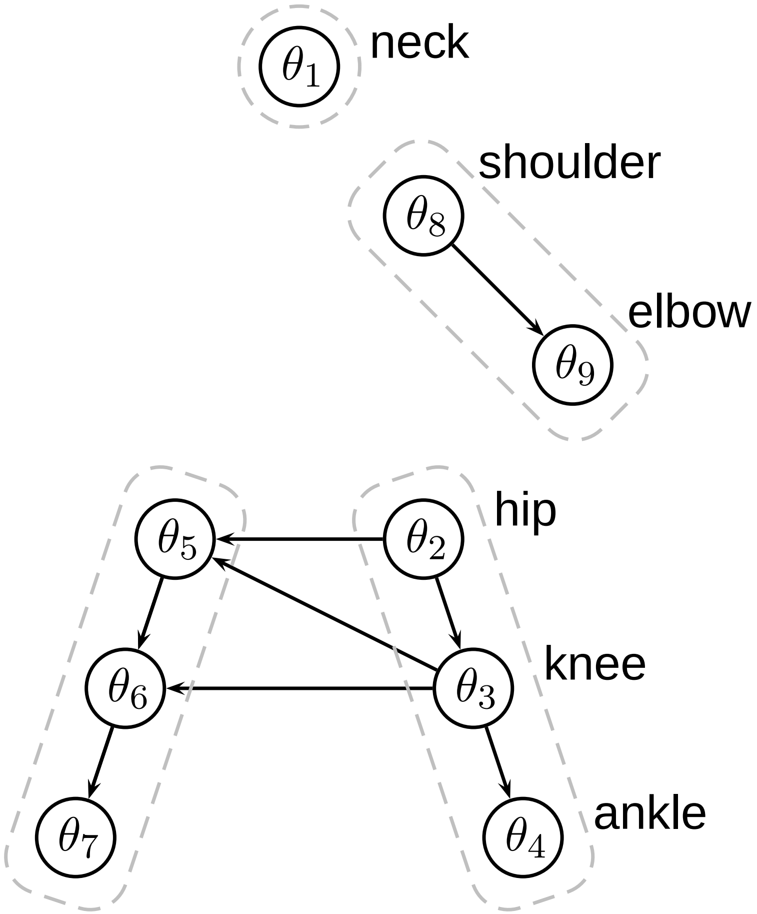
Two deformation mechanisms to specify $p(\mathbf{v}_k|\mathbf{v}_{1:k-1})$.

I. Rotation of Joints

Joint vertex $\mathbf{v}$ and its parent edge $\mathbf{e}$ define joint angle θ.

$$\boxed{\mathbf{v} = \rho \cdot Rot(\theta) \cdot \vec{\mathbf{e}}} + \boxed{p(\theta_{1:9}),\ p(\rho_{1:9})}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{e}_k, \theta\text{'s visited})}$$



$$p(\Theta) = p(\theta_{1:9})$$

Shape Prior (cont.)

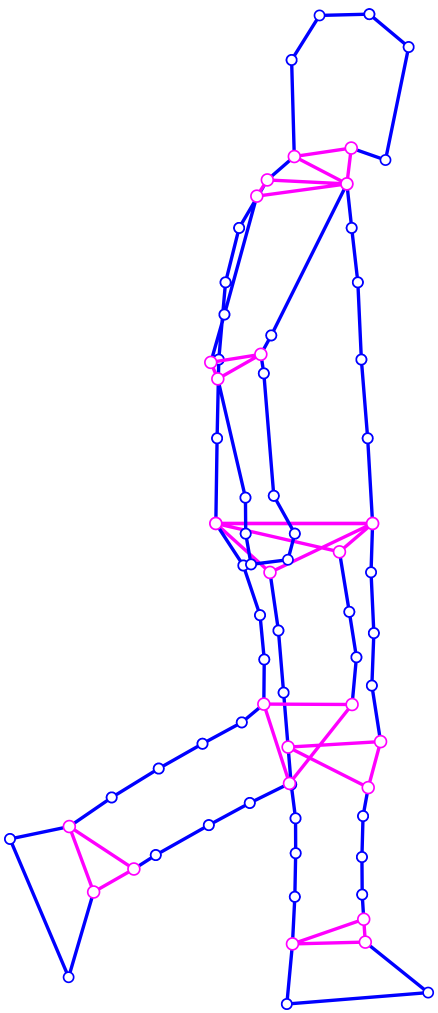
Two deformation mechanisms to specify $p(\mathbf{v}_k|\mathbf{v}_{1:k-1})$.

I. Rotation of Joints

Joint vertex $\mathbf{v}$ and its parent edge $\mathbf{e}$ define joint angle θ.

$$\boxed{\mathbf{v} = \rho \cdot Rot(\theta) \cdot \vec{\mathbf{e}}} + \boxed{p(\theta_{1:9}),\ p(\rho_{1:9})}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{e}_k, \theta\text{'s visited})}$$



$$\tilde{\Theta} \sim p(\Theta)$$

Shape Prior (cont.)

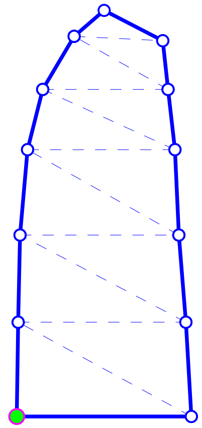
Two deformation mechanisms to specify $p(\mathbf{v}_k|\mathbf{v}_{1:k-1})$.

I. Rotation of Joints

Joint vertex $\mathbf{v}$ and its parent edge $\mathbf{e}$ define joint angle θ.

$$\boxed{\mathbf{v} = \rho \cdot Rot(\theta) \cdot \vec{\mathbf{e}}} + \boxed{p(\theta_{1:9}),\ p(\rho_{1:9})}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{e}_k, \theta\text{'s visited})}$$



II. Local Nonrigid Motion

Parent triangle $\mathbf{g} \longrightarrow$ affine trans. (A, b).

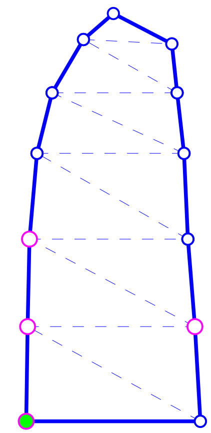
$$\boxed{\mathbf{v} = (A \cdot \bar{\mathbf{v}} + b) \circ \mathbf{n}} + \boxed{p(\mathbf{n}_k)}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{g}_k)}$$

Shape Prior (cont.)

Two deformation mechanisms to specify $p(\mathbf{v}_k|\mathbf{v}_{1:k-1})$.

I. Rotation of Joints



Joint vertex $\mathbf{v}$ and its parent edge $\mathbf{e}$ define joint angle θ.

$$\boxed{\mathbf{v} = \rho \cdot Rot(\theta) \cdot \vec{\mathbf{e}}} + \boxed{p(\theta_{1:9}),\ p(\rho_{1:9})}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{e}_k, \theta\text{'s visited})}$$

II. Local Nonrigid Motion

Parent triangle $\mathbf{g} \longrightarrow$ affine trans. (A, b).

$$\boxed{\mathbf{v} = (A \cdot \bar{\mathbf{v}} + b) \circ \mathbf{n}} + \boxed{p(\mathbf{n}_k)}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{g}_k)}$$

Shape Prior (cont.)

Two deformation mechanisms to specify $p(\mathbf{v}_k|\mathbf{v}_{1:k-1})$.

I. Rotation of Joints

Joint vertex $\mathbf{v}$ and its parent edge $\mathbf{e}$ define joint angle θ.

$$\boxed{\mathbf{v} = \rho \cdot Rot(\theta) \cdot \vec{\mathbf{e}}} + \boxed{p(\theta_{1:9}),\ p(\rho_{1:9})}$$

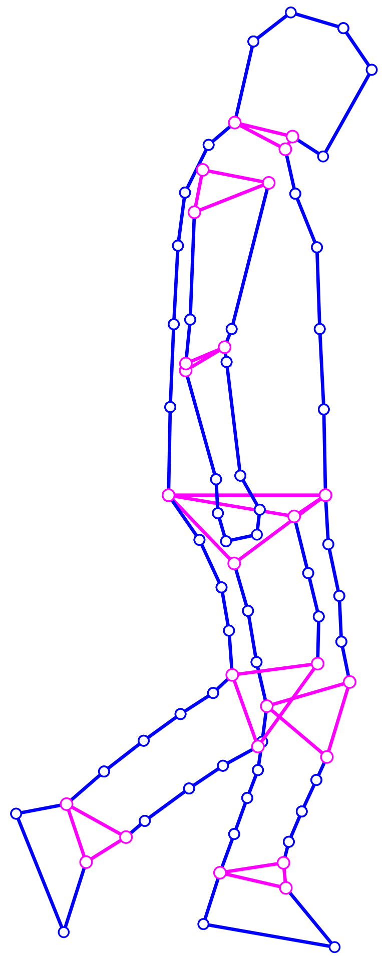
$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{e}_k, \theta\text{'s visited})}$$

II. Local Nonrigid Motion

Parent triangle $\mathbf{g} \rightarrow$ affine trans. (A, b).

$$\boxed{\mathbf{v} = (A \cdot \bar{\mathbf{v}} + b) \circ \mathbf{n}} + \boxed{p(\mathbf{n}_k)}$$

$$\Rightarrow \boxed{p(\mathbf{v}_k|\mathbf{v}_{1:k-1}) = p(\mathbf{v}_k|\mathbf{g}_k)}$$



$$\tilde{\Omega} \sim p(\Omega)$$

Image Likelihood

Cover $\mathbf{v}_{1:K}$ by set of clusters $\mathcal{C}$.

- cluster size from 2 to 6

Define $\phi(\mathbf{v}_C) \propto p(f_C|\mathbf{v}_C), C \in \mathcal{C}$.

- independence assumption

Four types of image features f_C

- Edge Gradient Map
- Foreground Mask
- Skin Color Mask
- Appearance Consistency

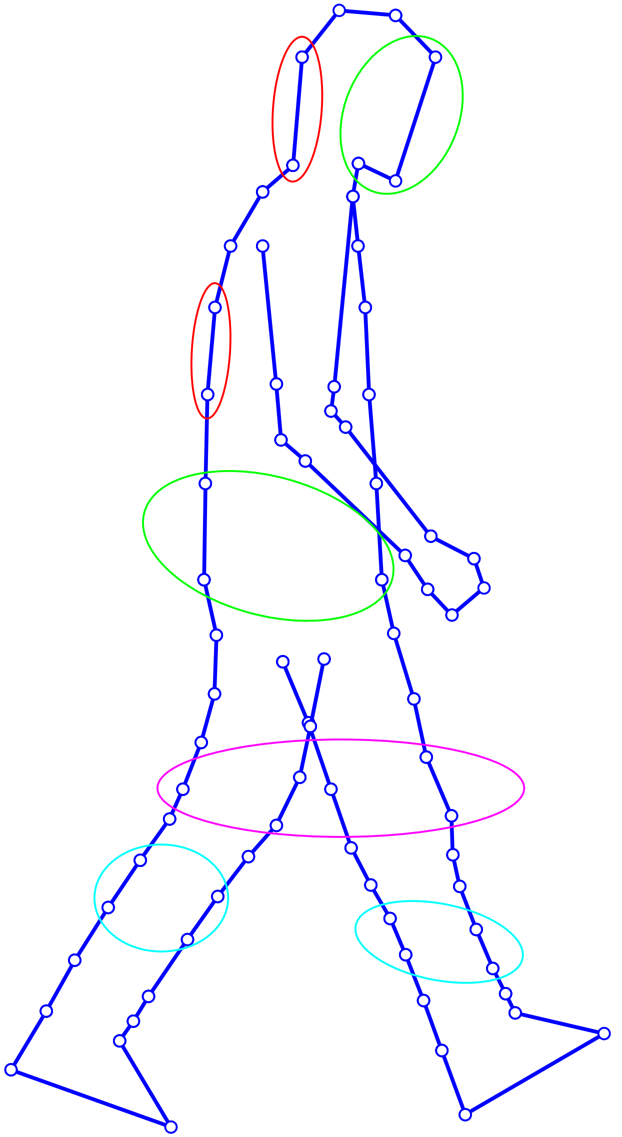


Image Likelihood

Cover $\mathbf{v}_{1:K}$ by set of clusters $\mathcal{C}$.

- cluster size from 2 to 6

Define $\phi(\mathbf{v}_C) \propto p(f_C|\mathbf{v}_C), C \in \mathcal{C}$.

- independence assumption

Four types of image features f_C

- Edge Gradient Map
- Foreground Mask
- Skin Color Mask
- Appearance Consistency

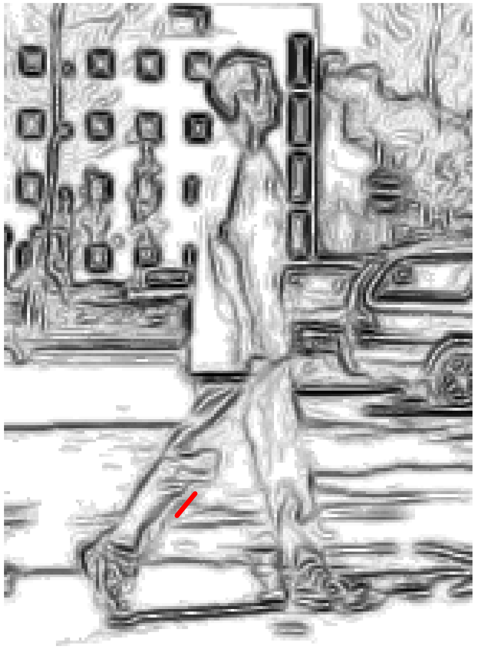


Image Likelihood

Cover $\mathbf{v}_{1:K}$ by set of clusters $\mathcal{C}$.

- cluster size from 2 to 6

Define $\phi(\mathbf{v}_C) \propto p(f_C|\mathbf{v}_C), C \in \mathcal{C}$.

- independence assumption

Four types of image features f_C

- Edge Gradient Map
- Foreground Mask
- Skin Color Mask
- Appearance Consistency

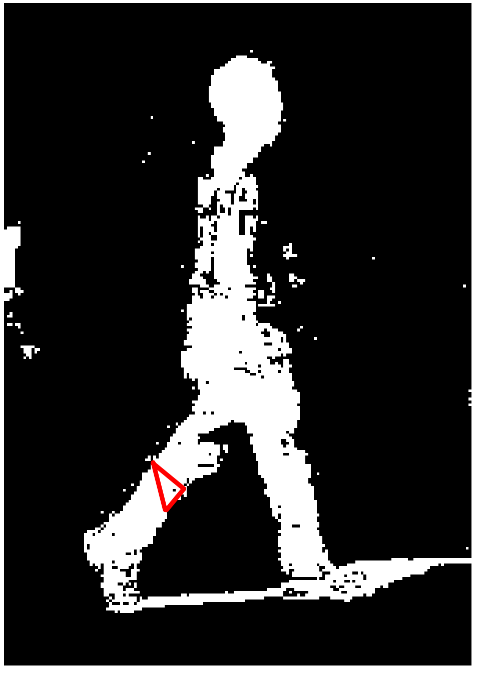


Image Likelihood

Cover $\mathbf{v}_{1:K}$ by set of clusters $\mathcal{C}$.

- cluster size from 2 to 6

Define $\phi(\mathbf{v}_C) \propto p(f_C|\mathbf{v}_C), C \in \mathcal{C}$.

- independence assumption

Four types of image features f_C

- Edge Gradient Map
- Foreground Mask
- Skin Color Mask
- Appearance Consistency

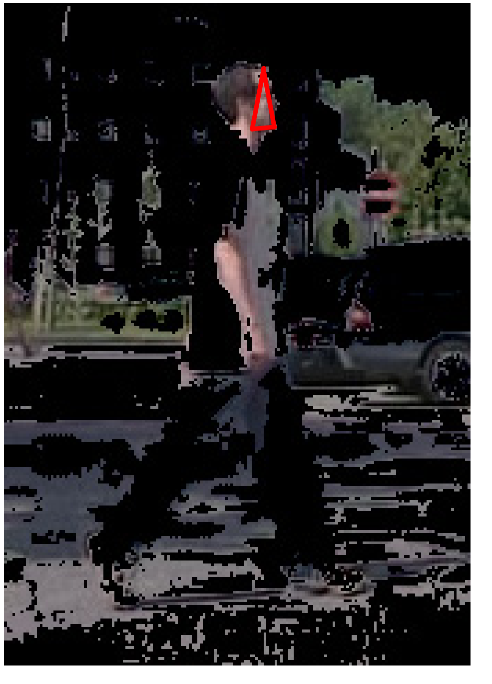


Image Likelihood

Cover $\mathbf{v}_{1:K}$ by set of clusters $\mathcal{C}$.

- cluster size from 2 to 6

Define $\phi(\mathbf{v}_C) \propto p(f_C|\mathbf{v}_C), C \in \mathcal{C}$.

- independence assumption

Four types of image features f_C

- Edge Gradient Map
- Foreground Mask
- Skin Color Mask
- Appearance Consistency

Posterior Distribution

$$p(\mathbf{v}_{1:K}|\mathcal{I}) \;\propto\; \underbrace{p(\Theta)\prod_{k\in\mathcal{J}} p(\rho_k)\prod_{k\notin\mathcal{J}} p(\mathbf{n}_k)}_{\text{Prior}} \;\underbrace{\prod_{C\in\mathfrak{C}} \phi(\mathbf{v}_C)}_{\text{Likelihood}}$$

$\mathcal{J}$	index set of joint vertices, $\lvert\mathcal{J}\rvert = 9$
Θ	joint angles, $\theta_{1:9}$
ρ_k	radius scaling in polar coordinates
$\mathbf{n}_k$	residual noise in Cartesian coordinates
$\mathfrak{C}$	index set of clusters
$\phi(\mathbf{v}_C)$	likelihood potential, $\propto p(f_C\lvert\mathbf{v}_C)$

Posterior Distribution

$$p(\mathbf{v}_{1:K}|\mathcal{I}) \;\propto\; \underbrace{p(\Theta)\prod_{k\in\mathcal{J}} p(\rho_k)\prod_{k\notin\mathcal{J}} p(\mathbf{n}_k)}_{\text{Prior}}\; \underbrace{\prod_{C\in\mathfrak{C}} \phi(\mathbf{v}_C)}_{\text{Likelihood}}$$

Inference Tools Available

- DP (Dynamic Programming)
- MF (Mean Field)
- BP (Belief Propagation)
- MCMC (Markov Chain Monte Carlo)
- SMC (Sequential Monte Carlo)

Posterior Distribution

$$p(\mathbf{v}_{1:K}|\mathcal{I}) \propto \underbrace{p(\Theta) \prod_{k \in \mathcal{J}} p(\rho_k) \prod_{k \notin \mathcal{J}} p(\mathbf{n}_k)}_{\text{Prior}} \underbrace{\prod_{C \in \mathfrak{C}} \phi(\mathbf{v}_C)}_{\text{Likelihood}}$$

Inference Tools

- DP ☹
- MF ☹
- BP ☹
- MCMC
- SMC

Negative

- continuous variables
- clique size as high as 9
- complex likelihood

Posterior Distribution

$$p(\mathbf{v}_{1:K}|\mathcal{I}) \propto \underbrace{p(\Theta)\prod_{k\in\mathcal{J}} p(\rho_k)\prod_{k\notin\mathcal{J}} p(\mathbf{n}_k)}_{\text{Prior}} \underbrace{\prod_{C\in\mathfrak{C}} \phi(\mathbf{v}_C)}_{\text{Likelihood}}$$

Inference Tools

- DP ☹
- MF ☹
- BP ☹
- MCMC
- SMC ☺

Negative

- continuous variables
- clique size as high as 9
- complex likelihood

Positive

- chain-like model structure

Sequential Importance Sampling

Proposal Distribution

$$\pi_k = \pi_{k-1} \cdot p\big(\mathbf{v}_k \big| \mathbf{v}_{1:k-1}\big)$$

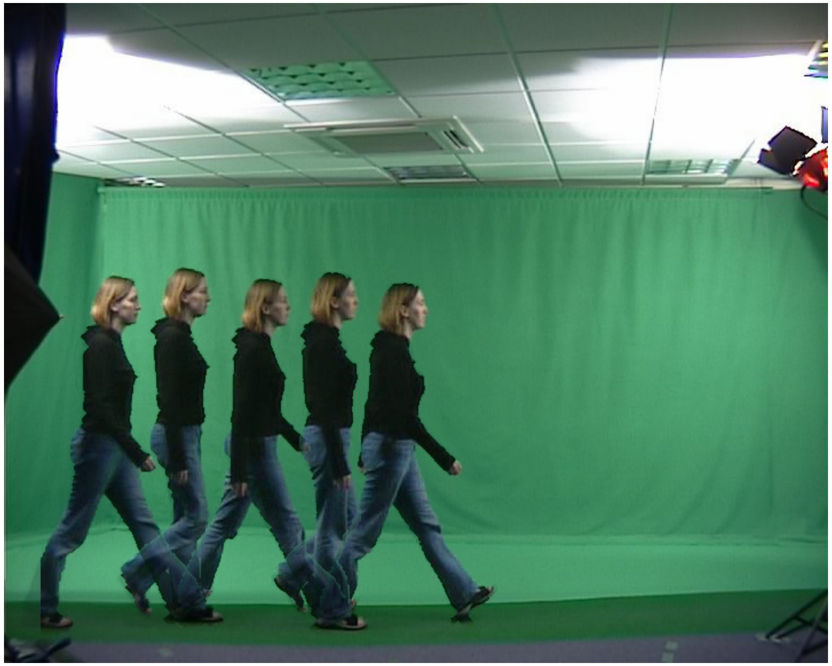
$$= \pi_{k-1} \cdot \begin{cases} p\big(\mathbf{v}_k \big| \mathbf{e}_k, \Theta_{k-1}\big) & \text{if } \mathbf{v}_k \text{ is joint vtx} \\ p\big(\mathbf{v}_k \big| \mathbf{g}_k\big) & \text{otherwise} \end{cases}$$

Importance Weights

$$w_k \propto w_{k-1} \cdot \prod_{C \in \mathcal{C}_k} \phi\big(\mathbf{v}_C\big)$$

Resampling Scheme: Stratified resampling

Experimental Dataset



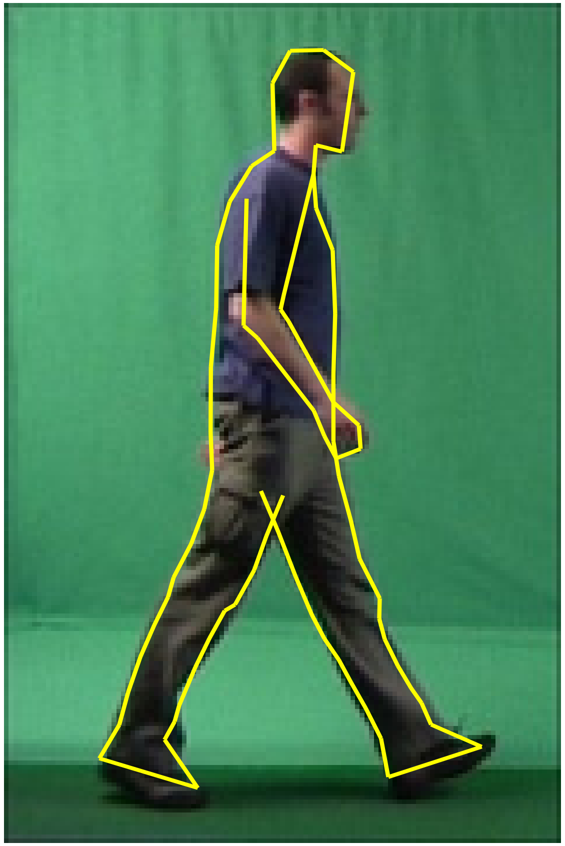
Training

- indoor, green backdrop, controlled lighting
- 28 subj, 112 seq, 3126 frm

Test

- outdoor, cluttered bg, natural lighting
- 10 subj, 10 seq, 963 frm
- 50 frames selected w/ ground truth

Bootstrap with Uniform Prior



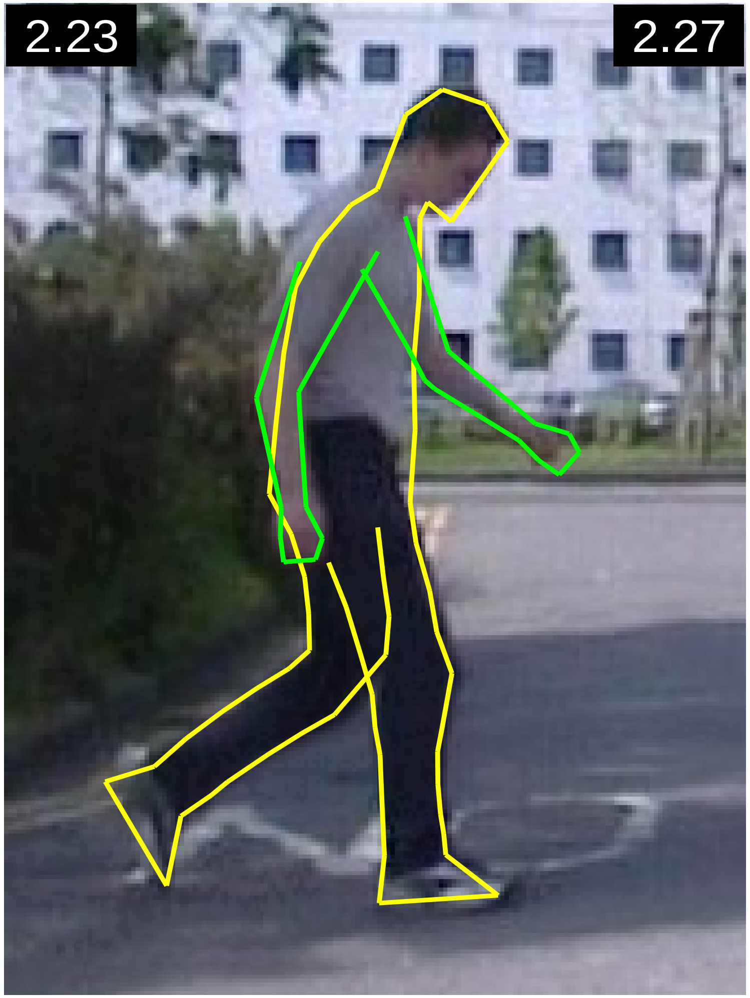
- Hand-label one frame of the indoor data.
- Fit training set using uniform $p(\Theta), p(\rho), p(\mathbf{n})$.
- Learned prior $\Rightarrow$ more challenging seqs.

Visualizing SMC Inference

Marginal distrib. of each vertex is summarized by its covariance ellipse (error ellipse).

Quantitative Evaluation on 50 Frames

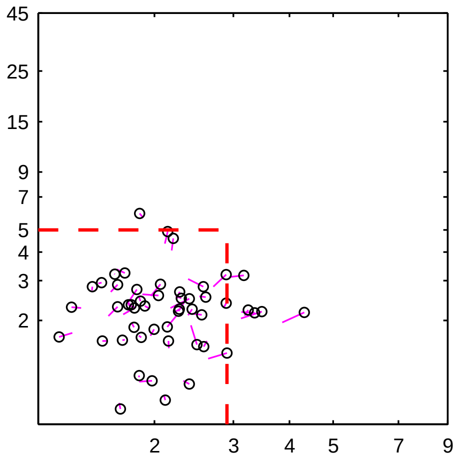
Body Error | Arm Error



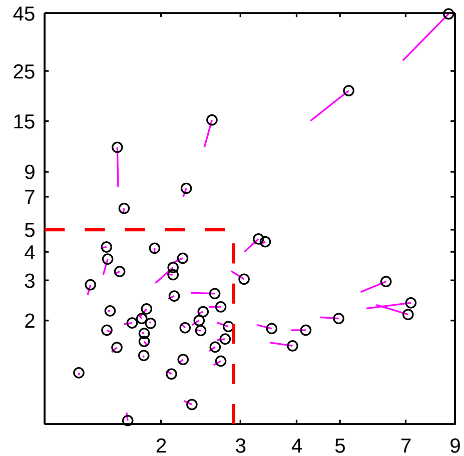
$\phi_e \phi_f \phi_s \phi_c$	Chamfer Distance	
	Body	Arm
● ○ ○ ○	4.00 ±3.52	4.41 ±4.08
○ ● ○ ○	2.53 ±0.91	6.25 ±5.95
○ ● ● ●	2.19 ±0.61	2.36 ±0.94
● ○ ● ●	2.77 ±1.62	4.13 ±6.83
● ● ○ ●	2.00 ±0.59	2.96 ±1.51
● ● ● ○	2.02 ±0.53	2.25 ±1.30
● ● ● ●	1.87 ±0.42	2.18 ±0.99

- error unit in pixels
- avg height ~ 200 pixels

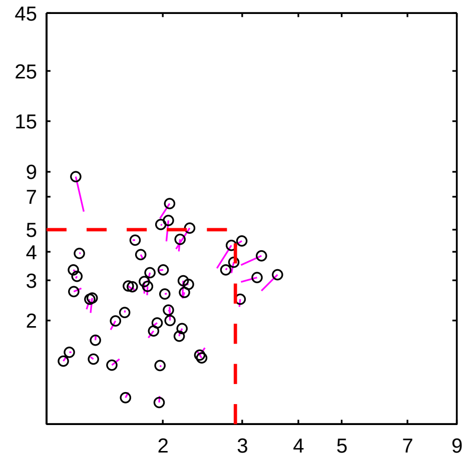
Each image cue has a role to play



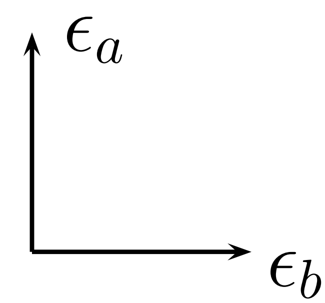
(a) ~~E~~, F, S, C
(+0.32,+0.18)



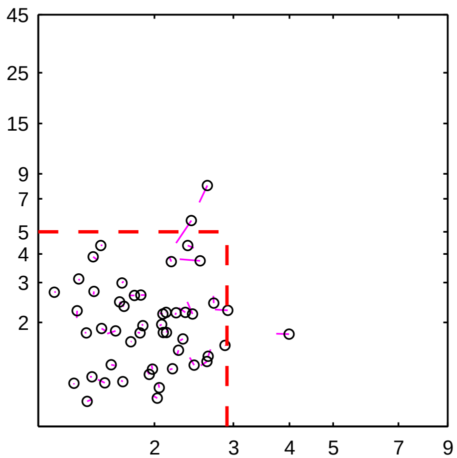
(b) E, ~~F~~, S, C
(+0.90,+1.95)



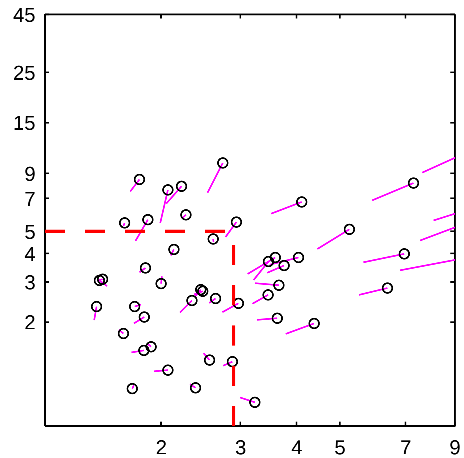
(c) E, F, ~~S~~, C
(+0.13,+0.78)



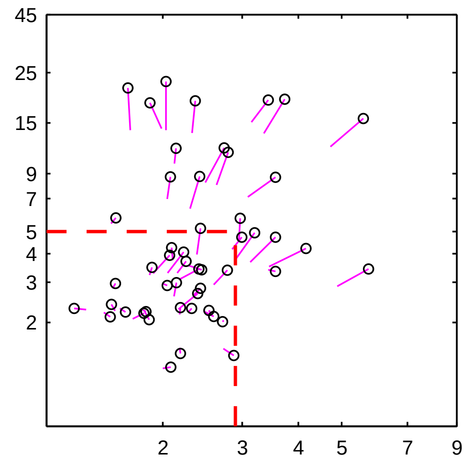
$(\Delta\bar{\epsilon}_b, \Delta\bar{\epsilon}_a)$



(d) E, F, S, ~~C~~
(+0.15,+0.07)



(e) E, ~~F~~, ~~S~~, ~~C~~
(+2.13,+2.23)



(f) ~~E~~, F, ~~S~~, ~~C~~
(+0.66,+4.07)

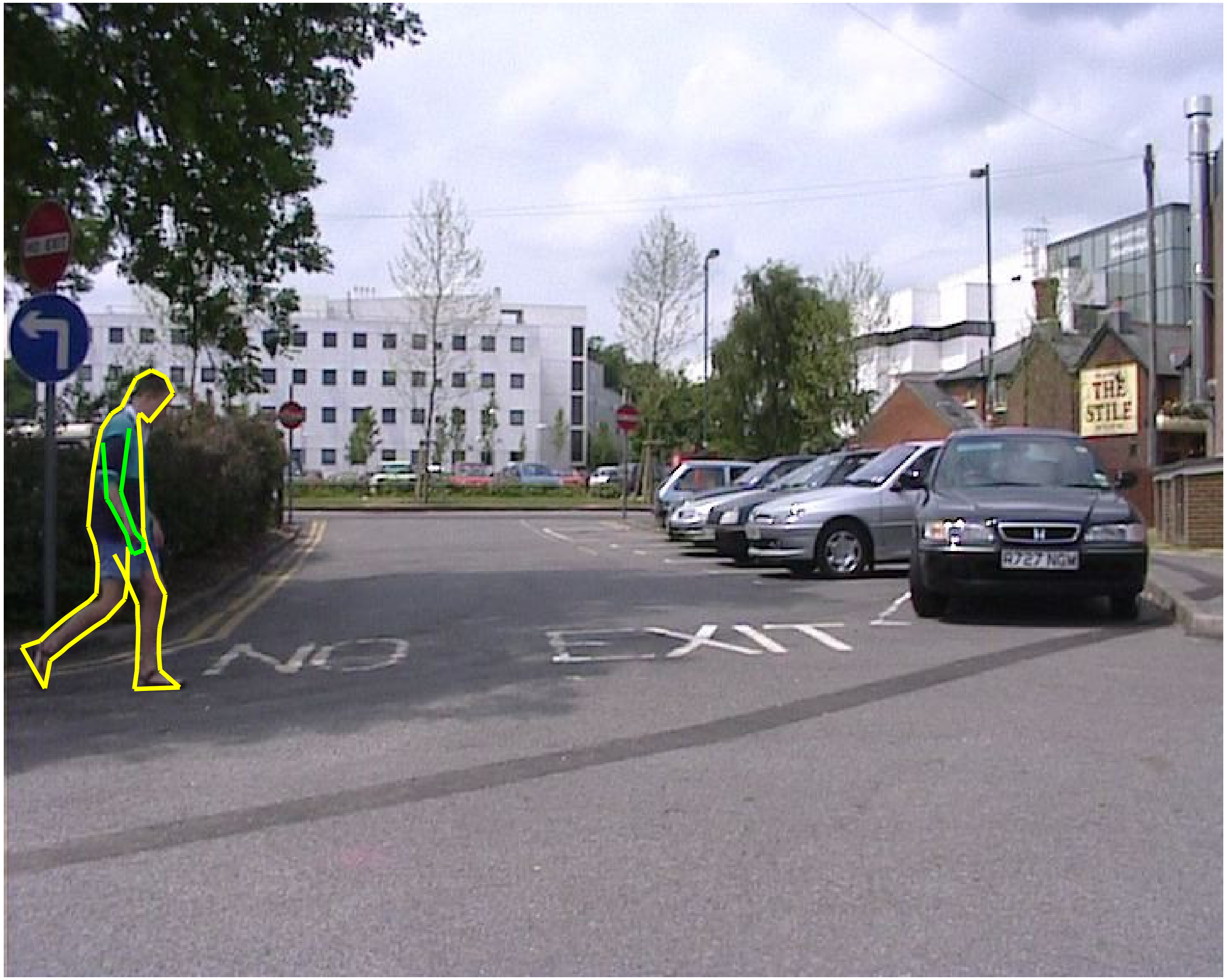
Edge

Foreground

Skin

Consistency

Performance on Video Sequences



01
02
03
04
05
06
07
08
09
10

Plotted are the posterior means.
Each frame is matched independently.

Summary

Contributions

- Novel Prob. Representation of Body Shapes
- Global Articulation $+$ *Local Deformation*
- Multiple-Cue Integration
- Efficient Spatial Inference via Sequential MC

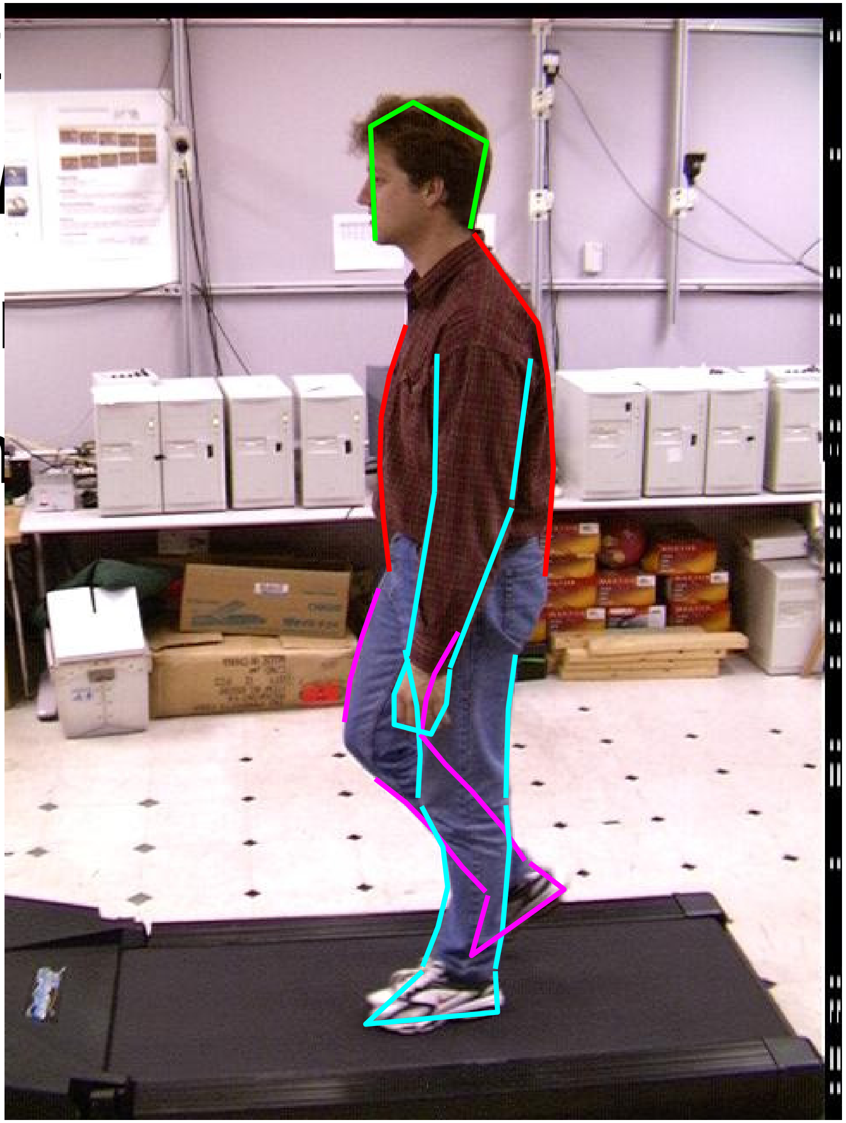
Work in progress

- motion and dynamics
- multiple viewpoints

Summary

Contributions

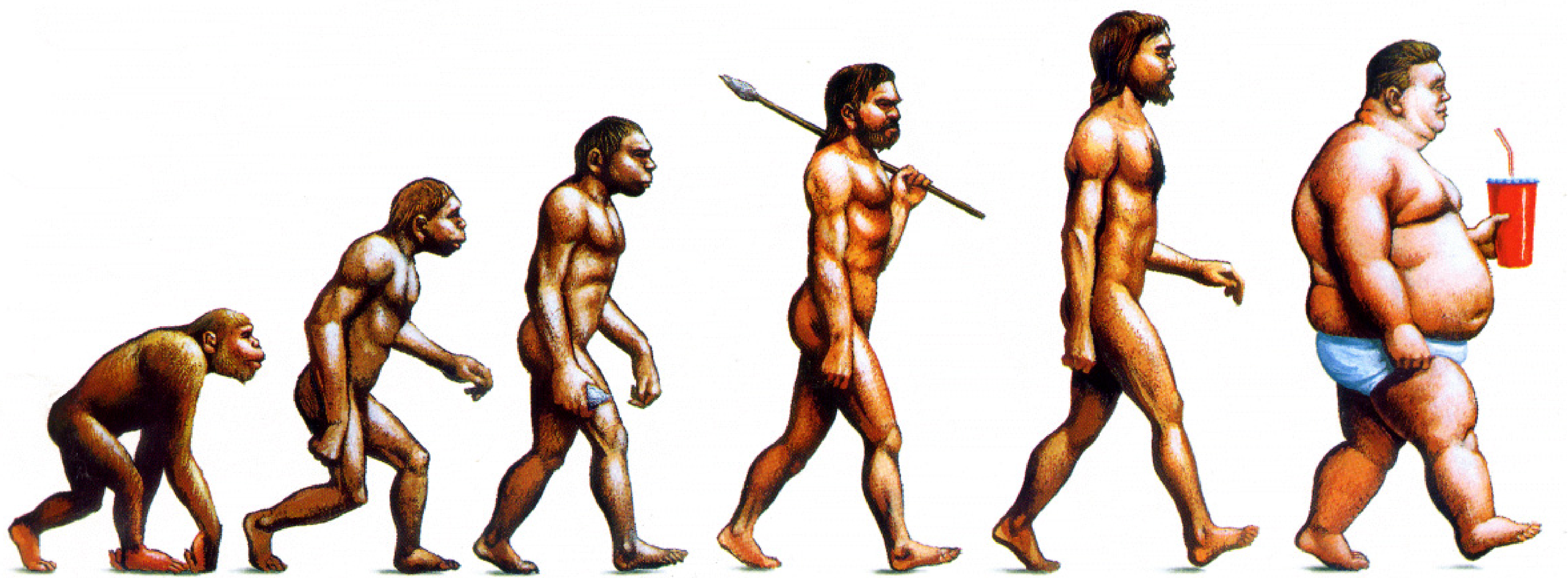
- Novel Prob. Represent
- Global Articulation $+$
- Multiple-Cue Integratio
- Efficient Spatial Inferen



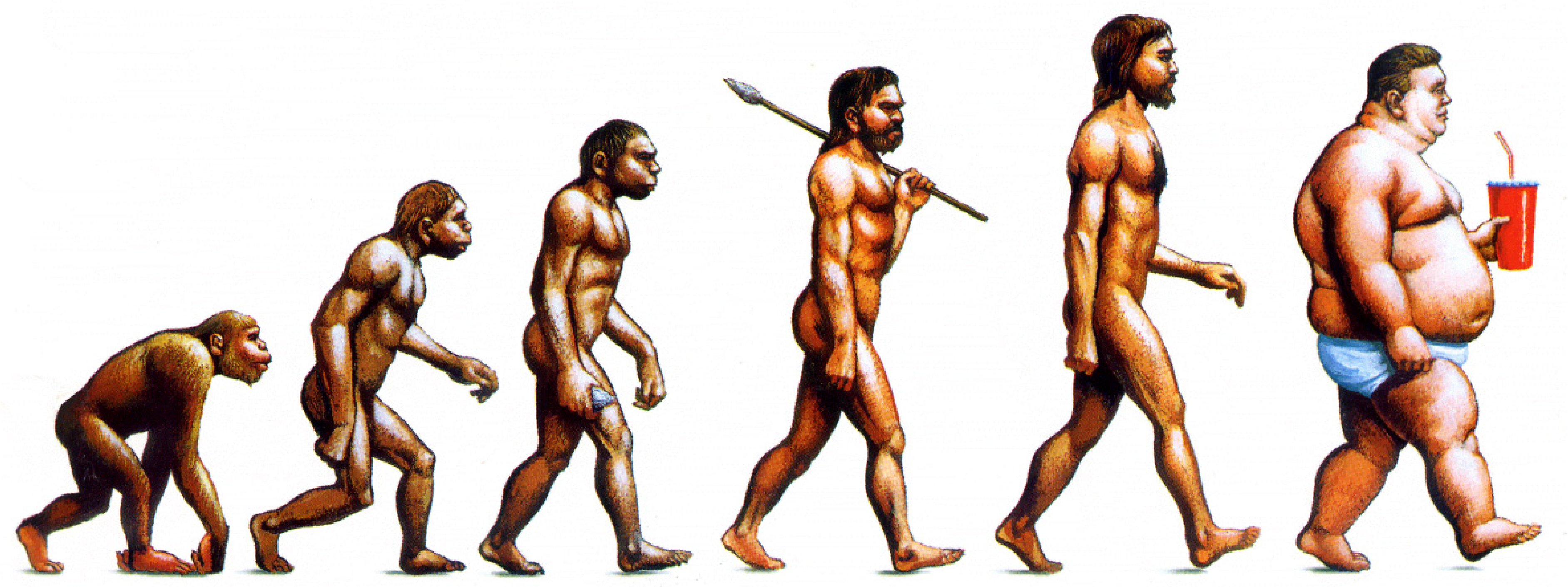
Work in progress

- motion and dynamics
- multiple viewpoints

The shape of things to come ...



The shape of things to come ...



The End!