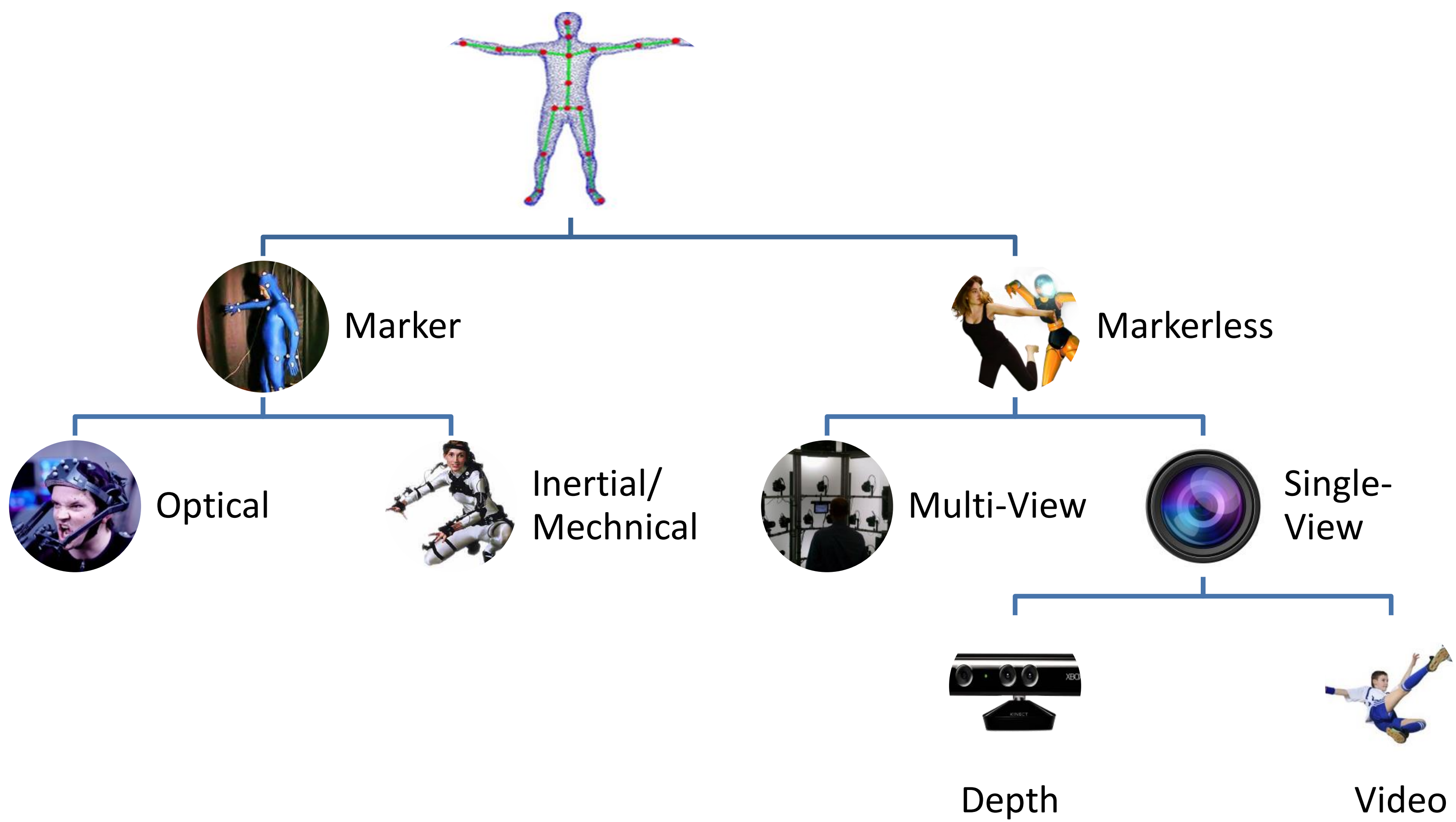
Motion Capture from RGB-D Camera

Ruigang Yang

ryang@cs.uky.edu

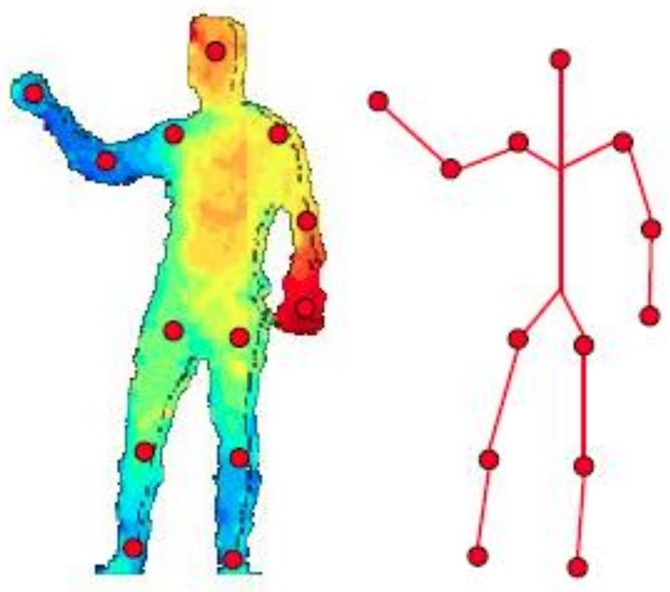
University of Kentucky

Motion/Performance Capture



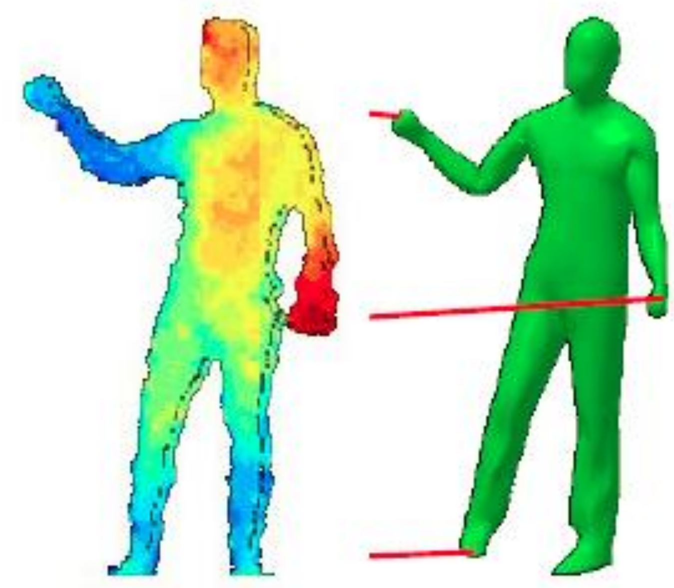
Mocap with a Single Depth Sensor

Discriminative



- **Shotton et al.,***CVPR,* 2011
- **…**

Generative

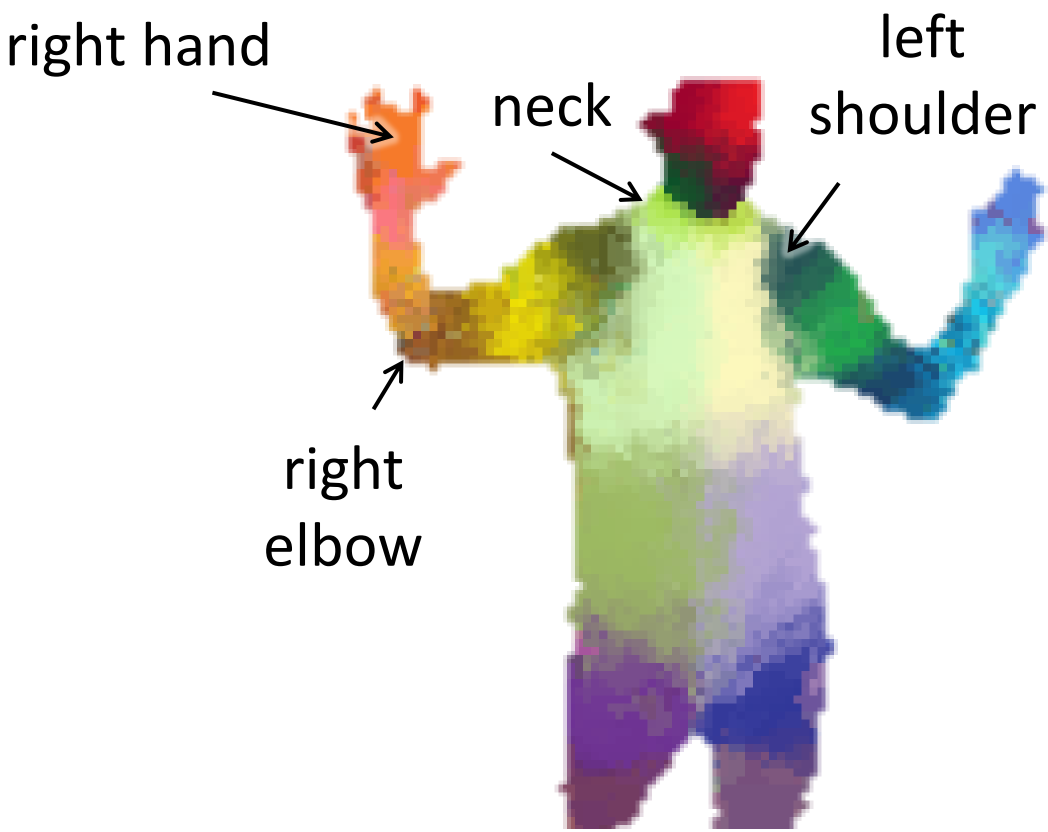


- **Ganapathi et al.** CVPR 2010
- …

Images from Christian Theobalt

The "Kinect" Approach

- Body Part Recognition



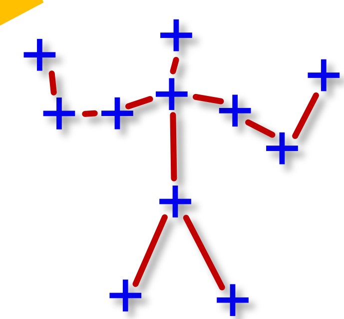
Slides from Shotton et al. 2011 CVPR Presentation

The Kinect pose estimation pipeline

capture
depth image &
remove bg

infer
body parts
per pixel

cluster pixels to
hypothesize
body joint
positions



fit model &
track skeleton

Classifying pixels

- Compute $P(C_x|\omega_x)$
  - pixels x
  - body part C_x
  - image window ω_x

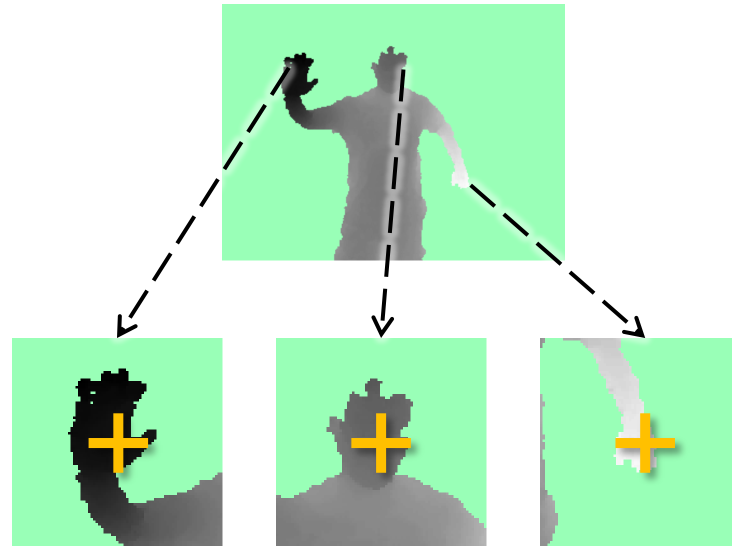


image windows move with classifier

- Discriminative approach
  - learn classifier $P(C_x|\omega_x)$ from training data

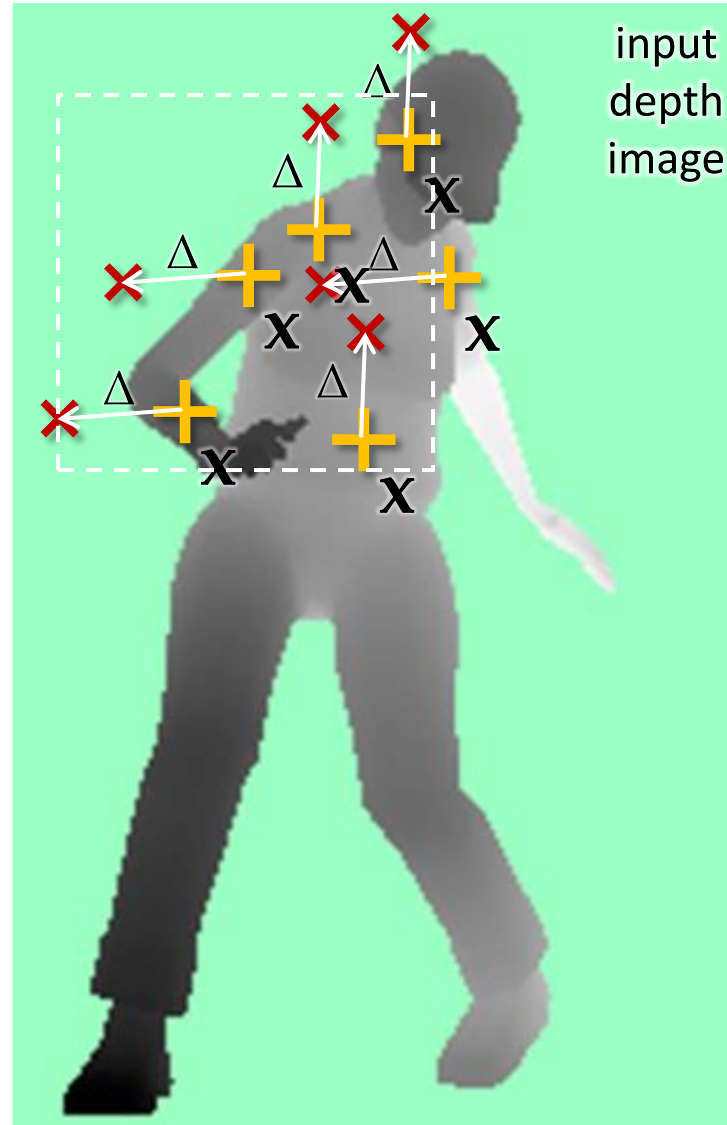
Fast depth image features

- Depth comparisons
  - very fast to compute

$$\underbrace{f(\overbrace{I}^{\text{image}},\underbrace{\mathbf{x}}_{\text{image coordinate}})}_{\text{feature response}} = \overbrace{d_I(\mathbf{x})}^{\text{depth}} - \overbrace{d_I(\mathbf{x}+\Delta)}^{\text{offset depth}}$$

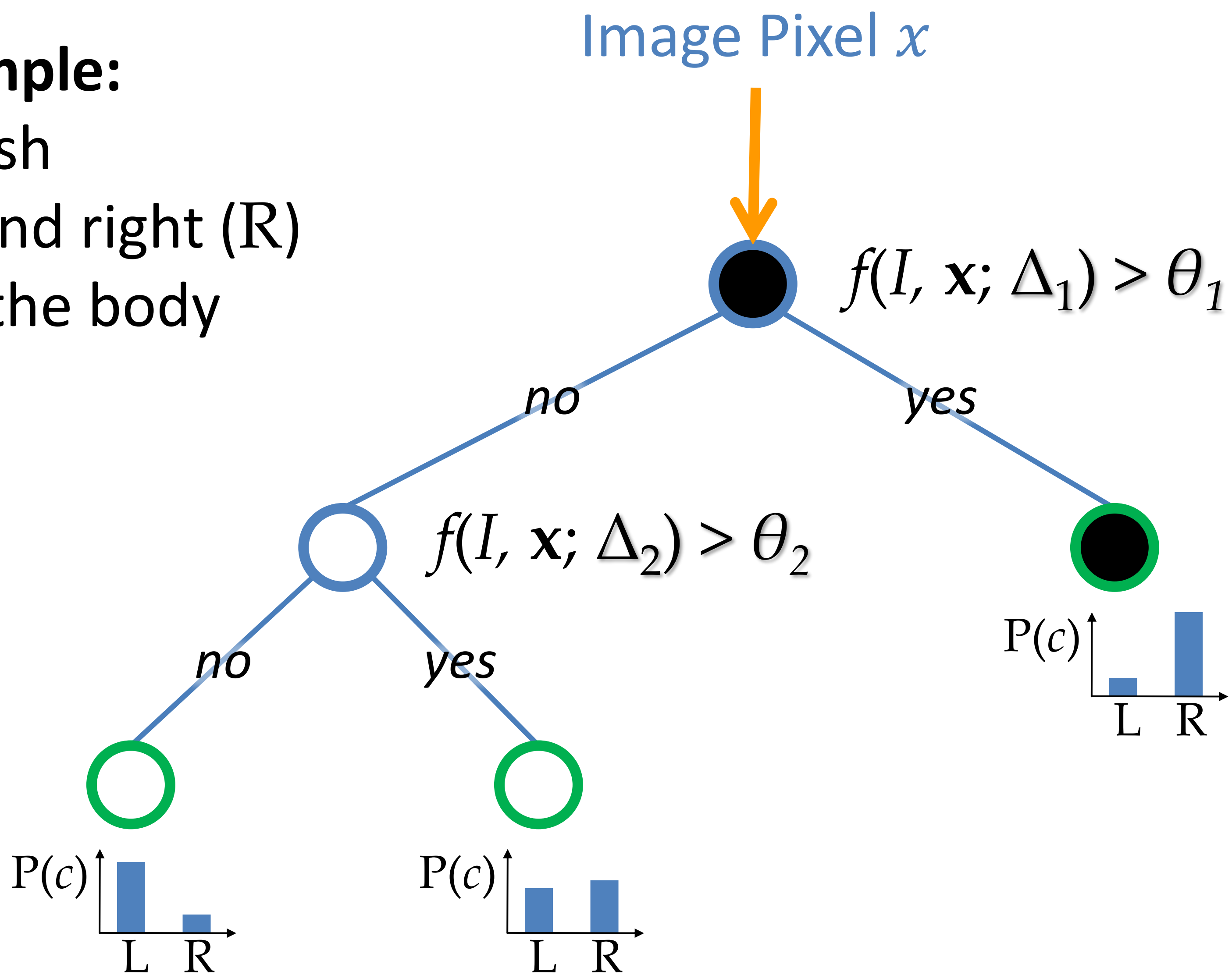
$$\Delta = \frac{\mathbf{v}}{d_I(\mathbf{x})}$$

Background pixels
d = large constant

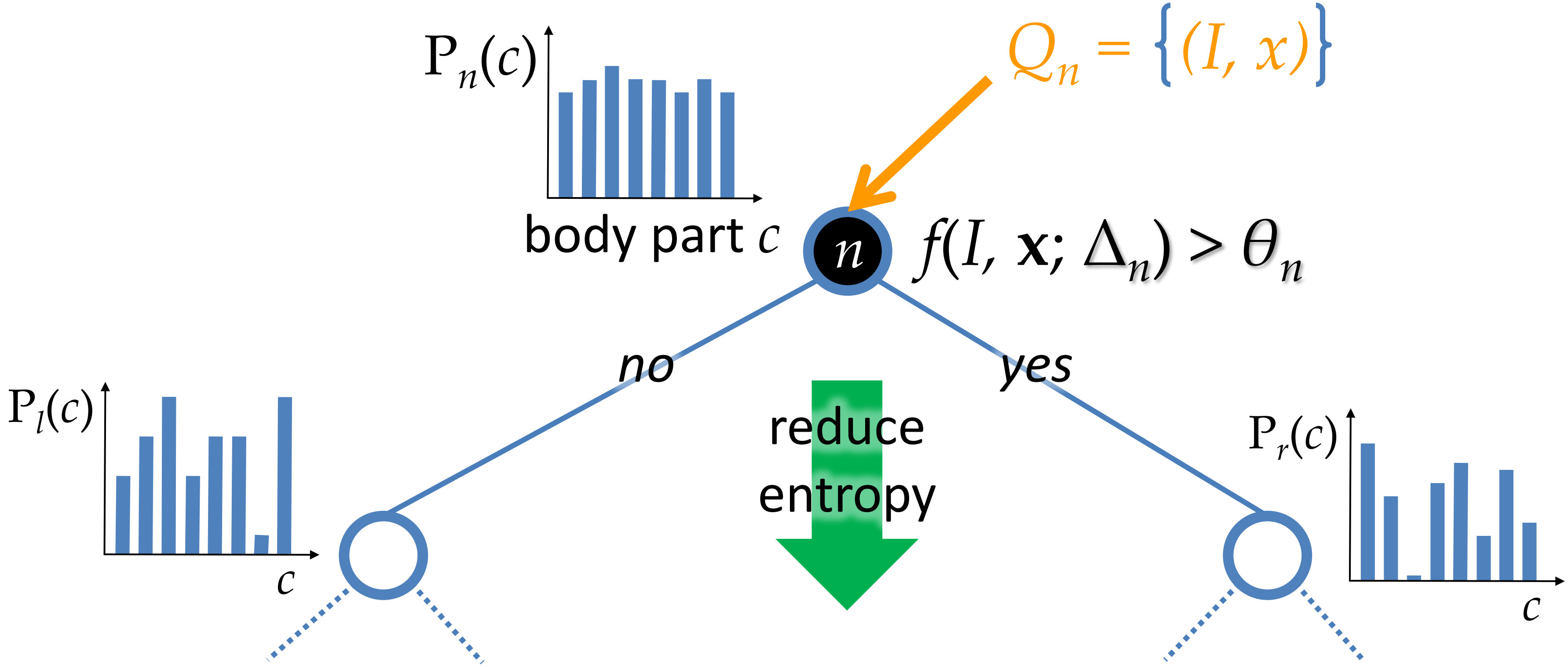


Decision tree classification

Toy example:
distinguish
left (L) and right (R)
sides of the body



Training decision trees [Breiman *et al.* 84]

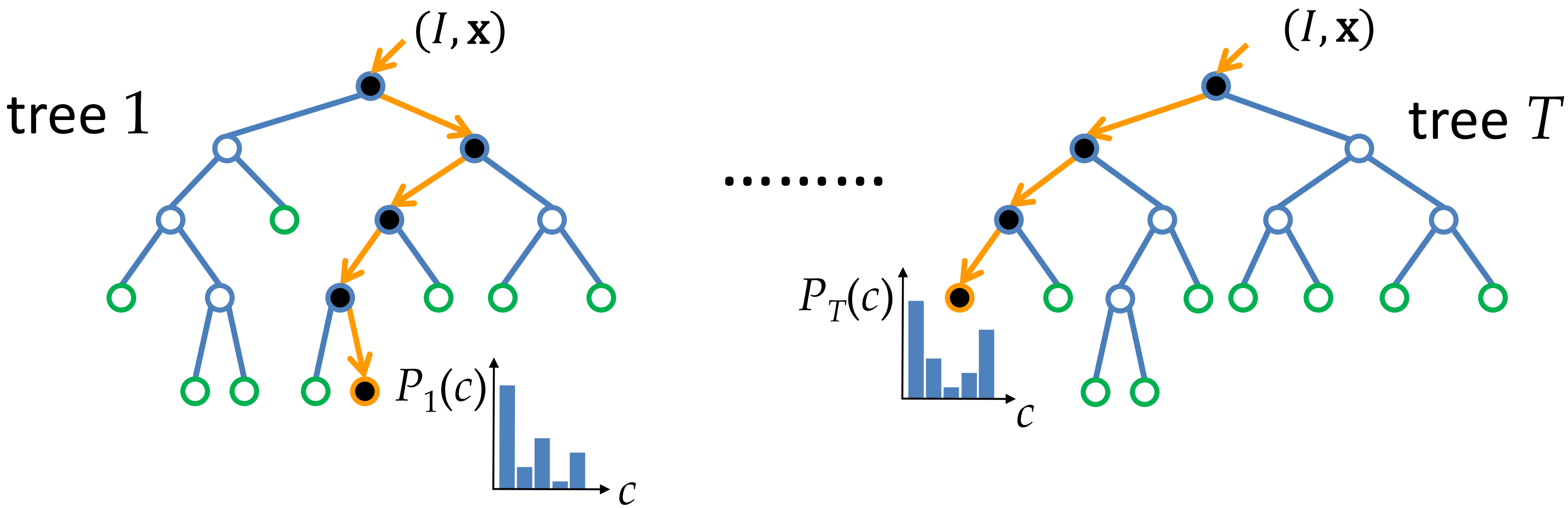


Take (Δ, θ) that maximises information gain:

$$\Delta E = -\frac{|Q_\mathrm{l}|}{|Q_n|}E(\mathrm{Q_l}) - \frac{|Q_\mathrm{r}|}{|Q_n|}E(\mathrm{Q_r})$$

Goal: drive entropy at leaf nodes to zero

Decision forest classifier [Breiman 01]



- Trained on different random subset of images
  - "bagging" helps avoid over-fitting
- Average tree posteriors $P(c|I,\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|I,\mathbf{x})$

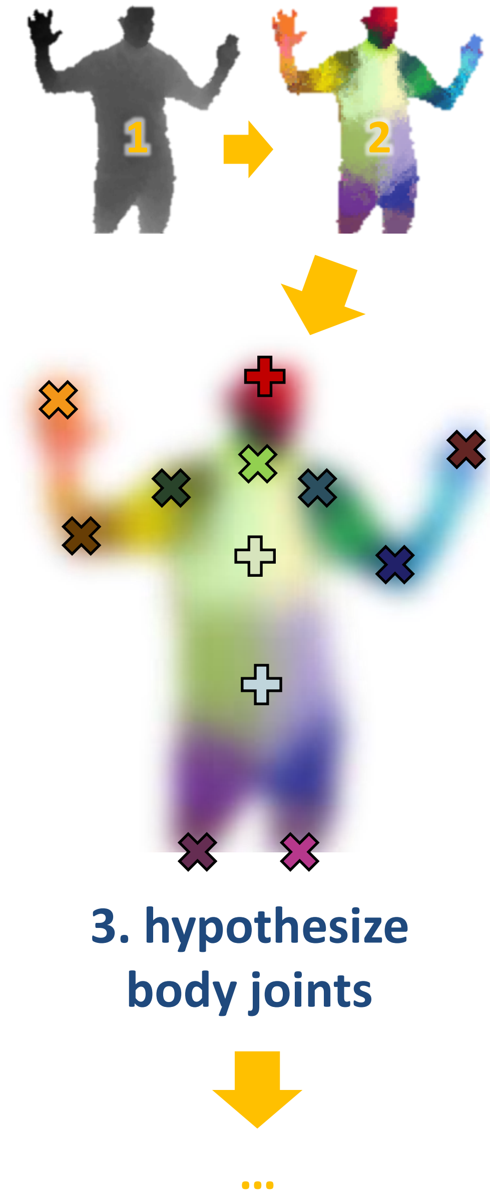
Body parts to joint hypotheses

- Define 3D world space density:

$$f_c(\overbrace{\bar{x}}^{\text{3D Coordinates}}) \propto \sum_i \overbrace{\omega_{ic}}^{\text{pixel weight}} \exp\left(-\left\|\frac{\bar{x}-\bar{x}_i}{\underbrace{b_c}_{\text{bandwidth}}}\right\|^2\right)$$

$$\omega_{ic} = \underbrace{P(c|I, x_i)}_{\text{inferred probability}} \underbrace{(d_I(x_i))^2}_{\text{depth at } i^{\text{th}} \text{ pixel}}$$

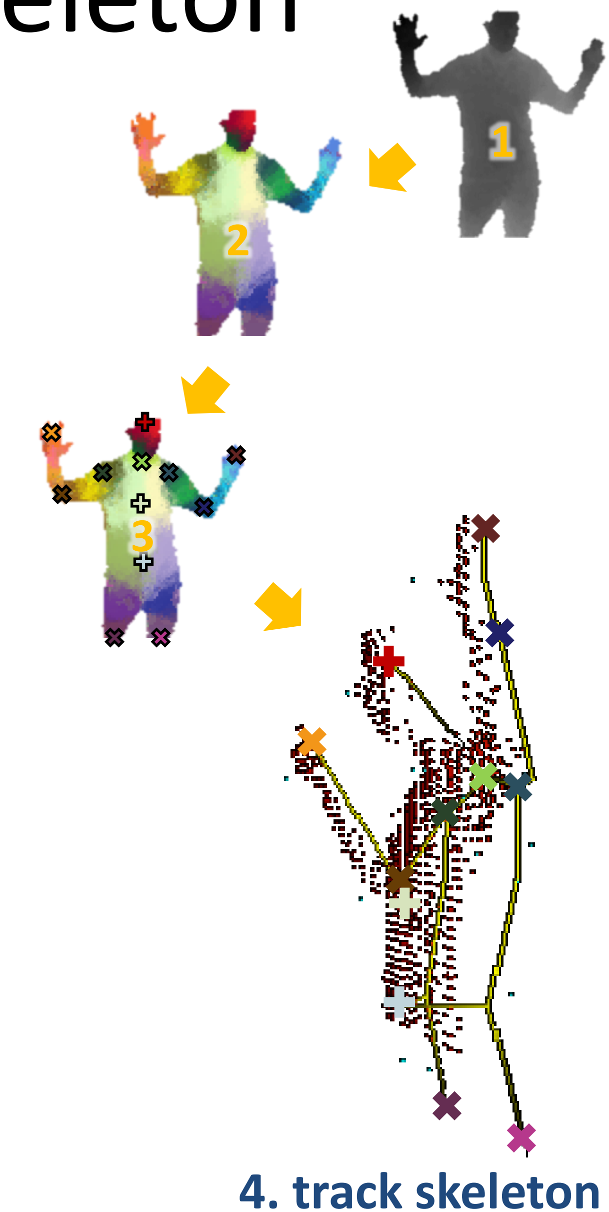
- Mean shift for mode detection



From proposals to skeleton

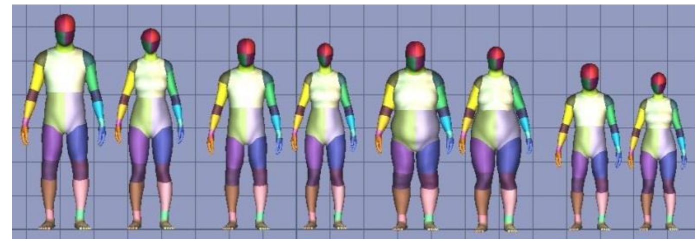
- Input
  - 3D joint hypotheses
  - kinematic constraints
  - temporal coherence

- Output
  - full skeleton
  - higher accuracy
  - invisible joints



Synthetic training data

Record mocap
500k frames
distilled to 100k poses



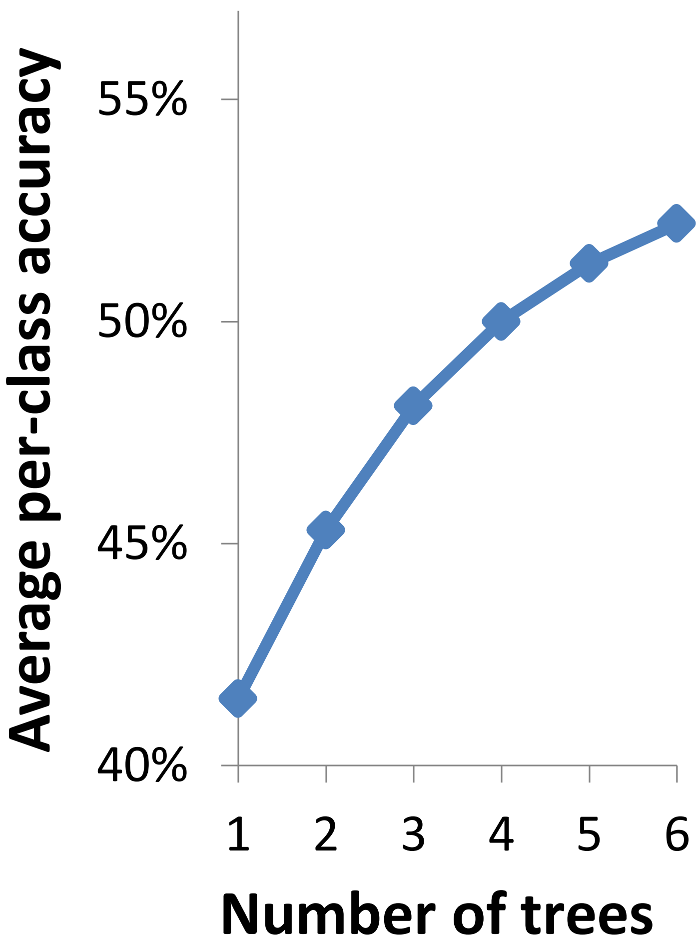
Train invariance to:

Synthetic vs real data

synthetic
(train & test)

real
(test)

Number of trees



ground truth

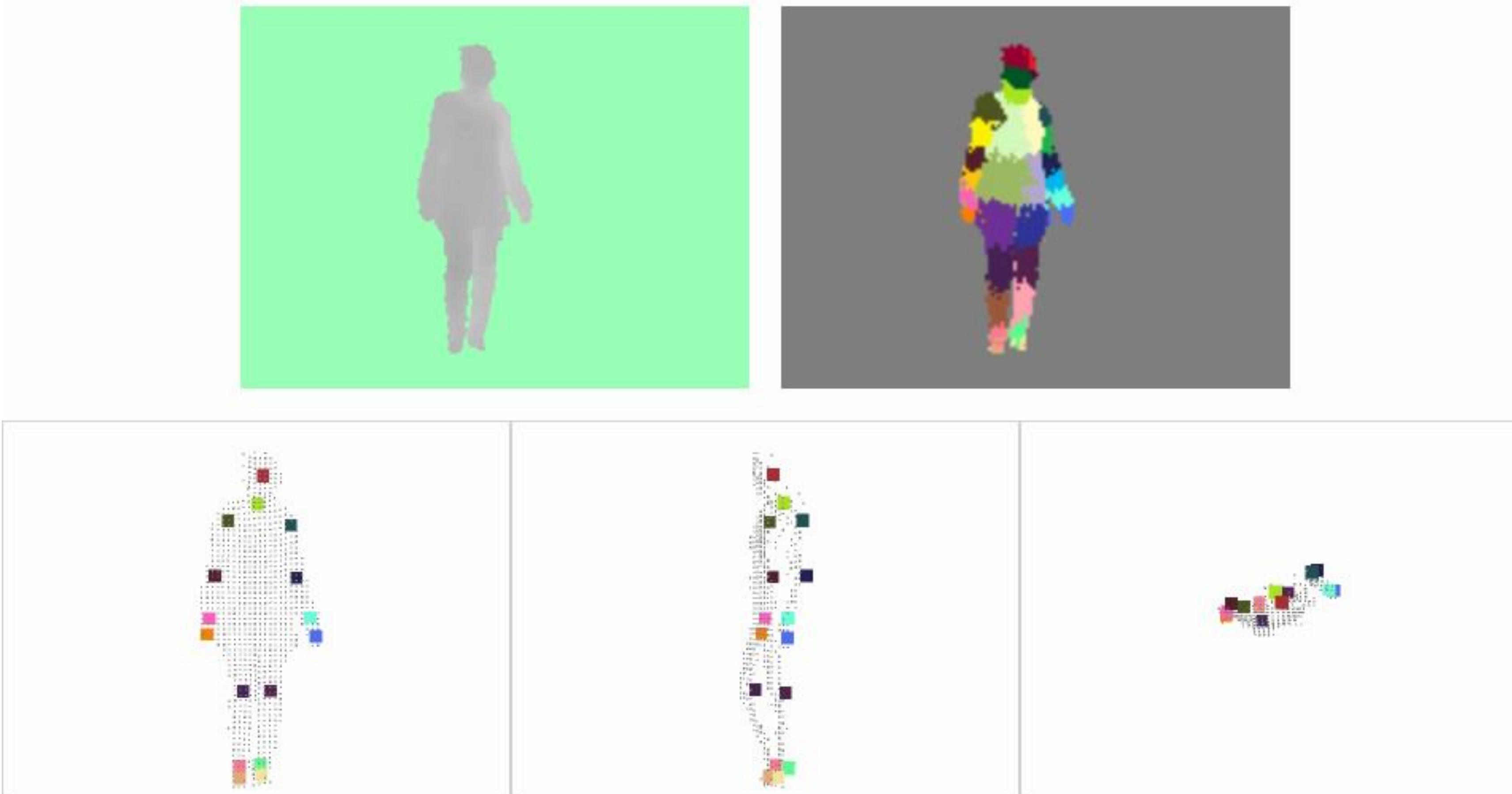
inferred body parts (most likely)

1 tree

3 trees

6 trees

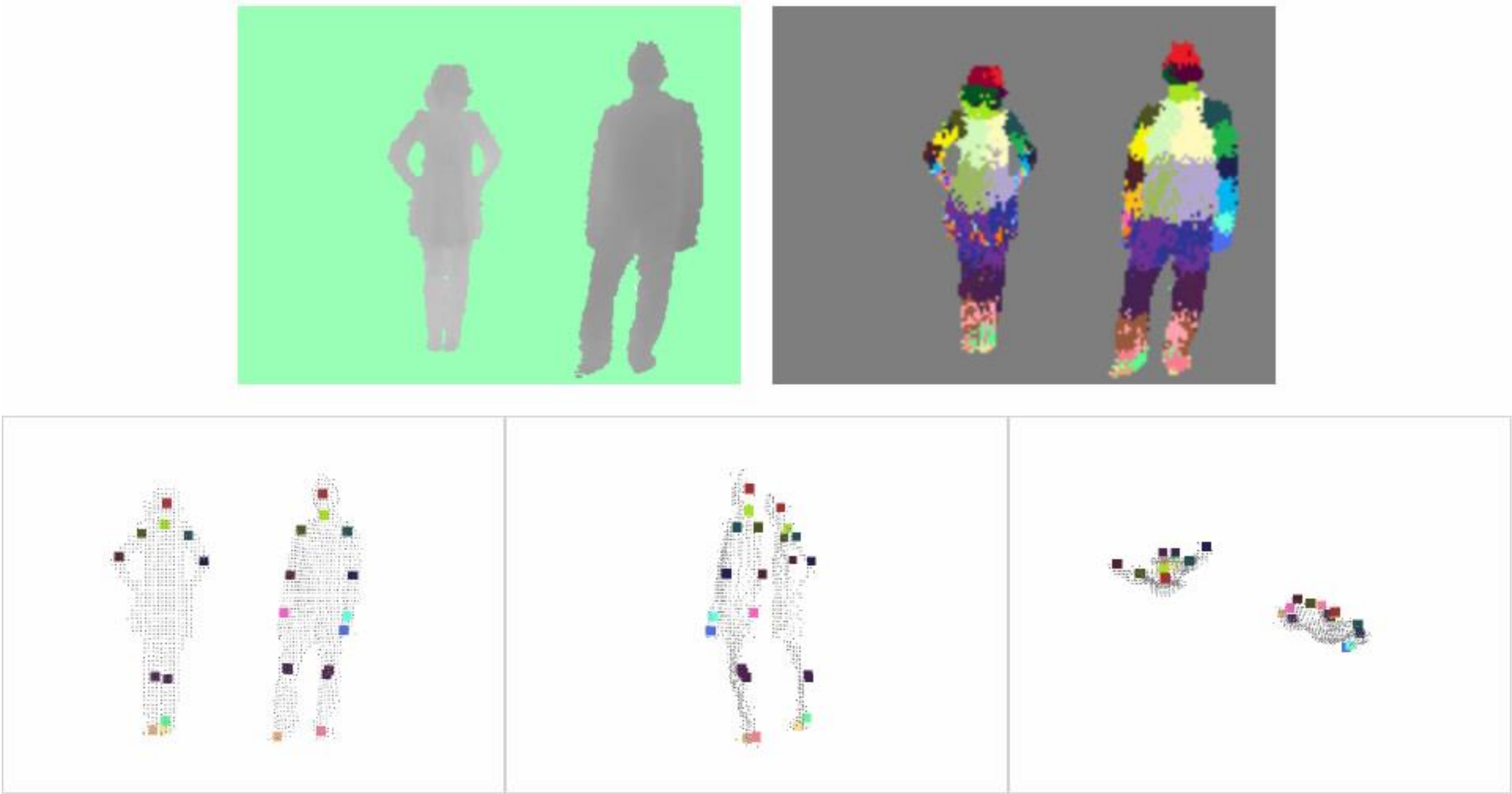
input depth inferred body parts



front view side view top view

inferred joint positions

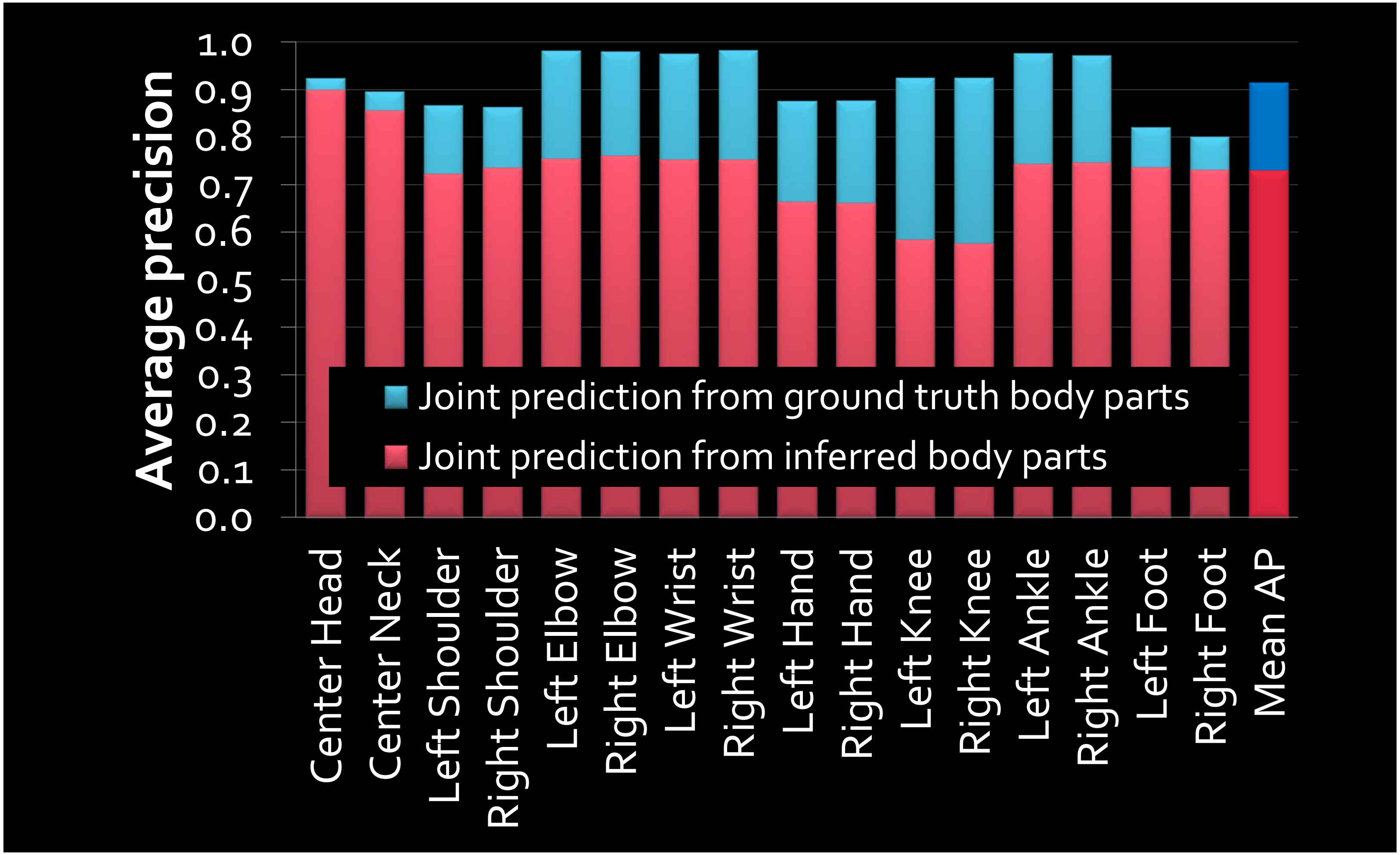
input depth **inferred body parts**



front view **side view** **top view**

inferred joint positions

Joint prediction accuracy



Summary

Pros:

– Frame-by-frame gives robustness

– Fast, simple machine learning

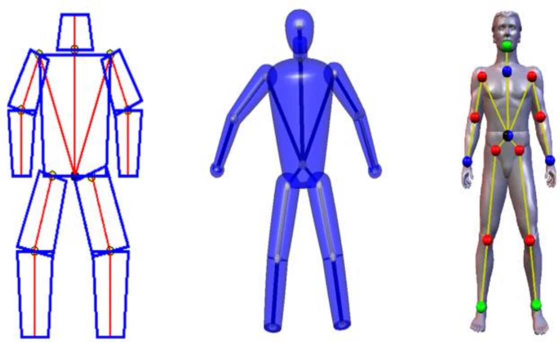
Cons:

– Accuracy can be improved.

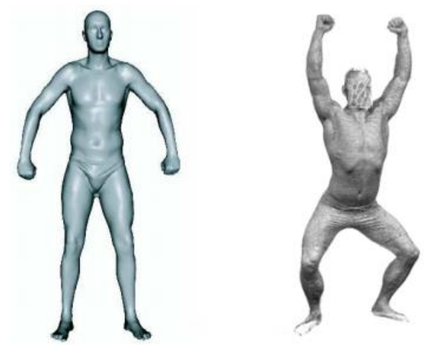
Threshold (m)	0.02	0.03	0.05	0.07	0.1	0.15	0.20
mAP: [2]	0.02	0.06	0.30	0.53	0.73	0.82	0.85

Generative Approaches

- Kinematic model

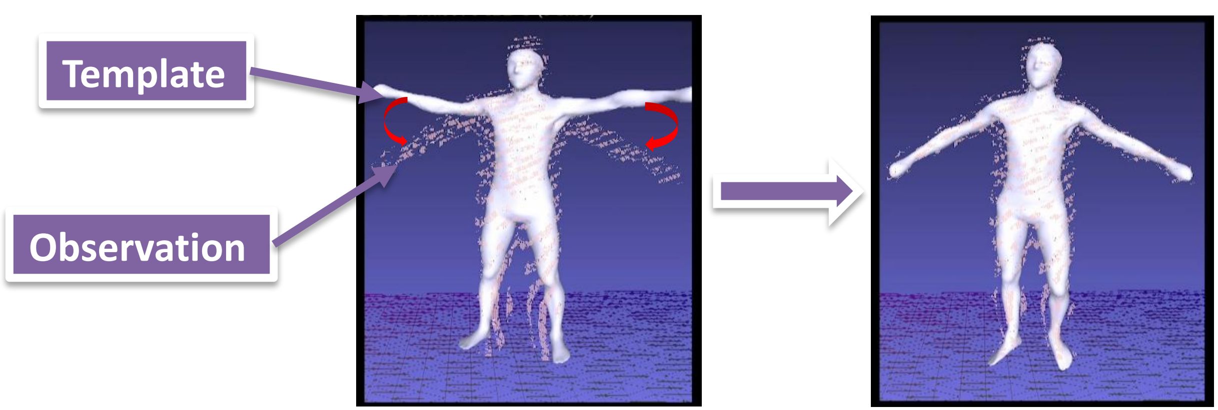


Single Templates (normally in a tree structure) *[Poppe et al. 2007]*



Statistical Models *[Anguelov et al. 2005, Hasler et al. 2009]*

- Maximize model-to-observation consistency



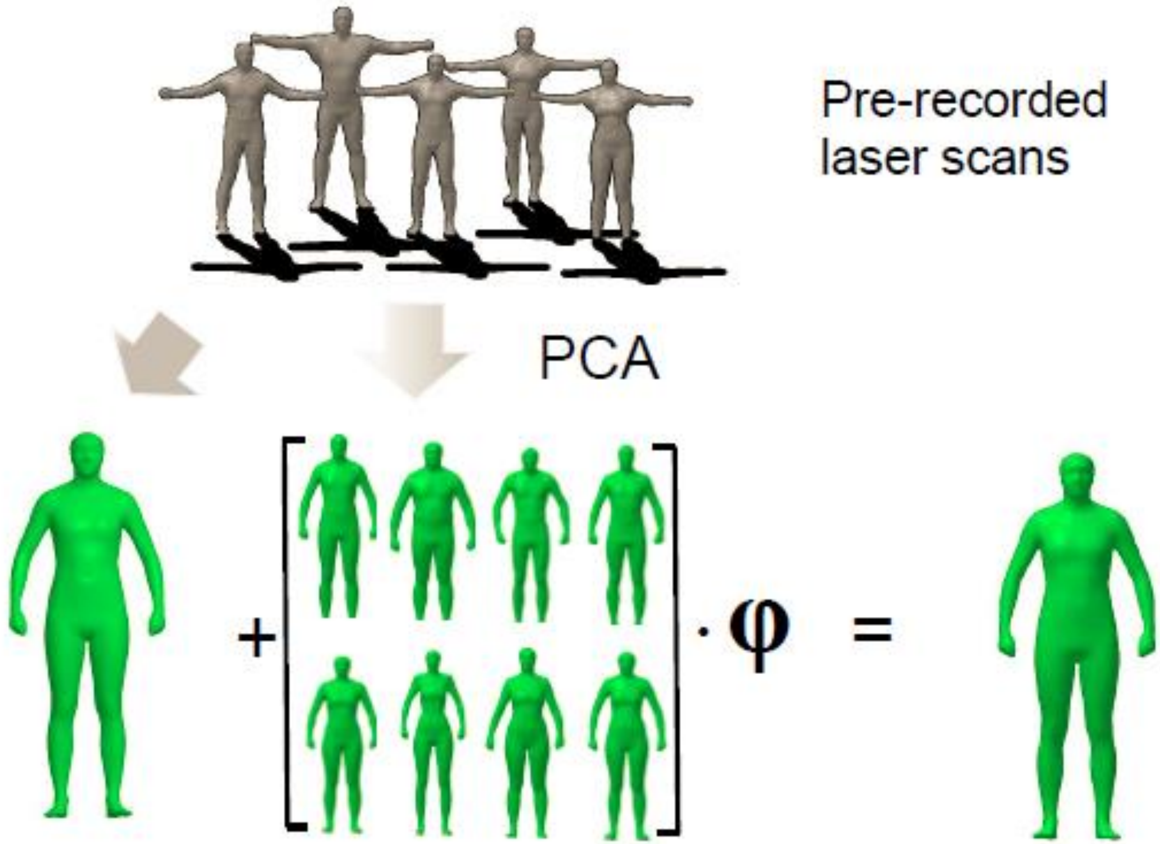
Linear Blend Skinning (LBS)

- Mesh + Skeleton

- Each vertex is controlled by several neighboring bones

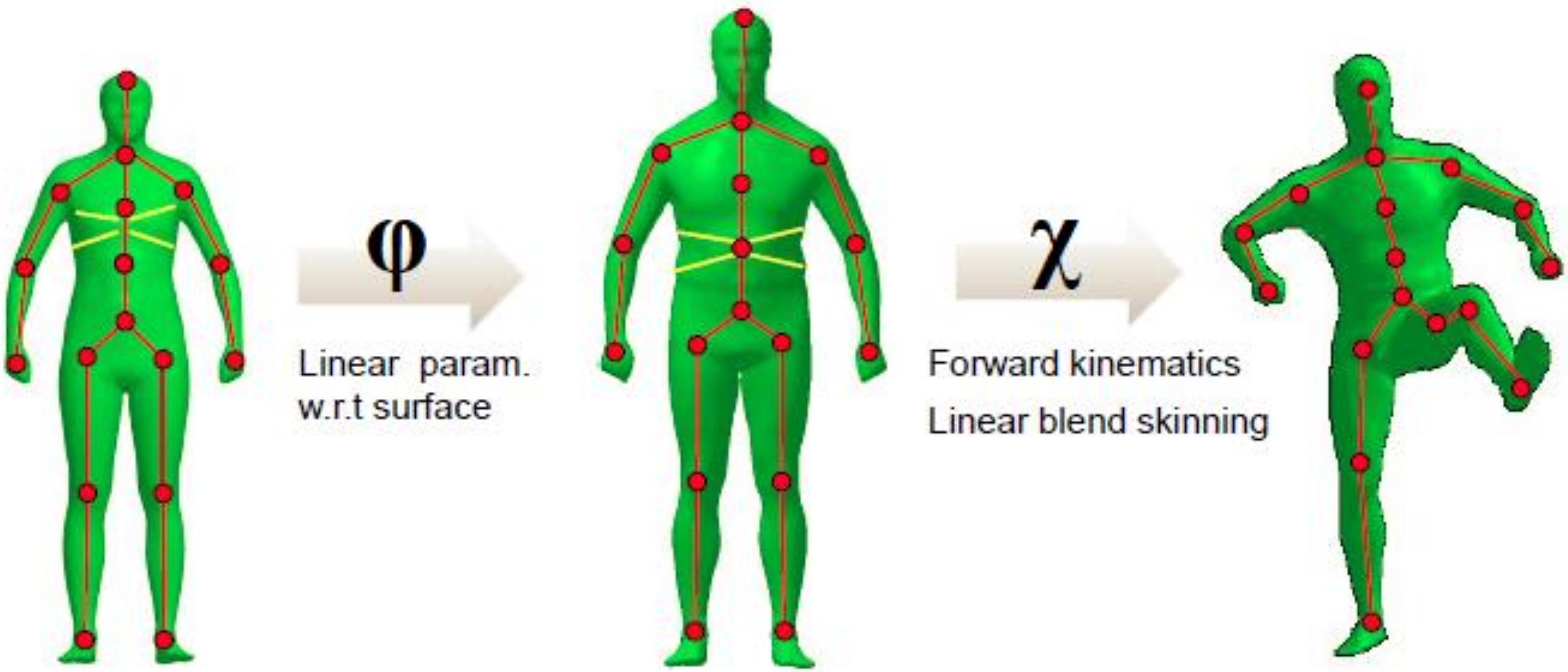
$$v_i = \left(\sum_k \overbrace{\boxed{\alpha_{i,k}}}^{\text{Skinning weights}} \underbrace{\boxed{T_k}}_{\text{Bone transformations}}\right) \boldsymbol{v}_i^0$$

Parametric Models



Images from Christian Theobalt

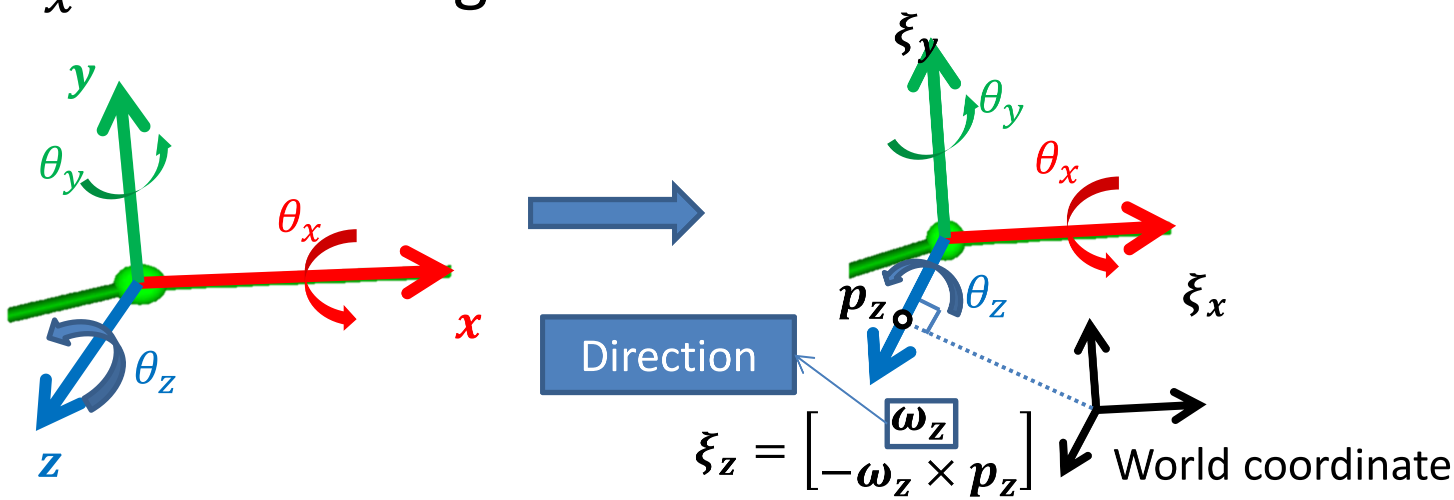
Pose Parameterization



Images from Christian Theobalt

Twist-based Representation of Transformations [Murray et al., 1994]

- $T_x = e^{\widehat{\xi_x}\theta_x}$
  - ξ_x: the twist representing rotation axis
  - θ_x: rotation angle



- Linearization
  - $T_x \approx (I + \widehat{\xi_x}\theta_x)$ if θ_x is small

Model-to-Observation Consistency

- Silhouette
- Texture
- Depth

Image-based

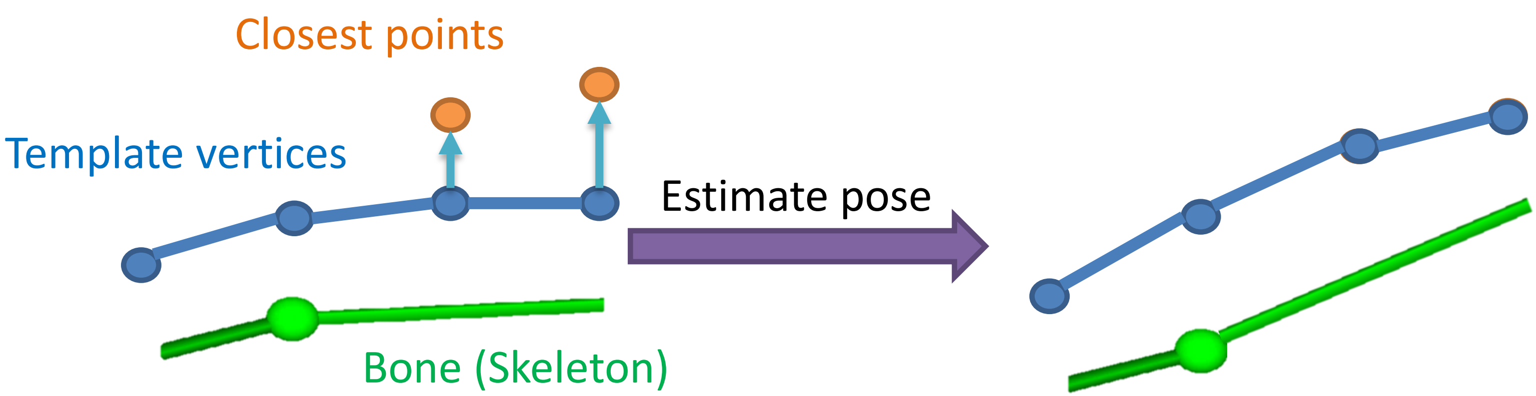
[*Gall et al. 2009,*
Gall et al. 2010,
Liu et al. 2013,
etc.]

Depth-based

[*Ganapathi et al. 2010,*
Ye et al. 2011
Baak et al. 2011
Helten et al. 2013
Wei et al. 2013
Ye et al. 2014
etc.]

Depth consistency and pose update

- Typically using ICP
  - [*Ganapathi et al. 2012, Helten et al. 2013, Wei et al. 2013*]



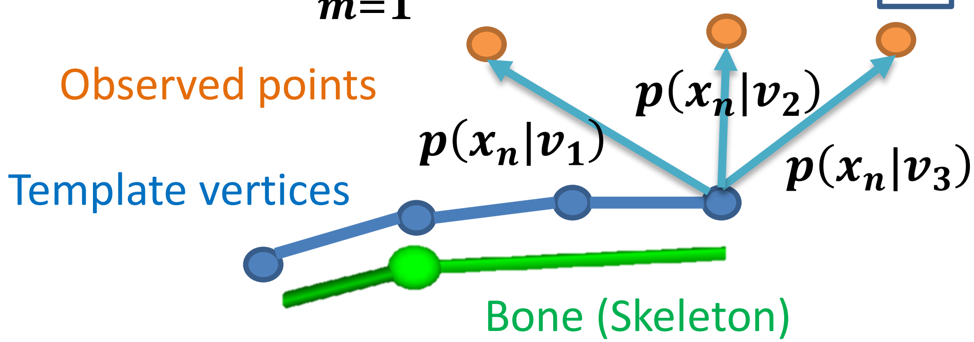
- Limitation: sensitive to local minima

Soft correspondences association

- Gaussian Mixture Model
  - Template vertices are Gaussian centroids
  - Observed points are sampling from the GMM

$$\boldsymbol{p(x_n) = \sum_{m=1}^{M} \frac{1-u}{M} p(x_n|v_m) + \boxed{\frac{u}{N}}}$$

Uniform distribution for outliers



- Pose estimation = find the pose that gives the $\boldsymbol{v_m}$ that achieves maximum joint probability $\prod_n \boldsymbol{p(x_n)}$

Maximize the joint probability

- Log likelihood

$$E(\boldsymbol{\Theta}, \boldsymbol{S}, \boldsymbol{\sigma}^2) = \sum_{n=1}^{N} \log\left(\sum_{m=1}^{M} \frac{1-u}{M} p(x_n|v_m) + \frac{u}{N}\right)$$

- Solve parameters (pose $\boldsymbol{\Theta}$) of $\boldsymbol{v_m}$ via EM
  - Negative complete log likelihood

Linear Blend Skinning

$$\boldsymbol{Q}(\boldsymbol{\Theta}, \boldsymbol{S}, \boldsymbol{\sigma}^2) \propto \sum_{n,m} p(v_m|x_n)\, ||x_n - \boxed{v_m(\boldsymbol{\Theta})}||^2$$

$$\xRightarrow{\text{Linearization}} \sum_{n,m} \boxed{p(v_m|x_n)}\, ||x_n - L(\boxed{v_o}, \boxed{\boldsymbol{\Theta}^{\mathbf{prev}}}, \boxed{\Delta\boldsymbol{\Theta}})||^2$$

Posterior

Template vertices in reference pose

Most recent pose

Incremental pose update

Pose Energy Function

- Negative complete log likelihood

$$\sum_{n,m} \boldsymbol{p}(\boldsymbol{v_m}|\boldsymbol{x_n})\left||\boldsymbol{x_n} - \boldsymbol{L}(\boldsymbol{v_o}, \boldsymbol{\Theta}^{\mathbf{prev}}, \boldsymbol{\Delta\Theta})|\right|^{\mathbf{2}}$$

- Regularization
  - Small pose update $\left||\boldsymbol{\Delta\Theta}|^{\mathbf{2}}\right|$
  - Prediction via auto-regression

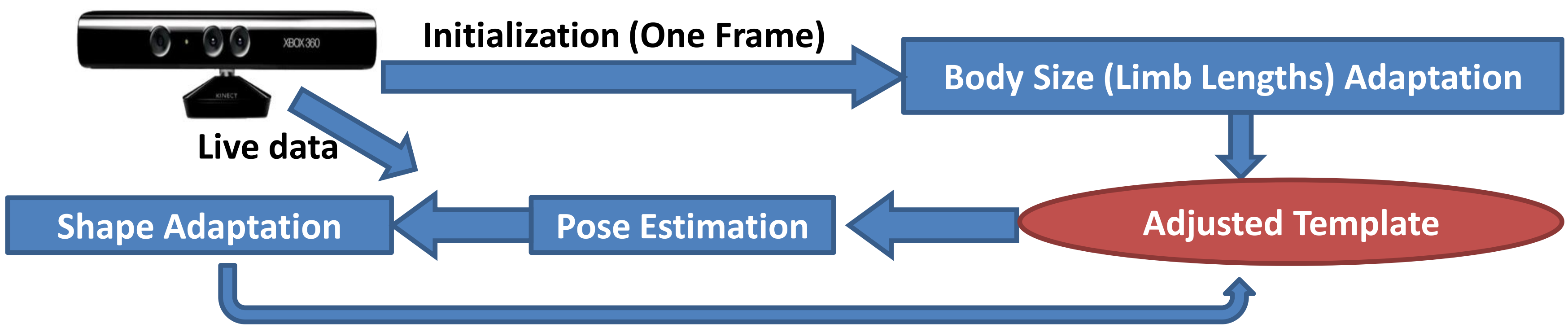
$$\left||\boldsymbol{\Theta}^{\mathbf{prev}} + \boldsymbol{\Delta\Theta} - \boldsymbol{\Theta}^{\mathbf{pred}}|\right|^{\mathbf{2}}$$

Pose Estimation

- Initialize the pose $\boldsymbol{\Theta}^t$ (e.g. from previous frame)
- Iterate until convergence
  - Compute template vertices $\{\boldsymbol{v_m}\}$ via LBS
  - E-step: compute posterior
  - M-step: minimize the pose energy function in previous slide over the pose update $\boldsymbol{\Delta\Theta}$
  - Increment the pose $\boldsymbol{\Theta}^t$ with $\boldsymbol{\Delta\Theta}$

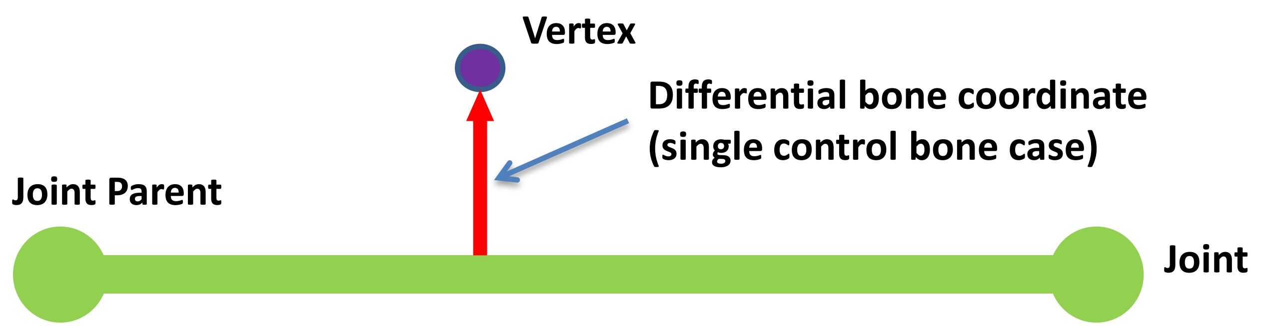
Template-Subject Consistency

- Body size and shape consistency between template and the subject is critical.
- Therefore, estimate
  - body size: limb length scales (higher or shorter)
  - shape: Vertex displacements (fatter or slimmer)
- System workflow



Limb length scales

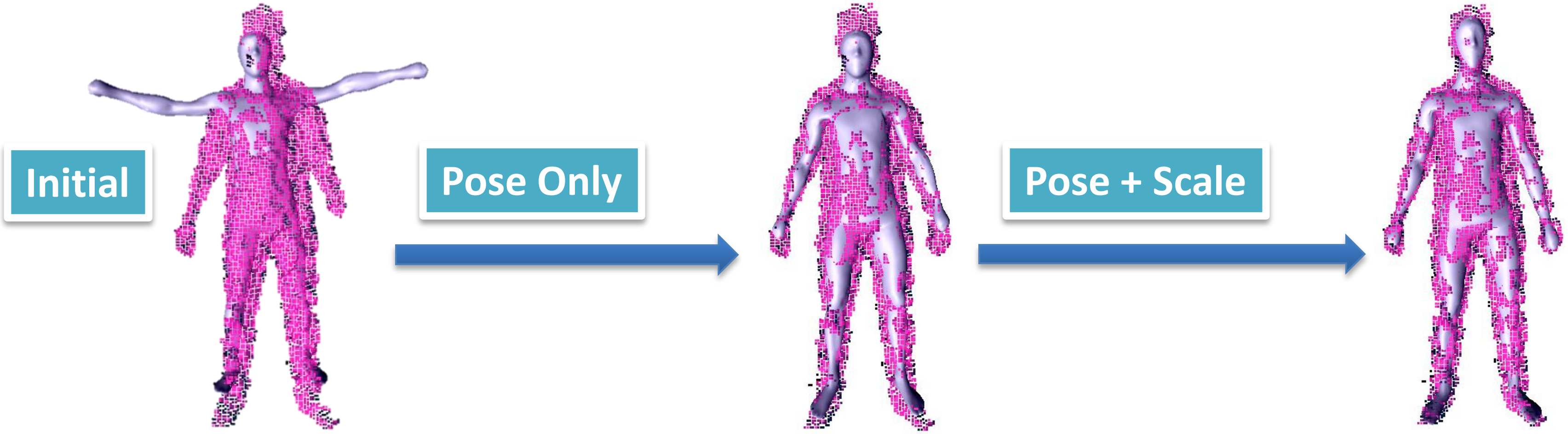
- Represent the vertex as a function of the limb length scales
- Differential bone coordinates [*Straka et al. 2012*]
  - Template vertex $\boldsymbol{v_m}$= LinearFunction(scales)



Limb length scales estimation

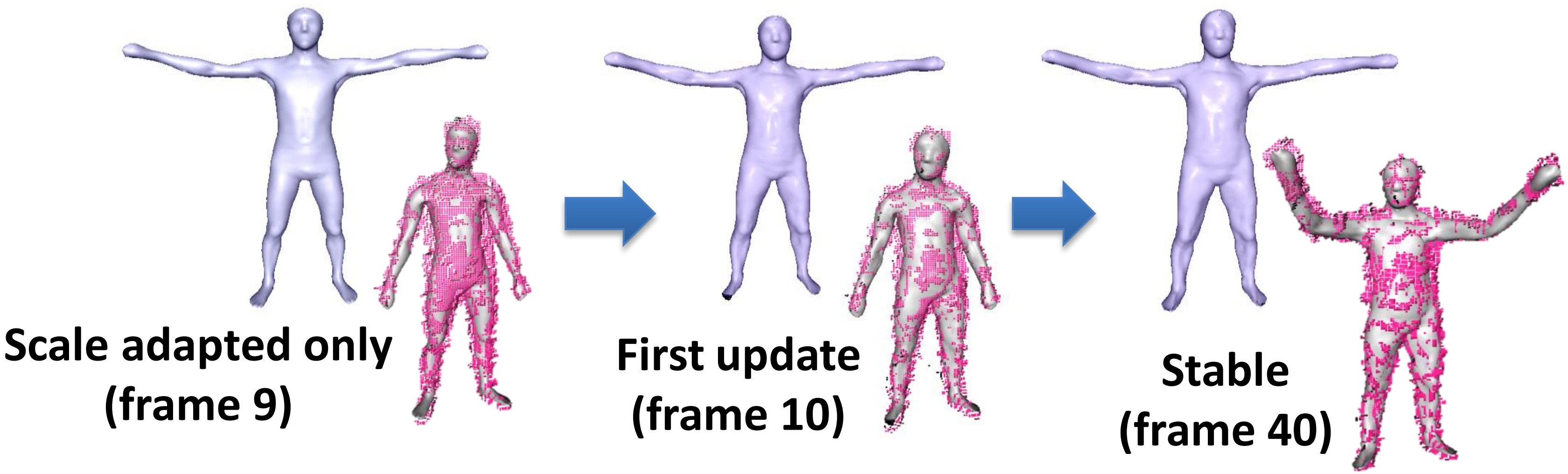
- Iterate between pose estimation and scale estimation
- Scale energy function
  - Negative complete log likelihood with $\boldsymbol{v_m}$ = Linearfunction(scales)
  - Regularization
    - Symmetric bones have similar scales
    - Connected bones have similar scales

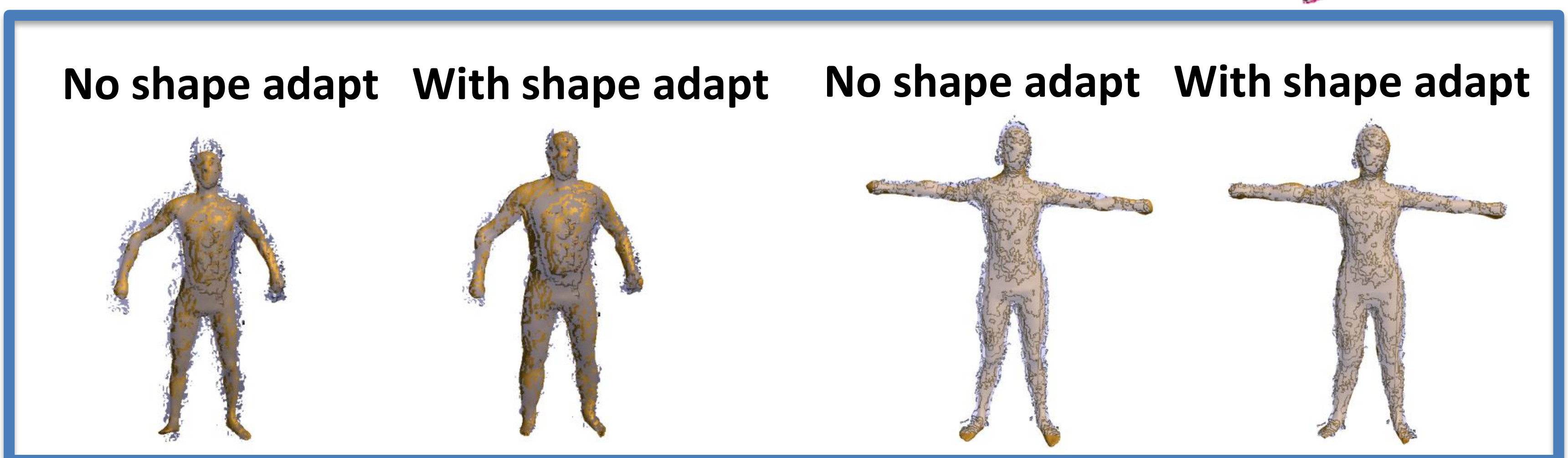
Limb length scales estimation (cont.)



Shape Adaptation (cont.)

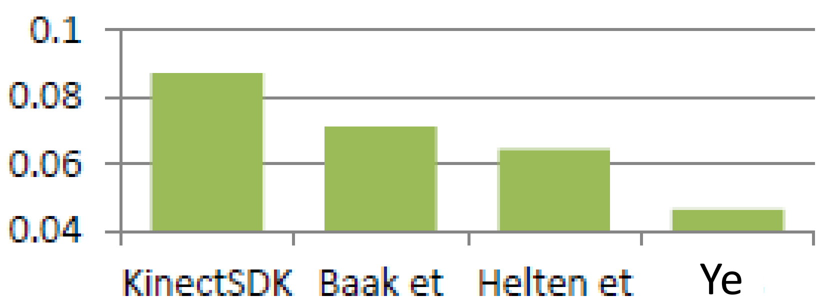
- Update for each 5 frames



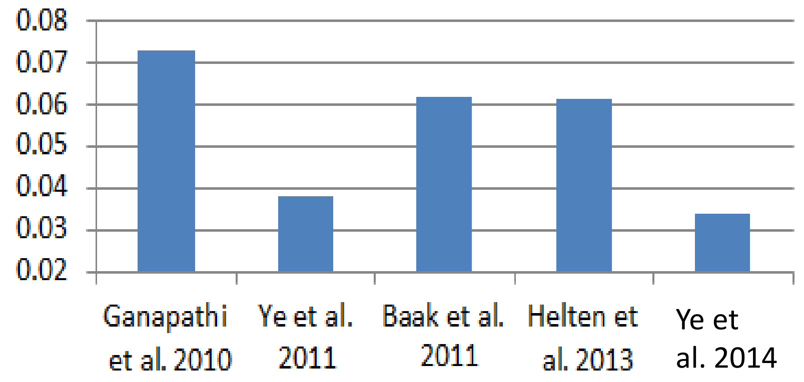


Evaluations

Comparison in terms of joint distance errors (unit = meter)

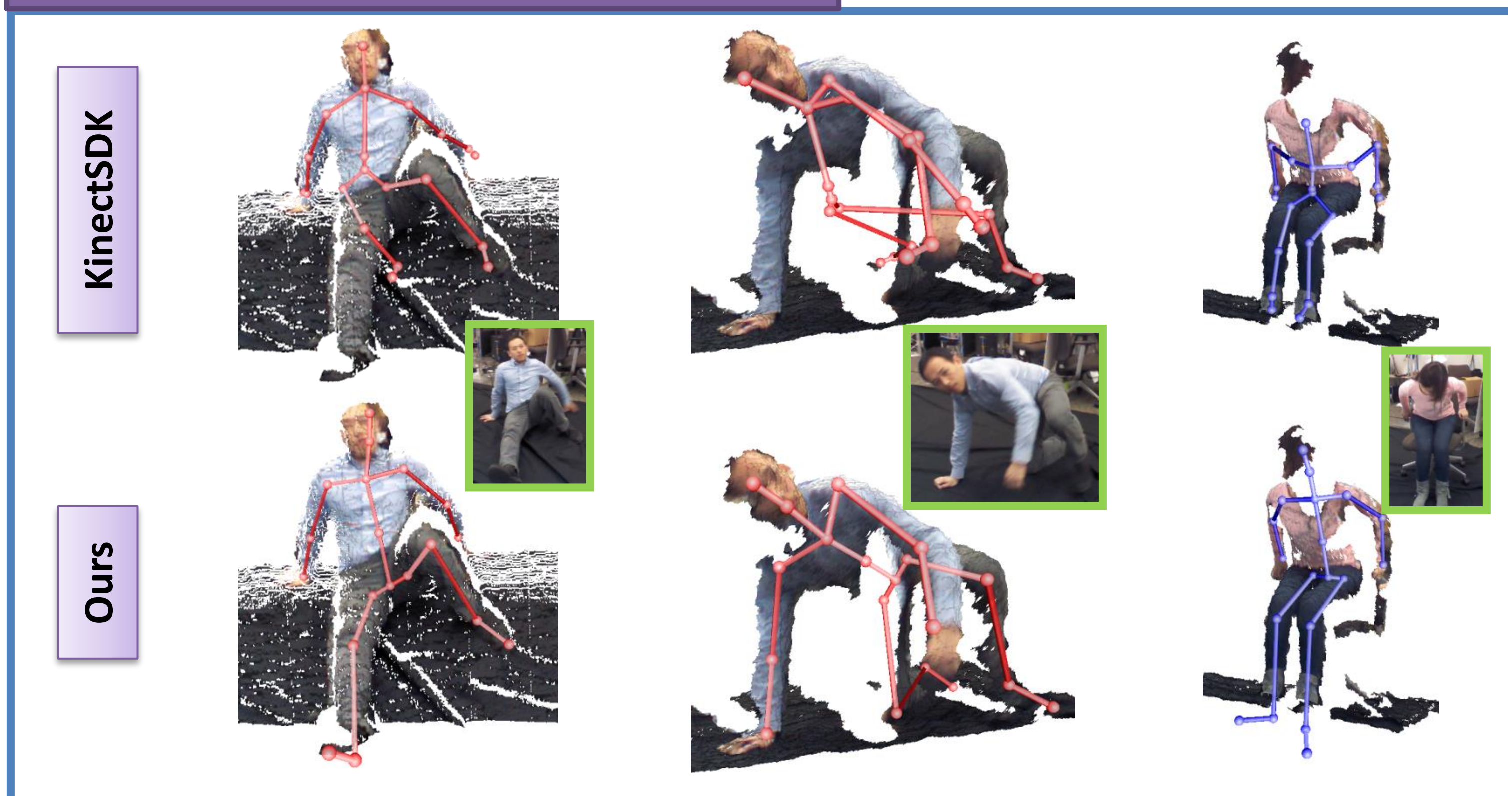


Comparison in terms of marker distance errors (unit = meter)



Qualitative Evaluations

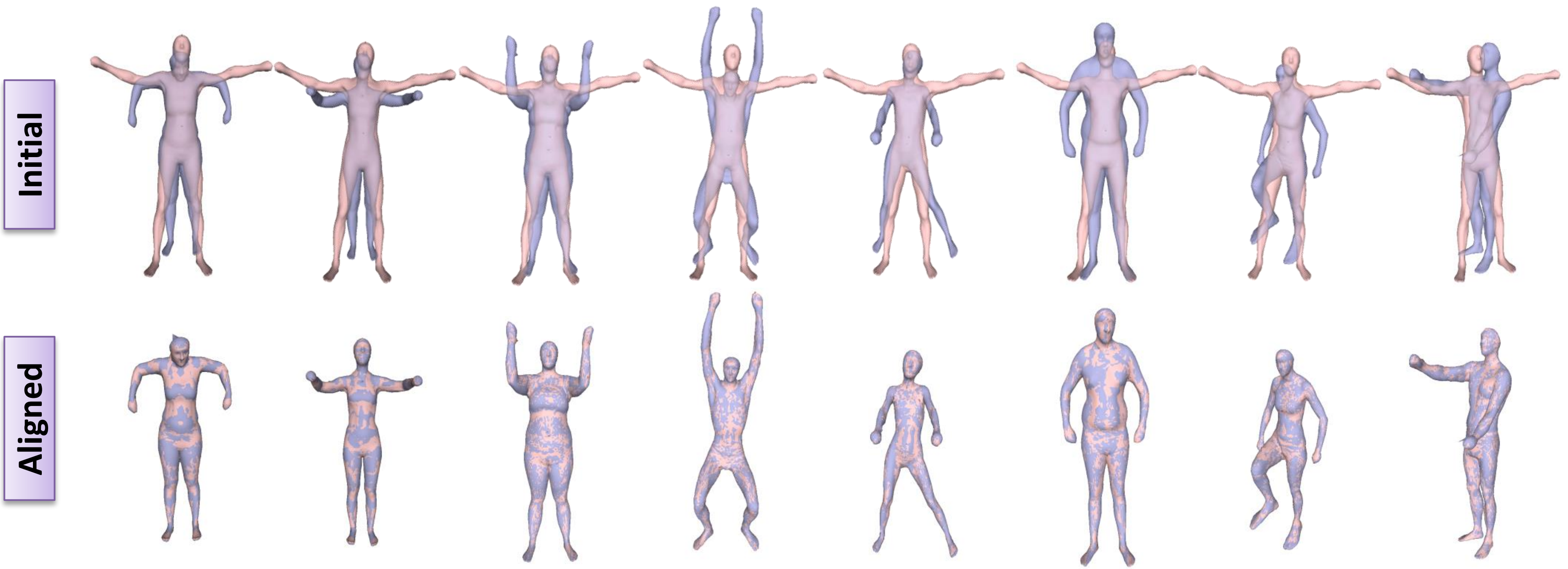
Comparisons with KinectSDK



Qualitative Evaluations - Video

Application: shape collection registration

- Align a single skin template to a collection of meshes



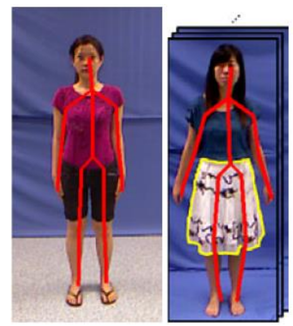
Applications

Health Care

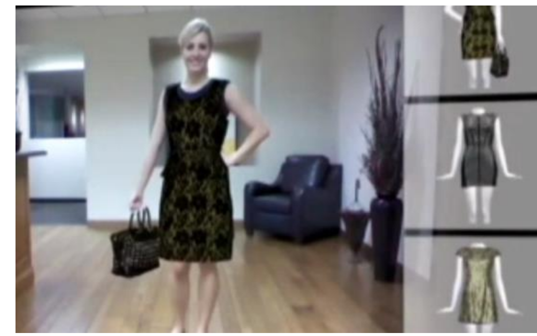
Reflexion Health

Jintronix

Virtual Try-On



Tan et al. 2013

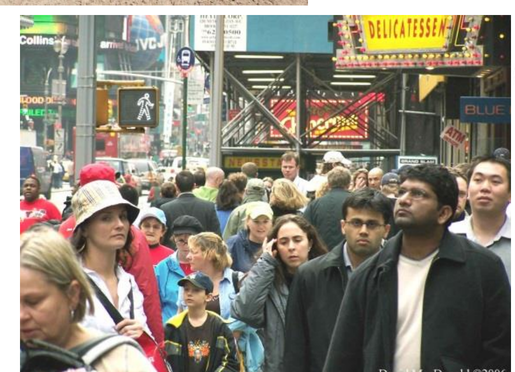


Virtual Accessory

Mao et al. 2014

Summary

- Advantages
  - Metric Input/Output
  - Fast and robust algorithms
- Challenges
  - Outdoor
  - Large Deformation
  - Crowd Mocap



Acknowledgment

- Christian Theobalt
- J. Shotton et al.
- Mao Ye

References

- Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
- [Gall et al 2009] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, Motion capture using joint skeleton tracking and surface estimation, in IEEE CVPR, 2009
- [Cui et al 2012] Y. Cui, W. Chang, T. N oll, and D. Stricker, Kinectavatar: Fully automatic body capture using a single kinect, in ACCV 2012 Workshop on Color Depth Fusion in Computer Vision, 2012
- [Change et al 2011] W. Chang and M. Zwicker, Global registration of dynamic range scans for articulated model reconstruction, ACM TOG, 2011
- [Weiss et al 2011] A. Weiss, D. Hirshberg, and M. J. Black, Home 3D body scans from noisy image and range data, in ICCV, 2011.
- [Strake et al 2012] M. Straka, S. Hauswiesner, M. R uther, and H. Bischof, Simultaneous shape and pose adaption of articulated models using linear optimization, in ECCV, 2012
- [Baak et al 2011] A. Baak, M. M uller, G. Bharaj, H.-P. Seidel, and C. Theobalt, A data-driven approach for realtime full body pose reconstruction from a depth camera, in ICCV, 2011
- [Ganapathi et al 2010] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, Real time motion capture using a single time-ofight camera, in IEEE CVPR, 2010
- [Ganapathi et al 2012] V. Ganapathi, C. Plagemann, D. Koller and S. Thrun, Real Time Human Pose Tracking from Range Data, in ECCV, 2012

References (Cont.)

- [de Aguiar et al 2008] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun. 2008. Performance capture from sparse multi-view video. In ACM SIGGRAPH 2008
- [Vlasic et al 2008] D. Vlasic, I. Baran, W. Matusik, and J. Popovic, Articulated mesh animation from multi-view silhouettes," in ACM SIGGRAPH 2008
- [Ballan et al 2008] L. Ballan and G. M. Cortelazzo, Marker-less Motion Capture of Skinned Models in a Four Camera Set-up using Optical Flow and Silhouettes, in 3DPVT, 2008
- [Gall et al 2011] J Gall, A Fossati L Van Gool, Functional categorization of objects using real-time markerless motion capture, in CVPR, 2011
- [Liu et al 2013] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.P. Seidel, C. Theobalt, Markerless Motion Capture of Multiple Characters Using Multiview Image Segmentation. IEEE TPAMI, 2013
- [Li et al 2013] H. Li, E. Vouga, A. Gudym, J. Barron, L. Luo and G. Gusev, 3D Self-Portraits , in ACM SIGGRAPH Asia, 2013
- [Li et al 2009] H. Li, B. Adams, L. Guibas, M. Pauly. Robust Single-View Geometry and Motion Reconstruction, in ACM SIGGRAPH, 2009
- [Shotton et al 2011] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, Real-time human pose recognition in parts from single depth images, in CVPR, 2011
- [Helten et al 2013a] T. Helten, A. Baak, G. Bharaj, M. Muller, H.P. Seidel, C. Theobalt, Personalization and Evaluation of a Real-time Depth-based Full Body Tracker, in 3DV, 2013
- [Helten et al 2013b] T. Helten, A. Baak, M. Muller, C. Theobalt, Full-Body Human Motion Capture from Monocular Depth Images, in LNCS, 2013
- [Ye et al 2011] M. Ye, X. Wang, R. Yang, L. Ren and M. Pollefeys. Accurate 3D Pose Estimation from a Single Depth Image. In ICCV, 2011
- [Ye and Yang 2014] M. Ye and R. Yang, Real-time Simultaneous Pose and Shape Estimation for Articulated Objects with a Single Depth Camera, in CVPR 2014
- [Mueslund et al 2006] T. B. Moeslund, A. Hilton, and V. Kr uger, A survey of advances in vision-based human