DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Tracking II: Visual 3D human motion tracking

Kim Steenstrup Pedersen

Plan for today

- Visual human motion tracking:
  - Marker-based versus marker-less tracking
  - Articulated visual 3D tracking

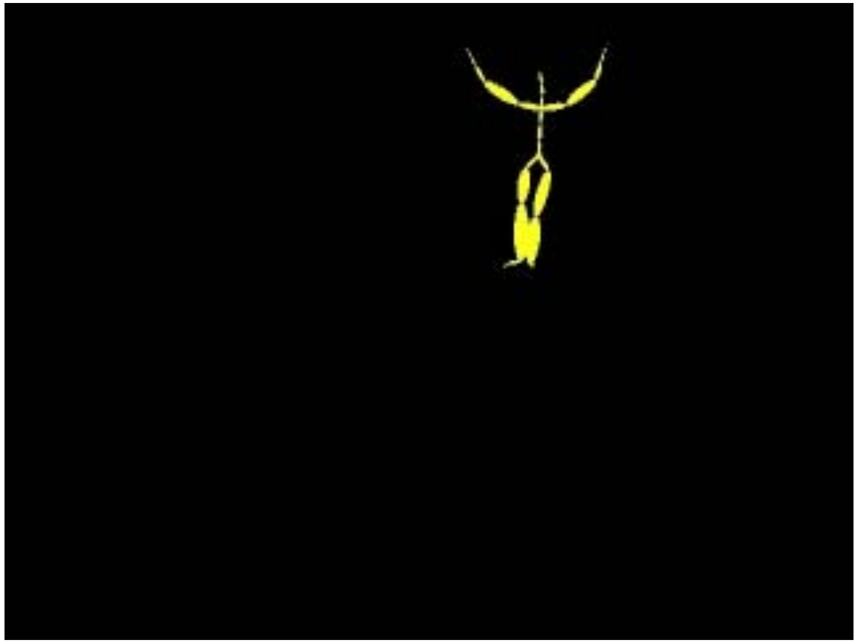
3D human motion tracking

- Def.: Estimation of 3D pose and motion of an articulated model of the human body from visual data – this is referred to as motion capture.
- Marker-based motion capture (MoCap):
  - Outcome: Tracking markers on joints in 3D giving joint positions.
  - Markers: Acoustic, inertial, LED, magnetic, reflective, etc.
  - Cameras or active sensors.
- Marker-less motion capture (MoCap):
  - Outcome: 3D joint positions or triangulated surfaces and relation to video sequence.
  - Multi-view (several cameras / views)
  - Monocular (single camera / view)
  - Camera / view types: Optical camera, stereo pair, time-of-flight cameras, etc.

Marker based motion capture:
System from Vicon (www.vicon.com)

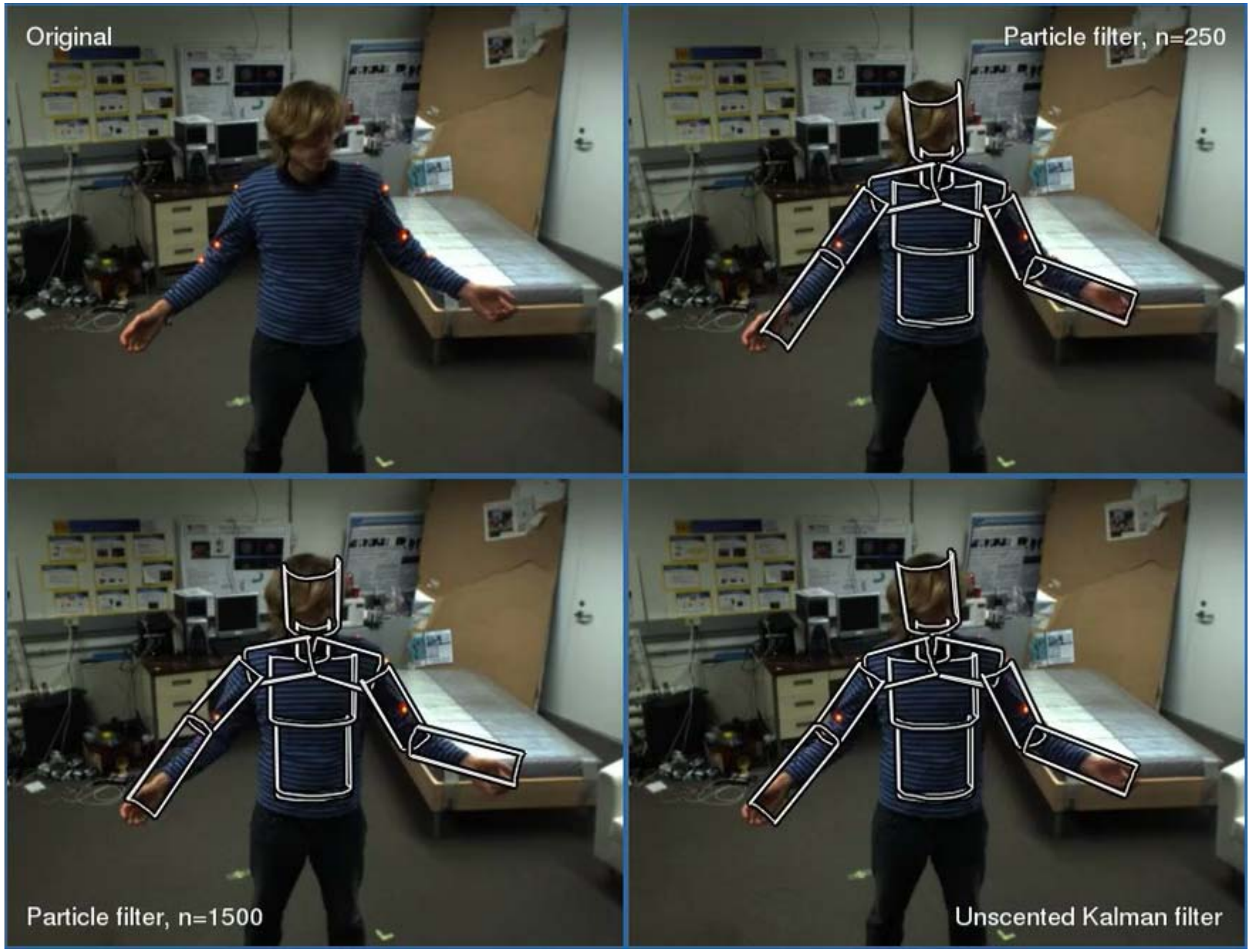
http://www.youtube.com/watch?v=2uDnW4AtFiE&feature=player_embedded

Marker based motion capture



[http://mocap.cs.cmu.edu/]

Marker-less motion capture: Using a stereo camera
(Markers for ground truth – NOT for tracking)

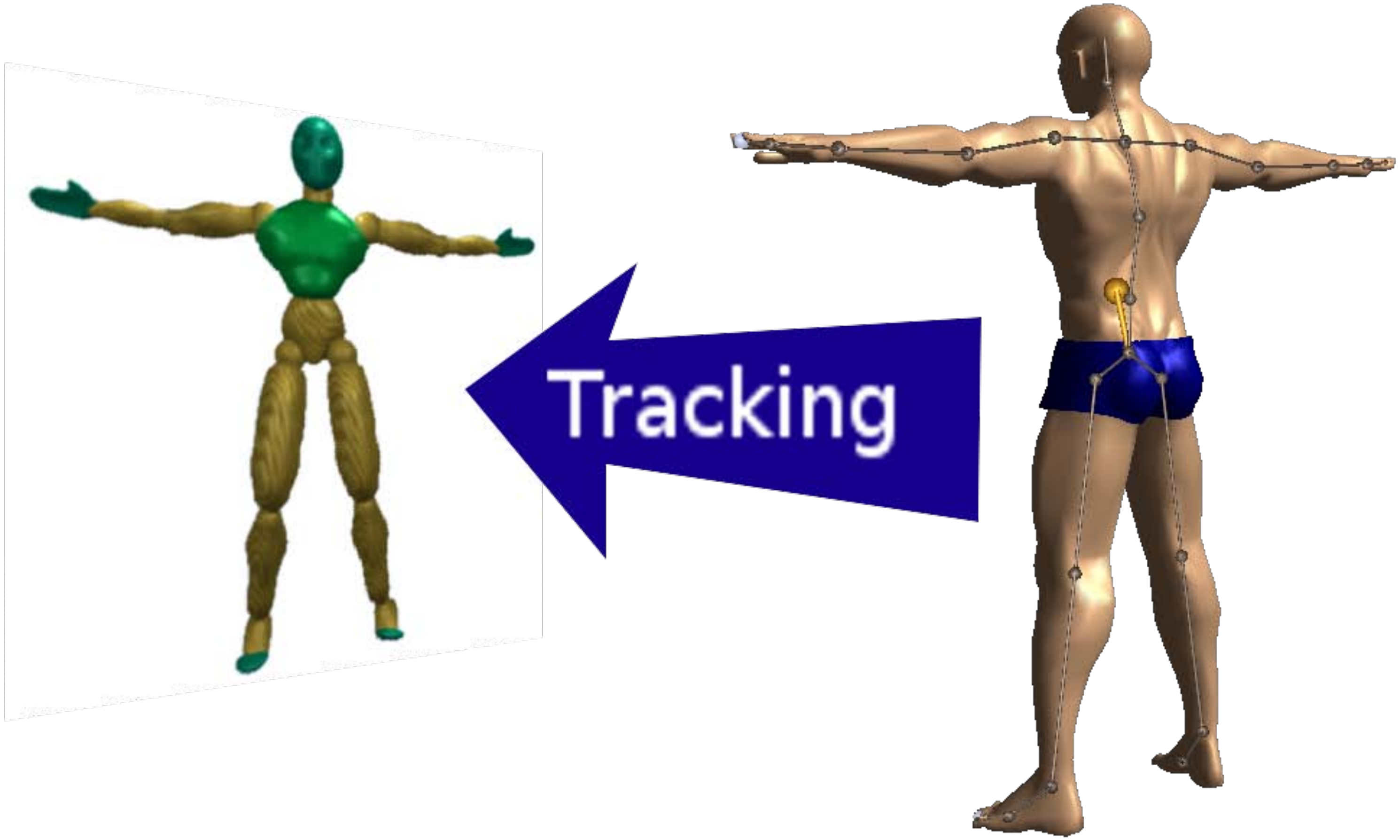


Why do we want to do tracking of human motion?

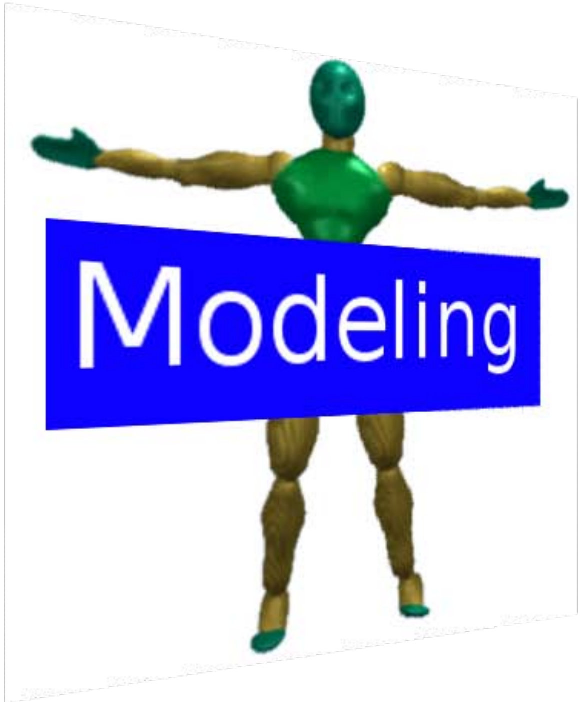
- Human computer interaction: Non-invasive interface technology
- Computer animation: Entertainment (movies and games), education, visualization
- Human motion analysis:
  - Surveillance: Suspicious behavior recognition, movement patterns
  - Biomechanical modeling
  - Physiotherapeutic analysis: Sports performance enhancement, patient treatment enhancement

Interactive physiotherapy: Interactive training with feedback at home

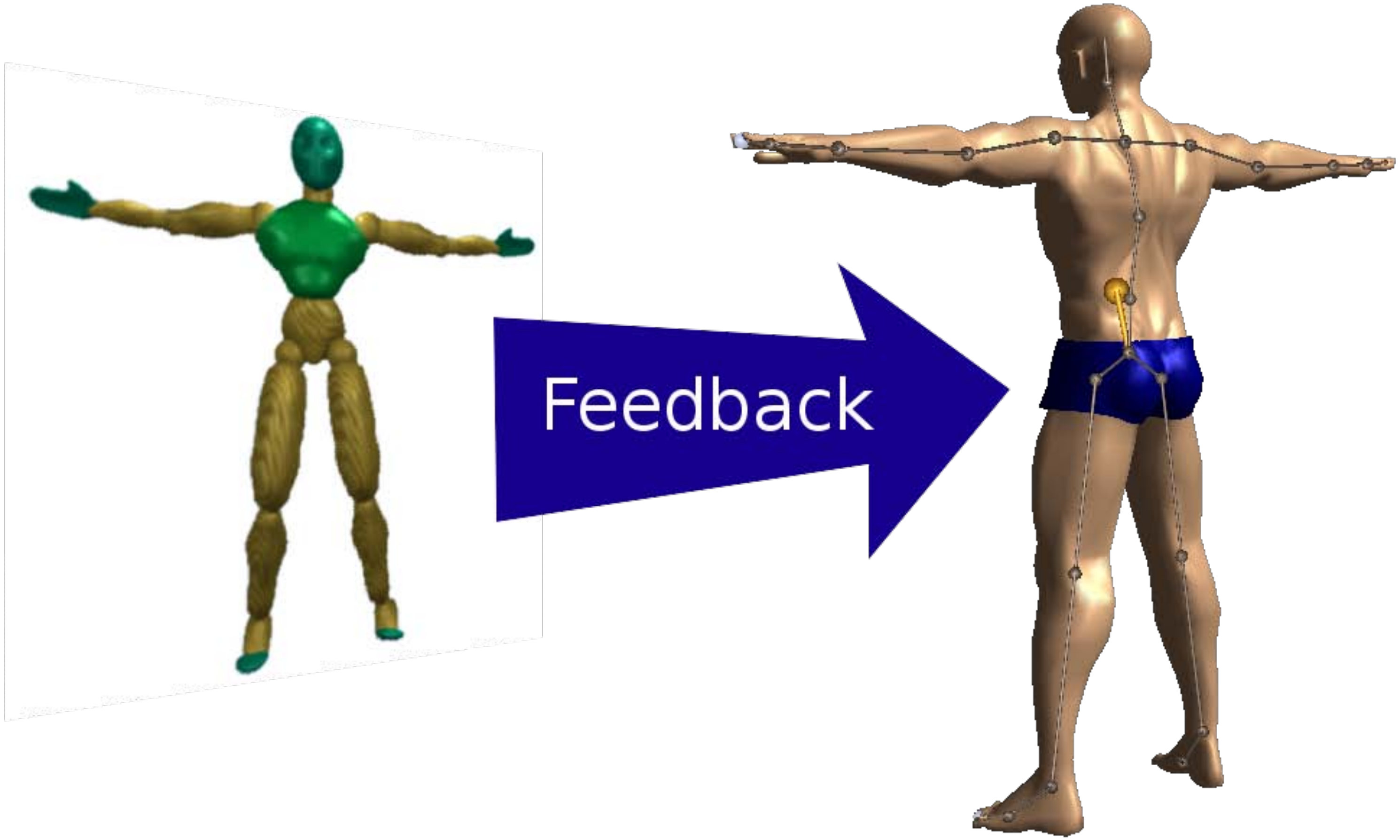
Vision for interactive physiotherapy: Measurements



Vision for interactive physiotherapy: Modeling of exercise



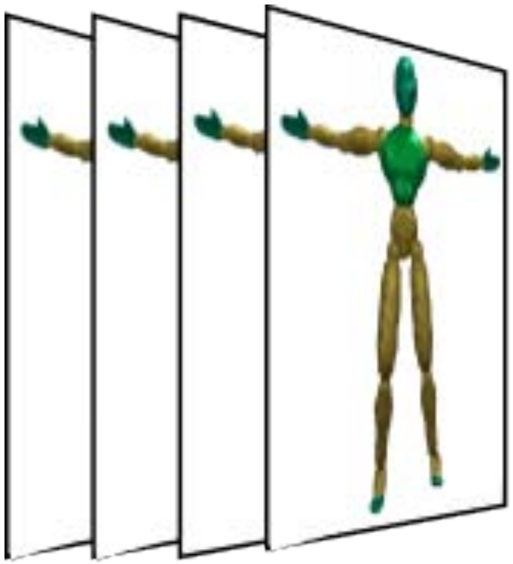
Vision for interactive physiotherapy: Feedback ("lift your arms higher")



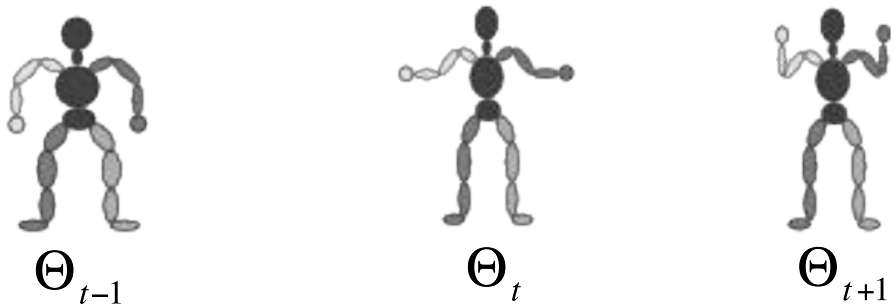
Demo

This is what we want to do
(Tracking = seq. estimation of pose from observations)

Video sequence $I_{1:t}$



Sequence of estimated poses



Approaches to visual marker-less tracking

- Model-based tracking
  - Introduce a model of the human body and estimate pose by estimating model parameters from observations.
  - Downsides: Relation between model and observation is important. Often these methods require re-initialization.

- Model-free tracking
  - Learning-based approach: Learn a direct map from observations to pose (e.g. a classifier)
  - Examplar-based approach: Infer pose by matching observations to examplar poses.
  - Downsides: Requires lots of training examples to either learn or construct examplar database.

Observations / Measurements

Types of observations:

- Features
- Texture / regions descriptors
- Silhouette
- 3D reconstruction / depth data (stereo, time-of-flight, active stereo (e.g. MS Kinect)).

Observations: Region-based texture descriptors

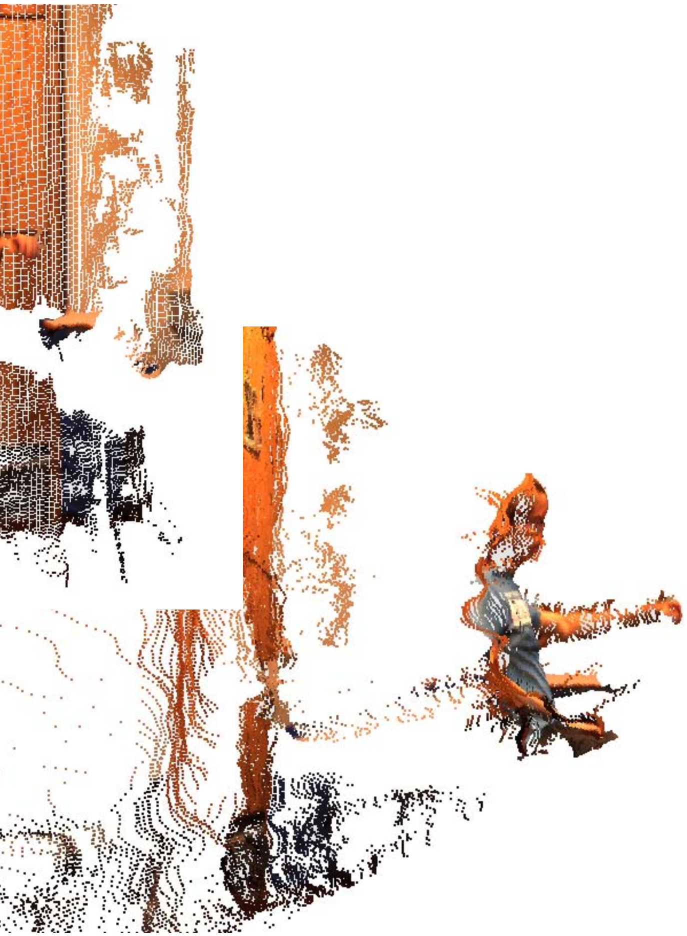
Observations: Silhouettes

Elgammal & Lee: T-PAMI, 2009

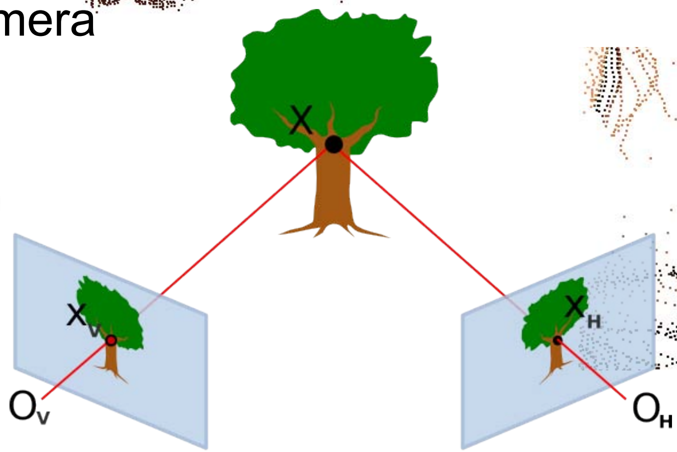
Using dense depth maps as visual observations

- Stereo cameras lead to dense depth maps and 3D point clouds.

We use a Point Grey Bumblebee2 stereo camera



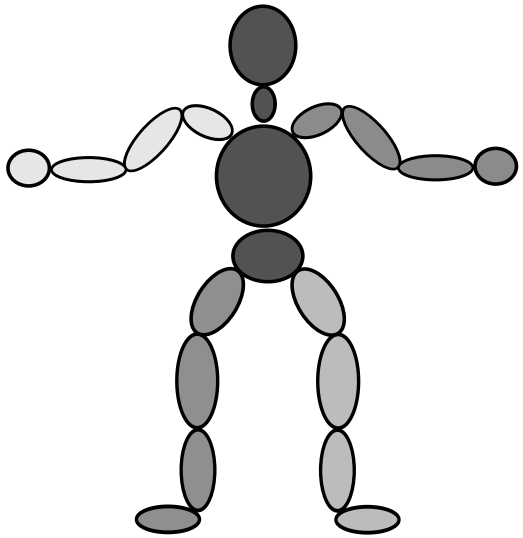
MS Kinect sensor



Show live depth map

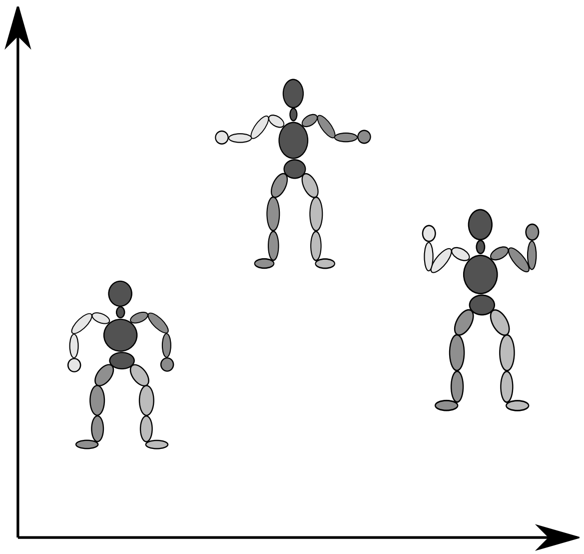
Human body model
(The model-based approach)

- The human body is commonly modeled as an articulated collection of rigid limbs connected with joints.
- **Common representation:**
  - Vector $\Theta = \left[\theta_1, \ldots, \theta_D\right]^T$ of joint angles together with some representation of global position and orientation.
  - Geometric shapes for modeling limb surface (boxes, ellipsoids).

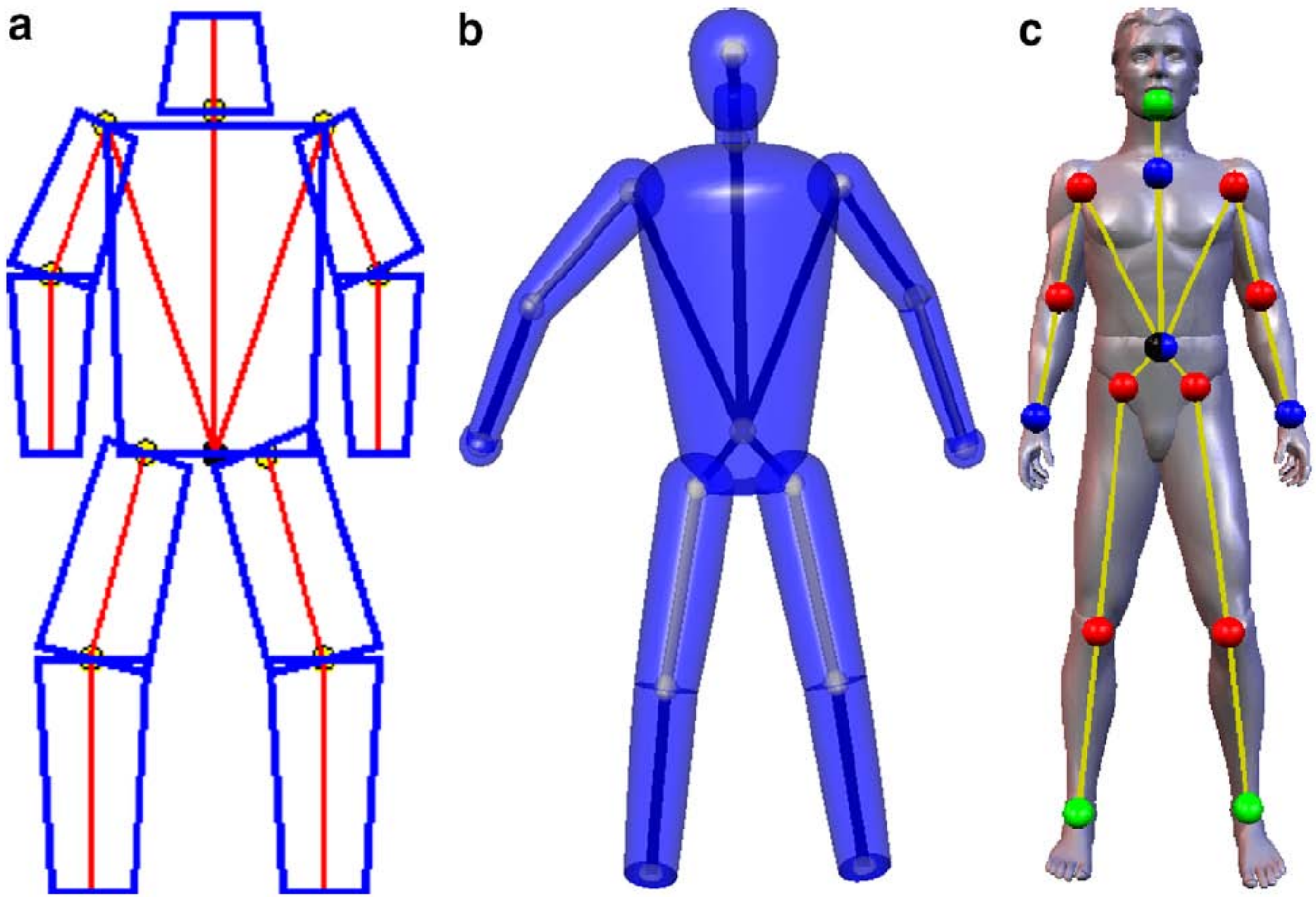


[Hauberg et al, 2009]

- Other representations:
  - Joint positions
  - End-effector positions
  - Pure surface models
  - …

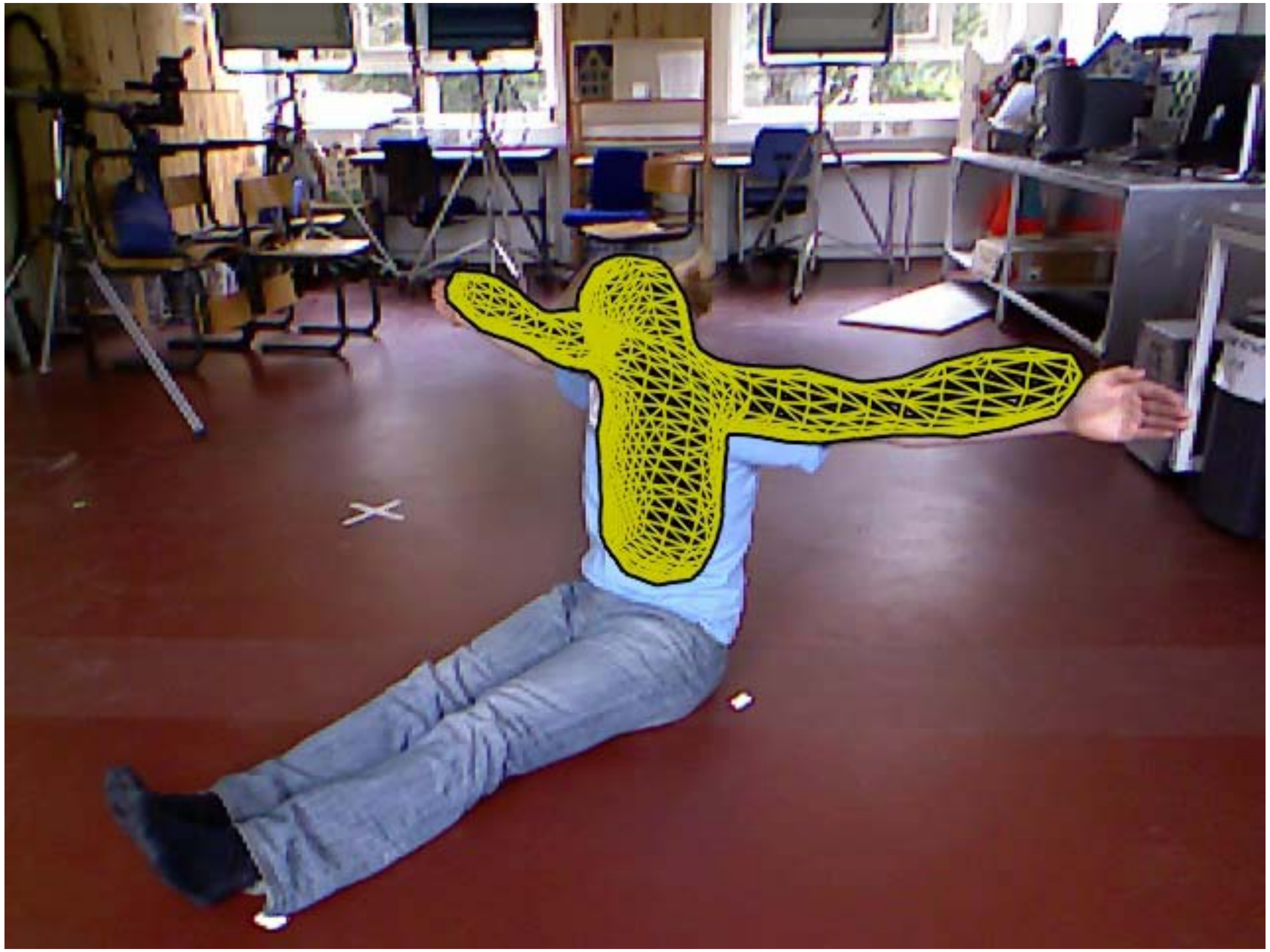


Modeling the skin surface



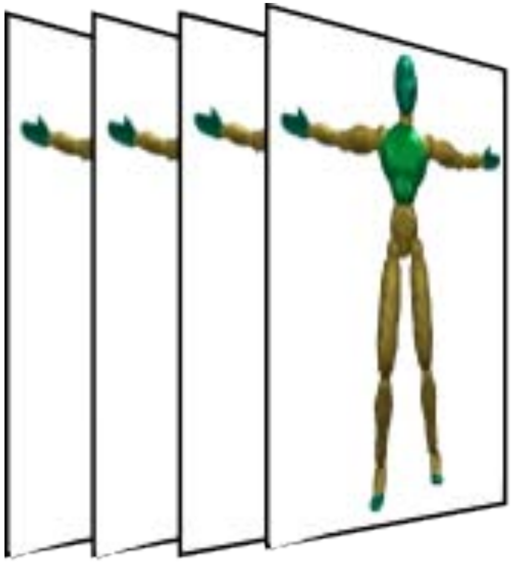
Poppe: CVIU, 2007

A polygonal mesh skin model in action

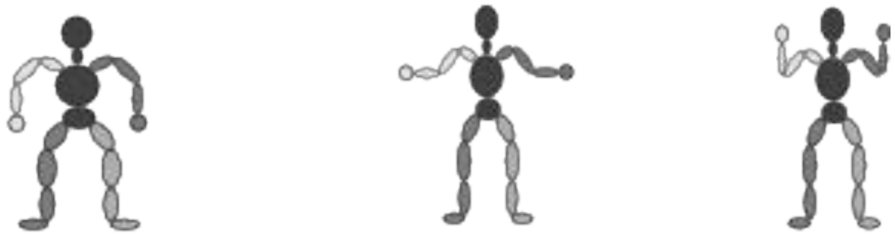


This is what we want to do
(Tracking = seq. estimation of model from observations)

Video sequence $I_{1:t}$

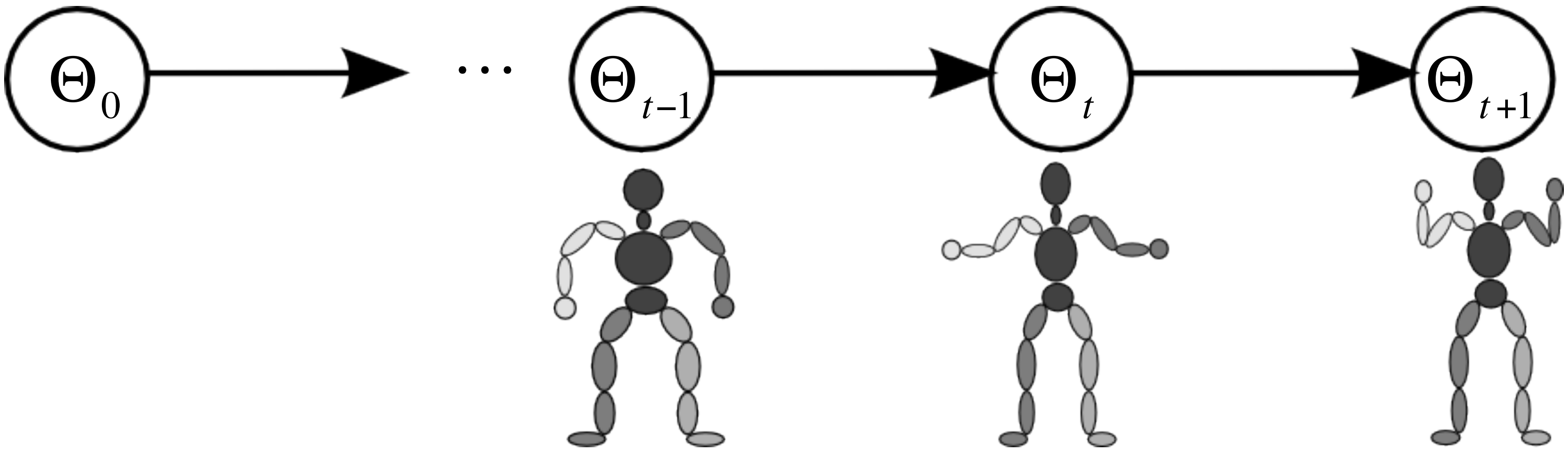


Sequence of estimated poses



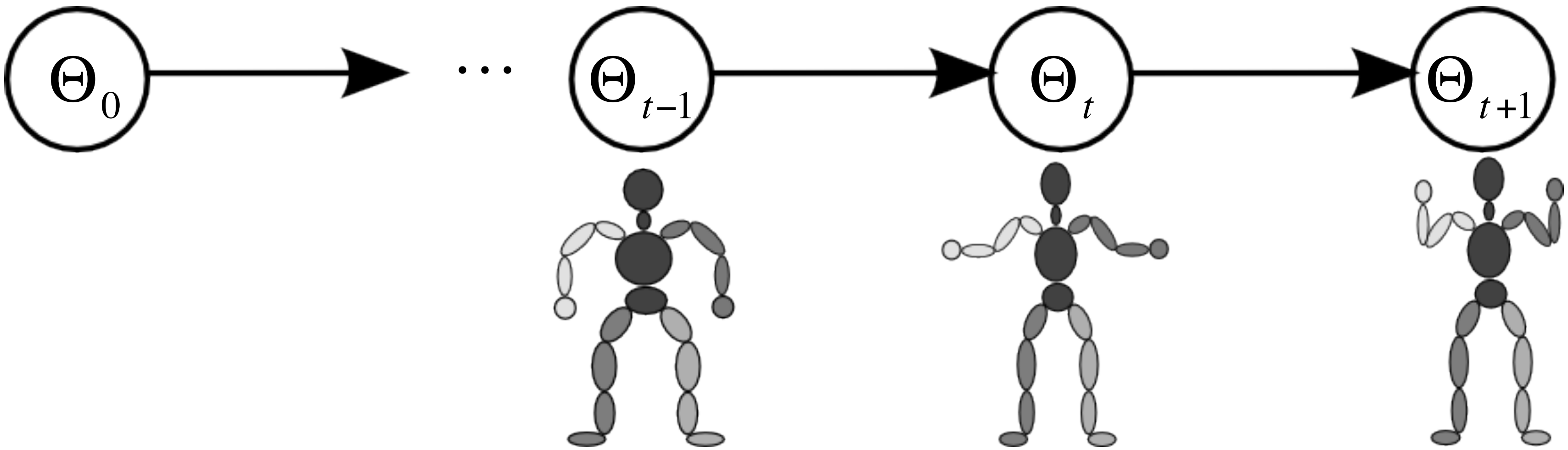
Θ_{t-1} Θ_{t} Θ_{t+1}

How do we include dynamics into our model?



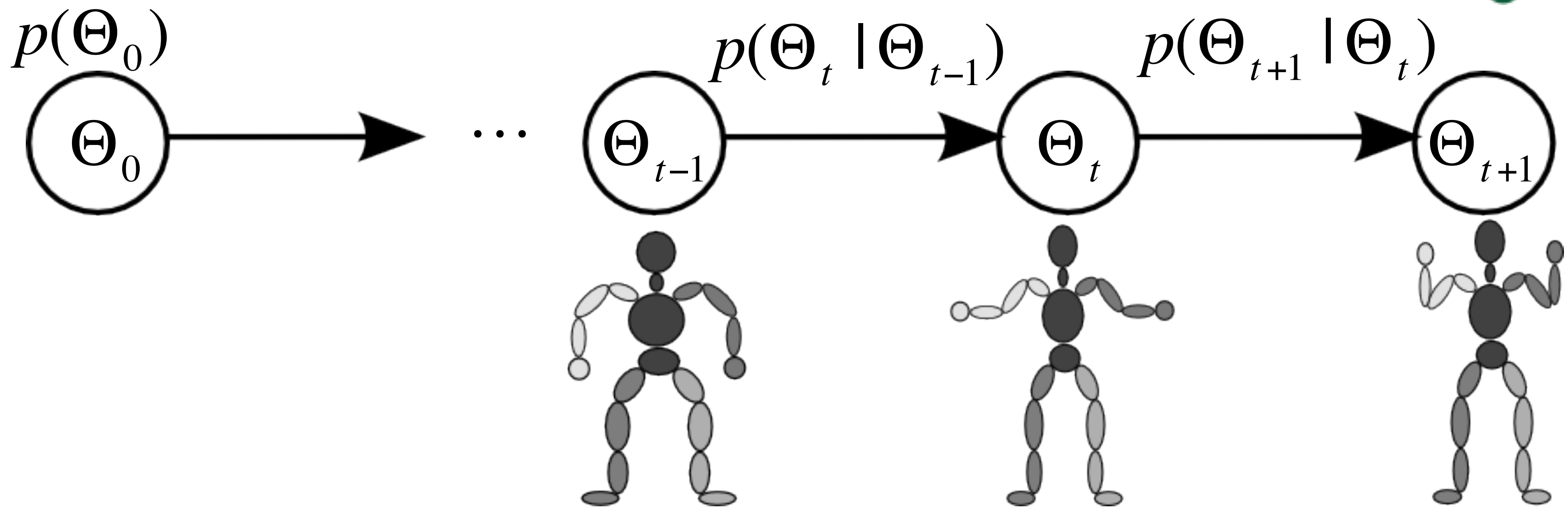
- Lets assume that the current state only depends on the immediate past state. We assume that somehow we can compute the new state given the old state!
- However, this update of states is stochastic (uncertain) – we are going to estimate it from noisy observations.

Enter probabilistic graphical models



- This is an example of a simple graphical model:
  - First order Markov chain
  - A directed acyclic graph (DAG) or tree, if you will.
- **Def. Graphical model:** Graph based model where *nodes* represent random quantities / variables and *edges* represents dependencies among variables.

Enter probabilistic graphical models



- If we know the transition conditional probability distributions and the prior distribution on the initial state, we can compute the probability distribution of state sequences (of particular sequences of stick figures):

$$p(\Theta_0,\ldots,\Theta_t) = p(\Theta_{0:t}) = p(\Theta_0)\prod_{i=1}^{t} p(\Theta_i \mid \Theta_{i-1})$$

Short-hand notation:

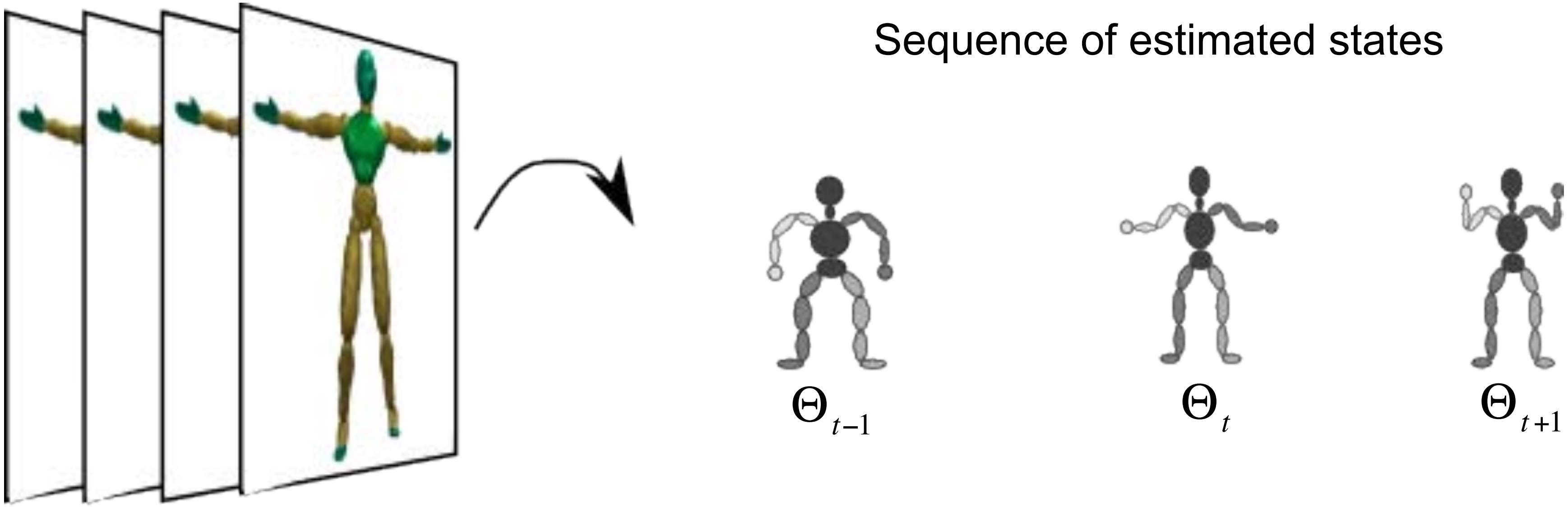
$$\Theta_{0:t} = \{\Theta_0,\ldots,\Theta_t\}$$

How to relate the model state with observations?

(Tracking = estimation of model from observations)

Video sequence $I_{1:t}$

Sequence of estimated states



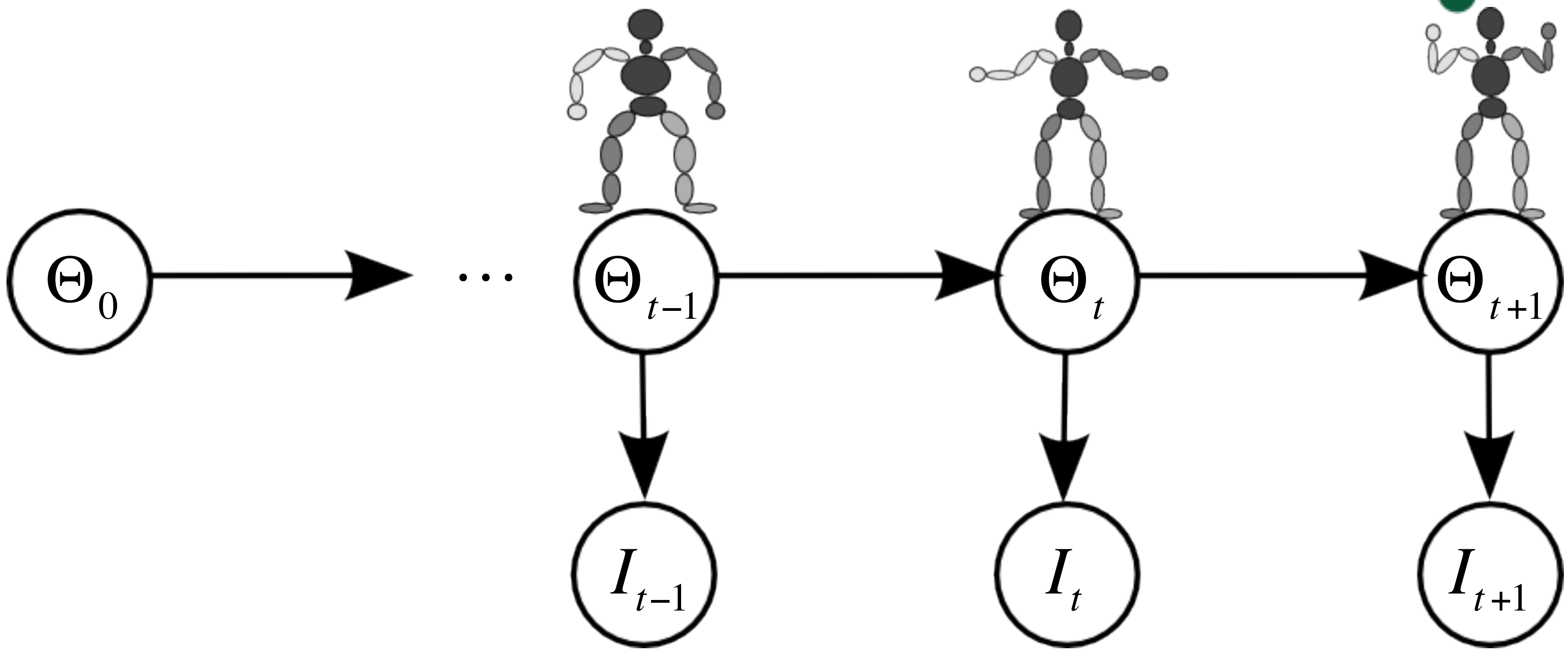
What do the image of a particular stick figure look like? Hard problem!

Lets introduce a potentially non-linear function for "*drawing the stick figure*" in image space and compare with the observed image. Something like this,

$$\left\| I_t - F(\Theta_t) \right\|^2$$

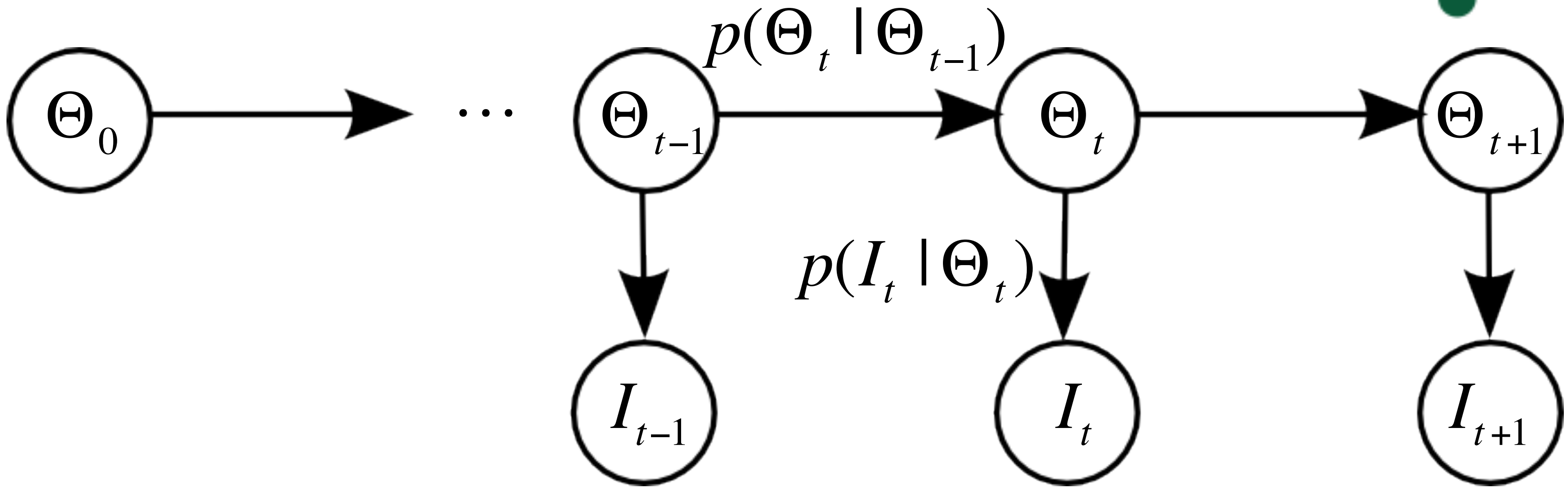
However our observations are noisy so we want a probabilistic model for this!

Enter hidden Markov models (HMM)



- This model states that we only observe the images directly and the states indirectly – they are hidden (latent variables). If states are discrete we have a HMM.
- First order Markov chain in the states.

We need a probabilistic observation model

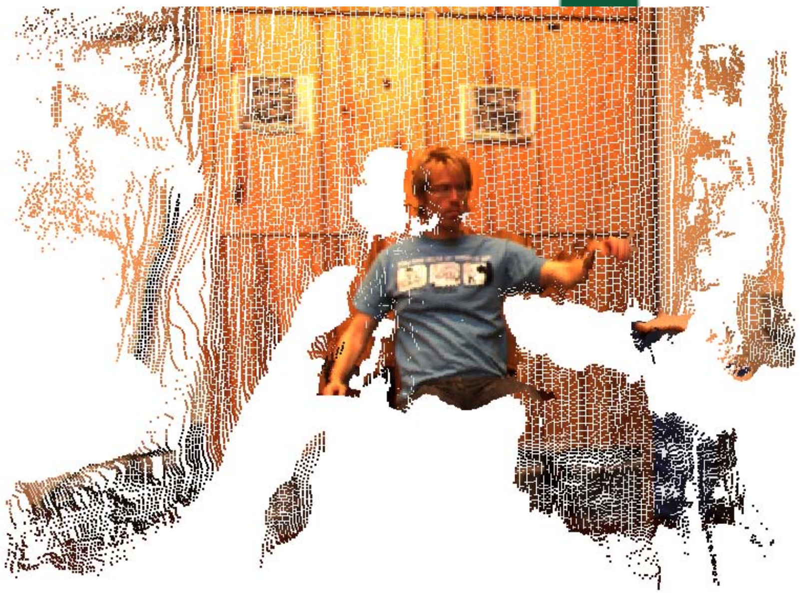


How about a Gaussian observation model?

$$p(I_t \mid \Theta_t) = \frac{1}{Z}\exp\left(-\frac{\left\| I_t - F(\Theta_t) \right\|^2}{2\sigma^2}\right)$$

Or perhaps more useful $\quad p(I_t \mid \Theta_t) = \frac{1}{Z}\exp\left(-H(I_t, \Theta_t)\right)$

Observational model using depth maps



- Observations are collections of 3D stereo points:

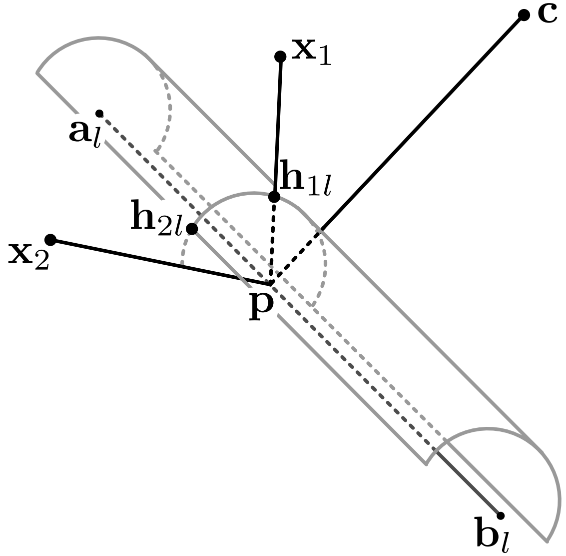
$$\mathbf{X}_t = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

- Introduce simple skin model: A capsule per bone.
- Observational model:

$$p(\mathbf{X}_t \mid \Theta_t) = \prod_{n=1}^{N} p(\mathbf{x}_n \mid \Theta_t)$$

$$p(\mathbf{x}_n \mid \Theta_t) \propto \exp\left(-\frac{D^2(\mathbf{x}_n, \Theta_t)}{2\sigma^2}\right)$$

$$p(I_t \mid \Theta_t) \equiv p(\mathbf{X}_t \mid \Theta_t)$$



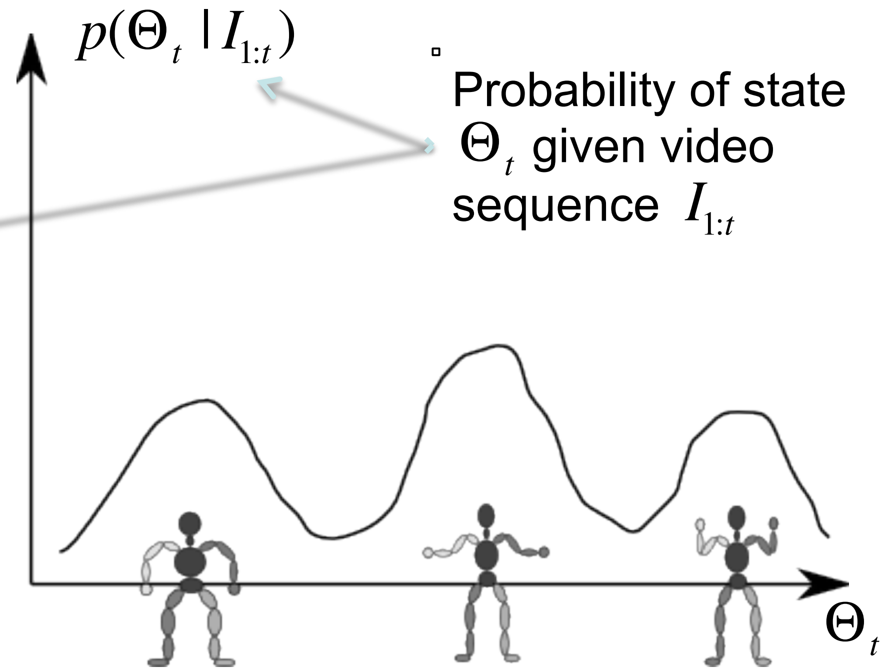
How to do tracking (estimation of pose)?

- The model gives us the joint distribution

$$p(I_{1:t}, \Theta_{0:t}) = p(\Theta_0) \prod_{i=1}^{t} p(I_i \mid \Theta_i)\, p(\Theta_i \mid \Theta_{i-1})$$

- If we want to do real-time tracking we need

$$p(\Theta_t \mid I_{1:t})$$



- And then take averages to compute a prediction of the current state.

Probability theory crash course

- by applying the sum and product rules we have

$$p(\Theta_t \mid I_{1:t}) = \frac{p(I_{1:t},\Theta_t)}{p(I_{1:t})}$$

and

$$p(I_{1:t},\Theta_t) = \int p(I_{1:t},\Theta_{0:t})d\Theta_{0:t-1}$$

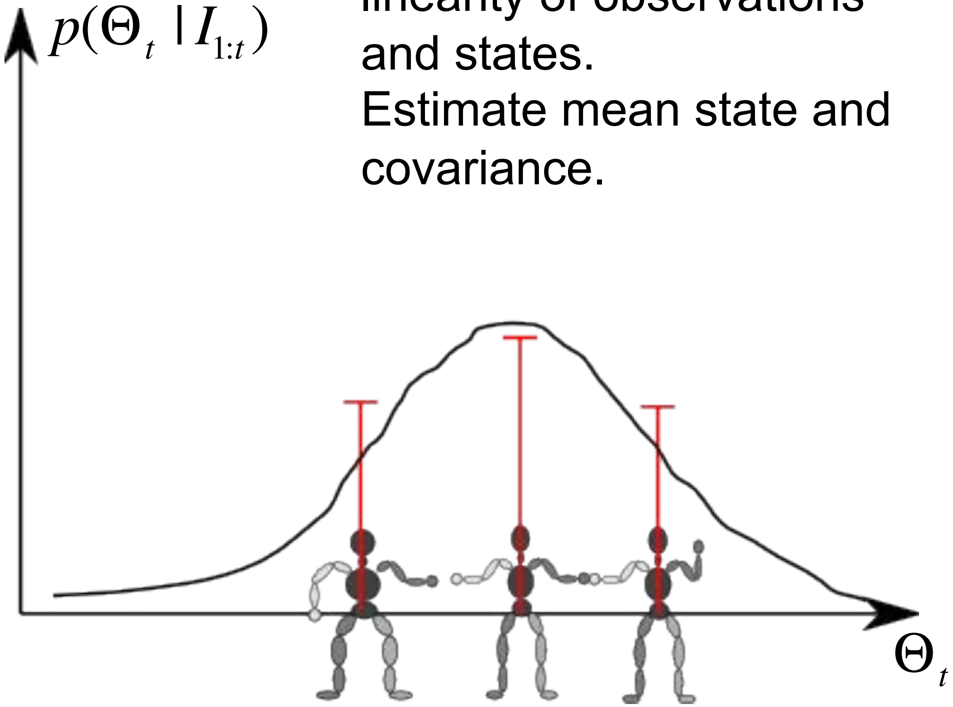
$$p(I_{1:t}) = \int p(I_{1:t},\Theta_{0:t})d\Theta_{0:t}$$

Hence what we need can be derived from the joint distribution.

So we need to sequentially estimate $p(\Theta_t \mid I_{1:t})$
(Details covered in Advanced topics in data modeling)

Kalman filtering

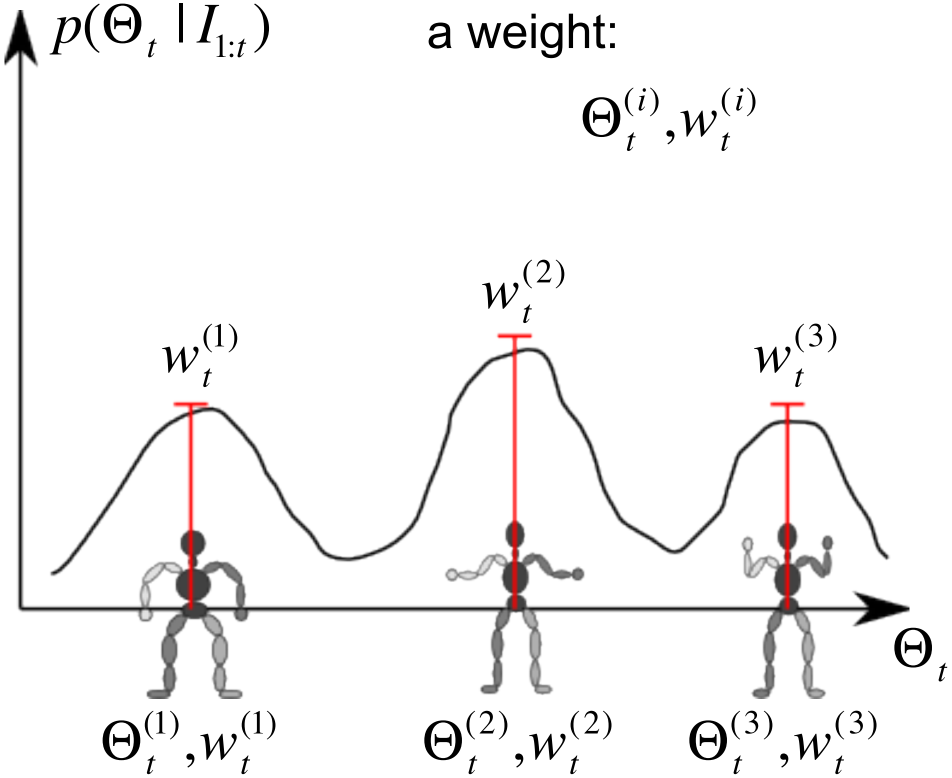
- Assume Gaussianity and linearity of observations and states.
  Estimate mean state and covariance.



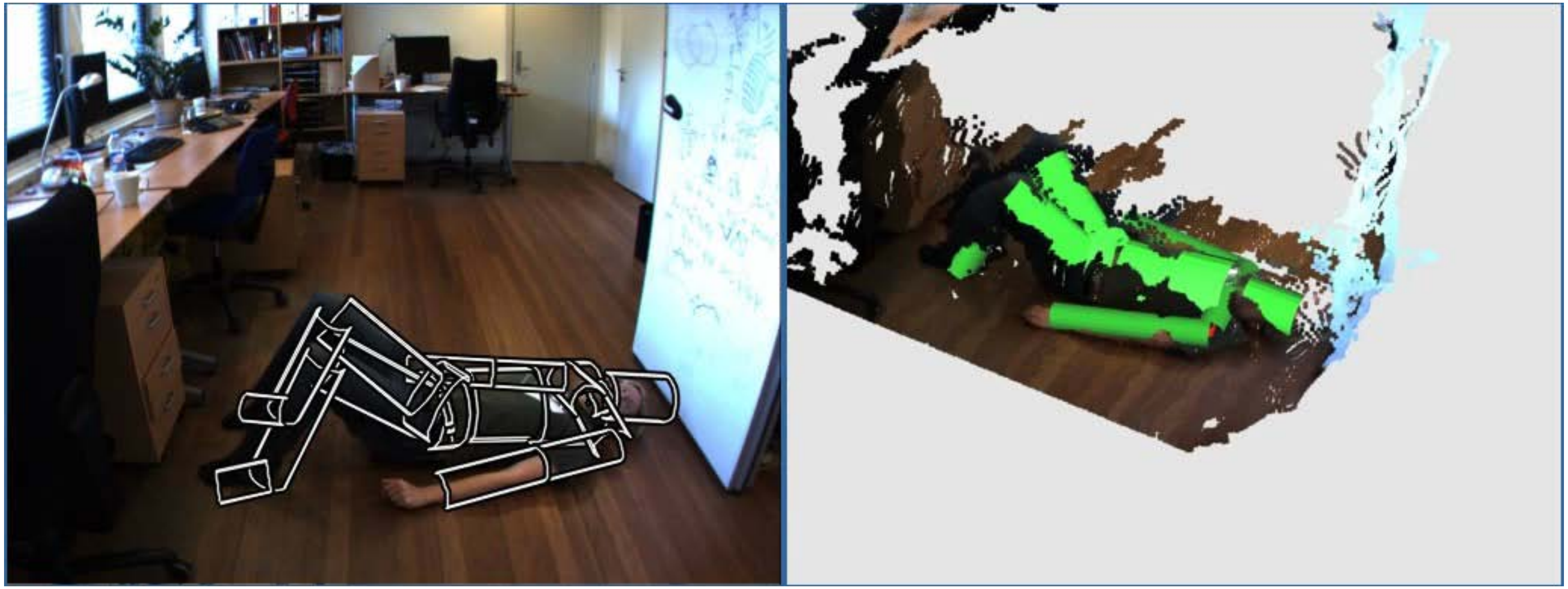
Particle filtering

- Particles: A state and a weight:

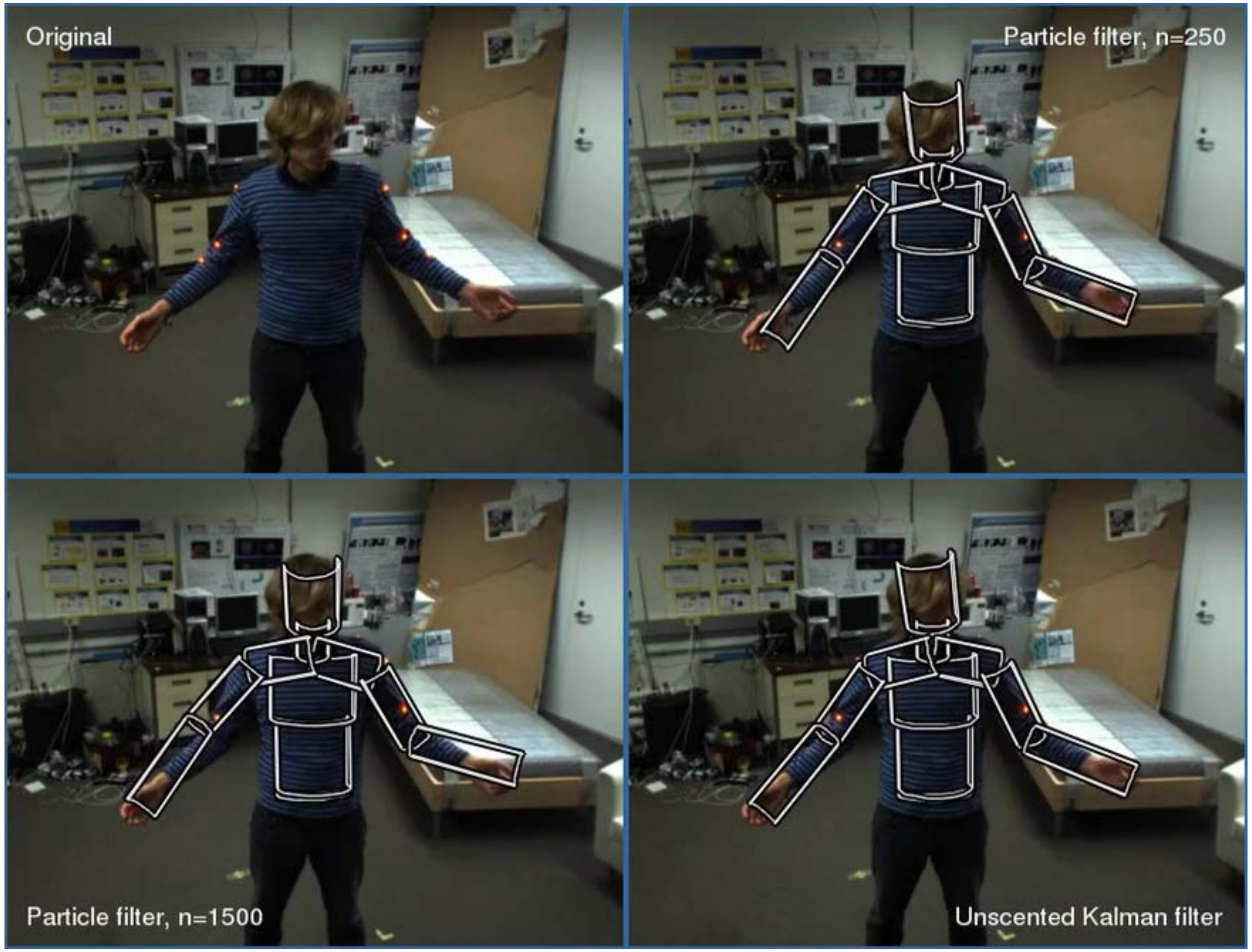
  $\Theta_t^{(i)}, w_t^{(i)}$



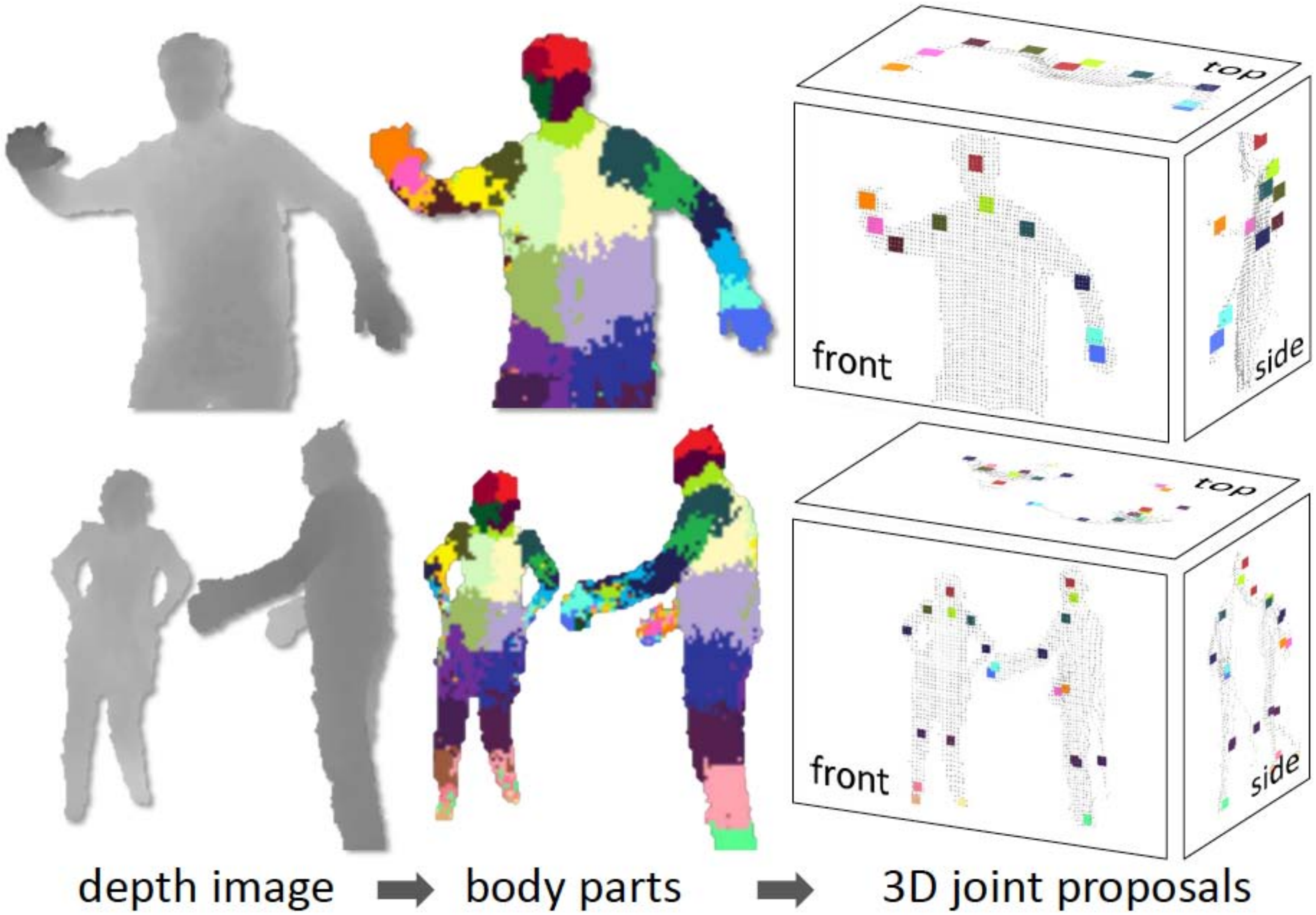
Voila! We now have a visual tracker



And here is another sequence showing different inference methods



Per frame part-based recognition of pose
(A model-free approach used in MS Kinect SDK)



Shotton et al.: CVPR 2011

Summary

- Visual tracking of 3D human motion:
  - Marker-based
  - Marker-less

- Marker-less tracking:
  - Model-based (what we saw in detail)
  - Model-free

- Model-based tracking requires:
  - Human body model
  - Relation between model and observations
  - Tracking is sequential filtering of the predicted pose (too avoid jitter in the estimated pose).

Literature

Reading material:

- R. Poppe: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding, 108 (1-2): 4-18, 2007.

Additional material:

- J. Shotton et al: Real-Time Human Pose Recognition in Parts from Single Depth Images. Proceedings of CVPR, 1297-1304, 2011.
- S. Hauberg and K. Steenstrup Pedersen: Predicting Articulated Human Motion from Spatial Processes. International Journal of Computer Vision, 94(3): 317-334, 2011.
- R. Poppe: A survey on vision-based human action recognition. Image and Vision Computing, 28(6): 976-990, 2010.

Remember to fill in the students course evaluation