Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences

Presentation of the thesis work of:
Hedvig Sidenbladh, KTH

Thesis opponent: Prof. Bill Freeman, MIT

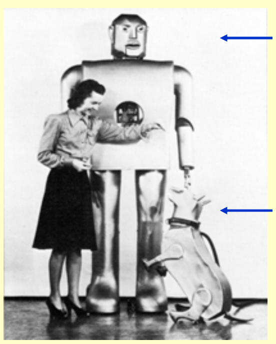
Thesis supervisors

- Prof. Jan-Olof Eklundh, KTH
- Prof. Michael Black, Brown University

Collaborators

- Dr. David Fleet, Xerox PARC
- Prof. Dirk Ormoneit, Stanford University

A vision of the future from the past.



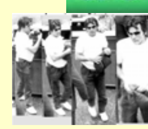
Elektro

Sparky

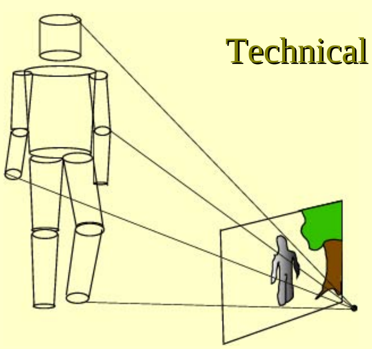
New York Worlds Fair, 1939
(Westinghouse Historical Collection)

Applications of computers looking at people

- Human-machine interaction
  - Robots
  - Intelligent rooms
- Video search
- Entertainment: motion capture for games, animation, and film.
- Surveillance



Technical Goal



Tracking a human in 3D

Why is it Hard?

The appearance of people can vary dramatically.

Why is it hard?

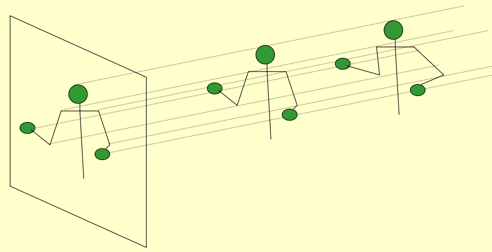


People can appear in arbitrary poses.

Structure is unobservable—
inference from visible parts.

Why is it hard?

Geometrically under-constrained.



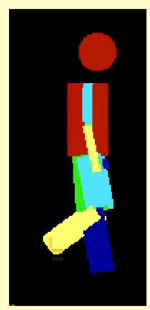
One solution:

- Use markers
- Use multiple cameras

http://www.vicon.com/animation/

State of the Art.

Bregler and Malik '98



- Brightness constancy cue
  - Insensitive to appearance
- Full-body required multiple cameras
- Single hypothesis

2D vs. 3D tracking

- Artist's models...

State of the Art.

Cham and Rehg '99

- Single camera, multiple hypotheses
- 2D templates (no drift but view dependent)

$I(\mathbf{x}, t) = I(\mathbf{x}+\mathbf{u}, 0) + \eta$

1999 state of art

(a) (b) (c)

Figure 4: (a) Tracker (in white) using constant velocity predictor drifts off track by frame 7. (b) SLDS-based tracker is on track at frame 7. Model (switching state) 3 has the highest likelihood. Black lines show prior mean and observation. (c) SLDS tracker at frame 20.

Pavlovic, Rehg, Cham, and Murphy, Intl. Conf. Computer Vision, 1999

State of the Art.

Deutscher, North, Bascle, & Blake '00

- Multiple hypotheses
- Multiple cameras
- Simplified clothing, lighting and background

Note: we can fake it with clever system design

M. Krueger, "Artificial Reality", Addison-Wesley, 1983.

Game videos...

Decathlete 100m hurdles

Black background

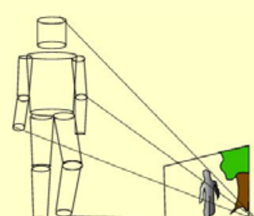
No other people in camera

Person at known distance and position.

Display tells person what motion to do.

Performance specifications

* No special clothing
* Monocular, grayscale, sequences (archival data)
* Unknown, cluttered, environment



Task: Infer 3D human motion from 2D image

Bayesian formulation

$$\boxed{p(\text{model} \mid \text{cues})} = \frac{\boxed{p(\text{cues} \mid \text{model})}\boxed{p(\text{model})}}{p(\text{cues})}$$

1. Need a constraining *likelihood* model that is also invariant to variations in human appearance.
2. Need a *prior* model of how people move.
3. *Posterior probability*: Need an effective way to explore the model space (very high dimensional) and represent ambiguities.

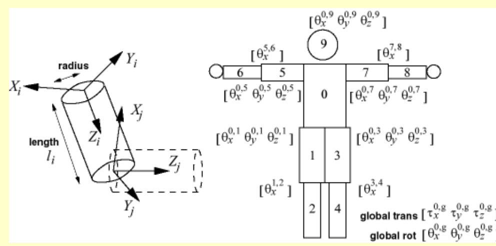
System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

Simple Body Model

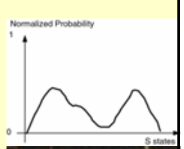


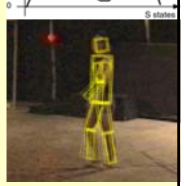
* Limbs are truncated cones

* Parameter vector of joint angles and angular velocities = ϕ

Multiple Hypotheses

- Posterior distribution over model parameters often multi-modal (due to ambiguities)
- Represent whole distribution:
  - sampled representation
  - each sample is a pose
  - predict over time using a particle filtering approach





Particle Filter

Posterior $p(\phi_{t-1} \mid \vec{\mathrm{I}}_{t-1})$ $\xrightarrow{\text{sample}}$ **Temporal dynamics** $p(\phi_t \mid \phi_{t-1})$

$\xrightarrow{\text{sample}}$ $p(\mathrm{I}_t \mid \phi_t)$ **Likelihood**

$\xrightarrow{\text{normalize}}$ $p(\phi_t \mid \vec{\mathrm{I}}_t)$ **Posterior**

Problem: Expensive represententation of posterior!
Approaches to solve problem:
• Lower the number of samples. (Deutsher et al., CVPR00)
• Represent the space in other ways (Choo and Fleet, ICCV01)

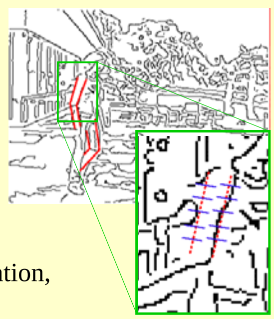
System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

What do people look like?

What do non-people look like?

Edge Detection?



- Probabilistic model?
- Under/over-segmentation, thresholds, …

Key Idea #1 (Likelihood)

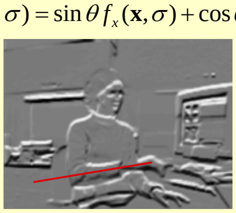
1. Use the 3D model to predict the location of limb boundaries (not necessarily features) in the scene.
2. Compute various filter responses *steered* to the predicted orientation of the limb.
3. Compute likelihood of filter responses using a statistical model *learned from examples*.

Edge Filters

Normalized derivatives of Gaussians (Lindeberg, Granlund and Knutsson, Perona, Freeman&Adelson, …)

Edge filter response steered to limb orientation:

$$f^e(\mathbf{x},\theta,\sigma) = \sin\theta\, f_x(\mathbf{x},\sigma) + \cos\theta\, f_y(\mathbf{x},\sigma)$$



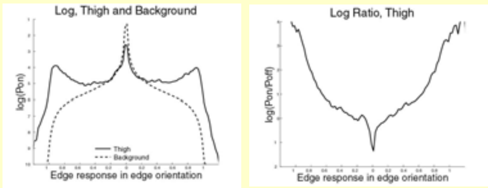
Filter responses steered to arm orientation.

Example Training Images

Edge Distributions

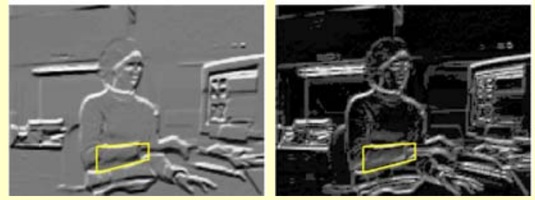
Edge response steered to model edge:

$$f_e(\mathbf{x},\theta,\sigma) = \sin\theta\, f_x(\mathbf{x},\sigma) + \cos\theta\, f_y(\mathbf{x},\sigma)$$



Similar to Konishi et al., CVPR 99

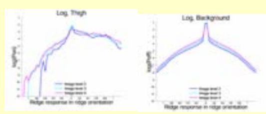
Edge Likelihood Ratio



Edge response Likelihood ratio

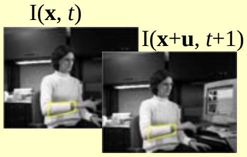
Other Cues

Ridges

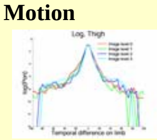


I($\mathbf{x}$, t)

I($\mathbf{x}$+$\mathbf{u}$, t+1)



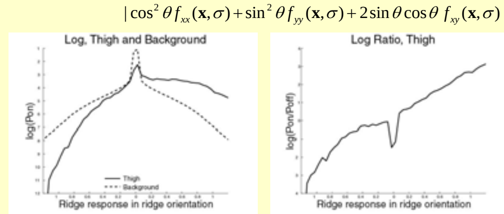
Motion



Ridge Distributions

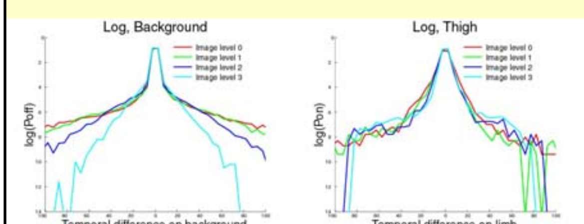
Ridge response steered to limb orientation

$$f_r(\mathbf{x},\theta,\sigma) = \ |\sin^2\theta\, f_{xx}(\mathbf{x},\sigma) + \cos^2\theta\, f_{yy}(\mathbf{x},\sigma) - 2\sin\theta\cos\theta\, f_{xy}(\mathbf{x},\sigma)| - |\cos^2\theta\, f_{xx}(\mathbf{x},\sigma) + \sin^2\theta\, f_{yy}(\mathbf{x},\sigma) + 2\sin\theta\cos\theta\, f_{xy}(\mathbf{x},\sigma)|$$



Ridge response only on certain image scales!

Motion distributions



Different underlying motion models

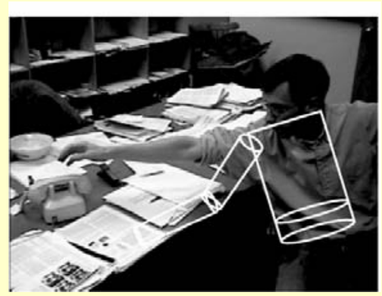
Likelihood Formulation

- Independence assumptions:
  - Cues: p(image | model) = p(cue1 | model) p(cue2 | model)
  - Spatial: p(image | model) = $\prod_{x \in \text{image}}$ p(image(x) | model)
  - Scales: p(image | model) = $\prod_{\sigma=1,\ldots}$ p(image(σ) | model)
- Combines cues and scales!
- Simplification, in reality there are dependencies

The power of cue combination

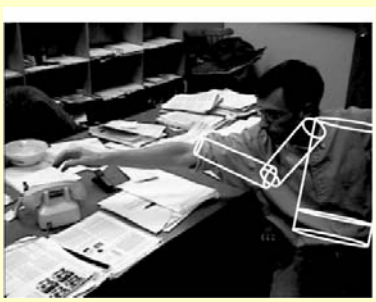
Using edge cues alone

Edge cues

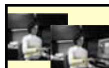


Using ridge cues alone

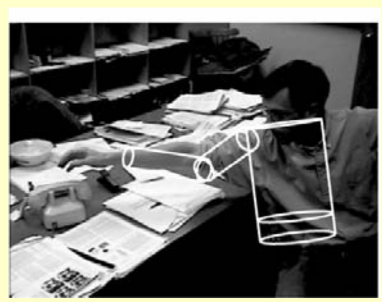
Ridge cues



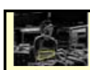
Using flow cue alone



Flow cues

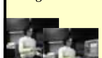


Using edge, ridge, and motion cues together

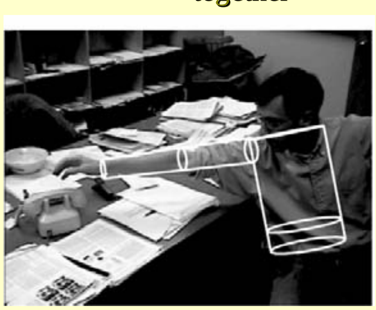


Edge cues

Ridge cues



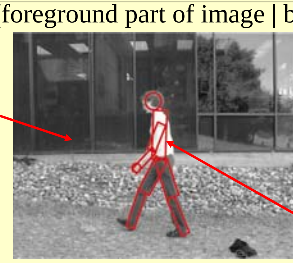
Flow cues



Key Idea #2

$$p(\text{image} \mid \text{foreground, background}) \propto \frac{p(\text{foreground part of image} \mid \text{foreground})}{p(\text{foreground part of image} \mid \text{background})}$$

Do not look in parts of the image considered background



Foreground part of image

Likelihood

$$p(image \mid fore, back) = \prod_{fore\ pixels} p(image \mid fore) \prod_{back\ pixels} p(image \mid back)$$

$$= \frac{\prod_{all\ pixels} p(image \mid back) \prod_{fore\ pixels} p(image \mid fore)}{\prod_{fore\ pixels} p(image \mid back)}$$

$$= \frac{const \prod_{fore\ pixels} p(image \mid fore)}{\prod_{fore\ pixels} p(image \mid back)}$$

Foreground pixels

Background pixels

System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

The Prior term

Bayesian formulation:

$$p(\text{model} \mid \text{cue}) \propto p(\text{cue} \mid \text{model})\, p(\text{model})$$

- Need a constraining likelihood model that is also invariant to variations in human appearance
- Need a good model of how people move

Very general model

- Constant velocity motions
- Not constrained by how people tend to move.

Constant velocity model

- All DOF in the model parameter space, ϕ, independent
- Angles are assumed to change with constant speed
- Speed and position changes are randomly sampled from normal distribution

Tracking an Arm

1500 samples
~2 min/frame

Moving camera, constant velocity model

Self Occlusion

1500 samples
~2 min/frame

Constant velocity model

System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

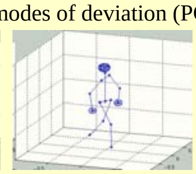
Very specific model

- Only handles people walking.
- Very powerful constraint on human motion.

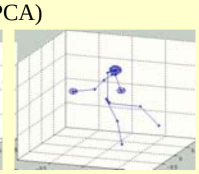
Models of Human Dynamics

- Action-specific model - Walking
  - Training data: 3D motion capture data
  - From training set, learn mean cycle and common modes of deviation (PCA)

Mean cycle



Small noise



Large noise

Walking Person

#samples from 15000 to 2500 by using the learned likelihood

2500 samples
~10 min/frame

Walking model

No likelihood

* how strong is the walking prior?
(or is our likelihood doing anything?)

System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

Example-based model

- Take lots of training data.
- Use "snippets" of the data as models for how people are likely to move.

Example-based model

Ten samples from the prior, drawn using approximate probabilistic tree search.

Tracking with only 300 particles.

Smooth motion prior.

Example-based motion prior.

Lessons Learned

- **Representation for probabilistic analysis.**
  - Probabilistic (Bayesian) framework allows
    - Integration of information in a principled way
    - Modeling of priors
  - Particle filtering allows
    - Multi-modal distributions
    - Tracking with ambiguities and non-linear models
- Models for human appearance (likelihood term).
- Models for human motion (prior term).

Lessons Learned

- Representation for probabilistic analysis.
- **Models for human appearance (likelihood term).**
  - Generic, learned, model of appearance
    - Combines multiple cues
    - Exploits work on image statistics
  - Use the 3D model to predict features
  - Model of foreground and background
    - Exploits the ratio between foreground and background likelihood
    - Improves tracking
- Models for human motion (prior term).

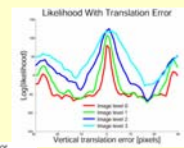
Lessons Learned

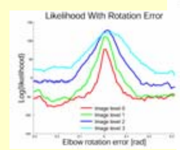
- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- **Models for human motion (prior term).**
  - Explored 3 different models; analyzed the tradeoffs between each.

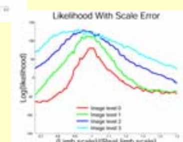
End

Decathlete javelin throw

Edges







Bayesian Inference

Exploit cues in the images. Learn *likelihood* models:

$$p(\text{image cue} \mid \text{model})$$

Build models of human form and motion. Learn *priors* over model parameters:

$$p(\text{model})$$

Represent the *posterior* distribution:

$$p(\text{model} \mid \text{cue}) \propto p(\text{cue} \mid \text{model})\, p(\text{model})$$