3D Articular Human Tracking from Monocular Video

From Condensation to Kinematic Jumps

Cristian Sminchisescu
Bill Triggs
GRAVIR-CNRS-INRIA, Grenoble

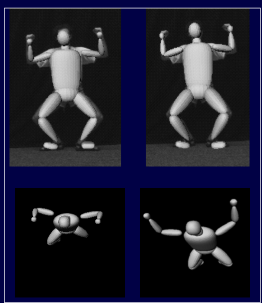
Goal: track human body motion in monocular video and estimate 3D joint motion

Why Monocular ?

- Movies, archival footage
- Tracking / interpretation of actions & gestures (HCI)
- Resynthesis, e.g. change point of view or actor
- How do humans do this so well?

Why is 3D-from-monocular hard?

Image matching ambiguities

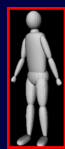


Depth ambiguities

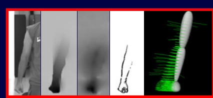
Violations of physical constraints

Overall Modelling Approach

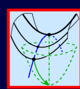
1. Generative Human Model
   - Complex, kinematics, geometry, photometry
   - Predicts images or descriptors



2. Model-image matching cost function
   - Associates model predictions to image features
   - Robust, probabilistically motivated

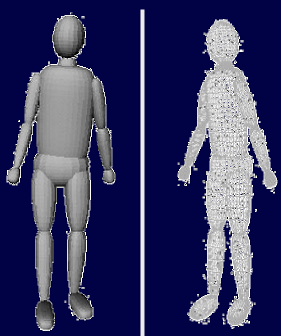


3. Tracking by search / optimization
   - Discovers well supported configurations of matching cost



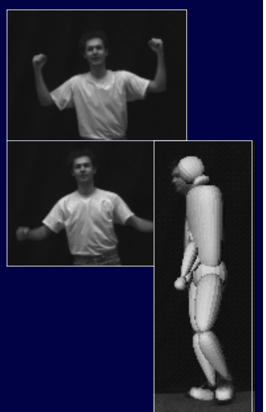
Human Body Model

- Explicit 3D model allows high-level interpretation
- 30-35 d.o.f. articular 'skeleton'
- 'Flesh' of superquadric ellipsoids with tapering & bending
- Model $\rightarrow$ image projection maps points on 'skin' through
  - kinematic chain
  - camera matrix
  - occlusion (z buffer)

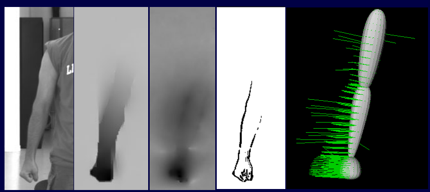


Parameter Space Priors

- Anthropometric prior
  - left/right symmetry
  - bias towards default human
- Accurate kinematic model
  - clavicle (shoulder), torso (twist)
  - robust prior stabilizes complex joints
- Body part interpenetration
  - repulsive inter-part potentials
- Anatomical joint limits
  - hard bounds in parameter space



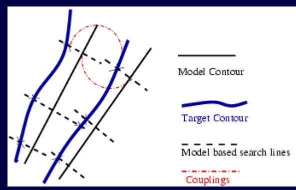
Multiple Image Features, Integrated Robustly



1. Intensity

- The model is `dressed' with the image texture under its projection (visible parts) in the previous time step
- Matching cost of model-projected texture against current image (robust intensity difference)

2.Contours



- Multiple probabilistic assignment integrates matching uncertainty
- Weighted towards motion discontinuities (robust flow outliers)
- Also accounts for higher order symmmetric model/data couplings
  - partially removes local, independent matching ambiguities

Cost Function Minima Caused By Incorrect Edge Assignments

Intensity + edges

Edges only

How many local minima are there?

Thousands ! – even *without* image matching ambiguities …

Tracking Approaches We Have Tried

- Traditional CONDENSATION
- Covariance Scaled Sampling
- Direct search for nearby minima
- Kinematic Jump Sampling

- 'Manual' initialization – already requires nontrivial optimization

Properties of Model-Image Matching Cost Function, 1

- High dimension
  - at least 30 – 35 d.o.f.
  - but factorial structure: limbs are quasi-independent
- Very ill-conditioned
  - depth d.o.f. often nearly unobservable
  - condition number O(1 : 10^4)
- Many many local minima
  - O(10^3) kinematic minima, times image ambiguity

Properties of Model-Image Matching Cost Function, 2

- Minima are usually well separated
  - fair random samples almost never jump between them
- But they often merge and separate
  - frequent passage through singular / critical configurations – frontoparallel limbs
  - causes mistracking!
- Minima are small, high-cost regions are large
  - random sampling with exaggerated noise almost never hits a minimum

Covariance Scaled Sampling, 1

Mistracking leaves us in the wrong minimum.

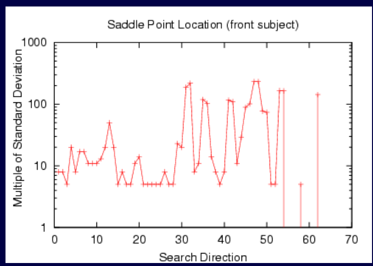
To make particle filter trackers work for this kind of cost function, we need :

- ***Broad sampling*** to reach basins of attraction of nearby minima
  - in CONDENSATION : exaggerate the dynamical noise
  - robust / long-tailed distributions are best
- Followed by ***local optimization*** to reach low-cost 'cores' of minima
  - core is small in high dim. problems, so samples rarely hit it
  - CONDENSATION style reweighting will kill them before they get there

Covariance Scaled Sampling, 2

- Sample distribution should be based on ***local shape of cost function***
  - the minima that cause confusion are much further in some directions than in others owing to ill-conditioning
  - in particular, kinematic flip pairs are aligned along ill-conditioned depth d.o.f.
- Combining these 3 properties gives ***Covariance Scaled Sampling***
  - long-tailed, covariance shaped sampling + optimization
  - represent sample distribution as robust mixture model

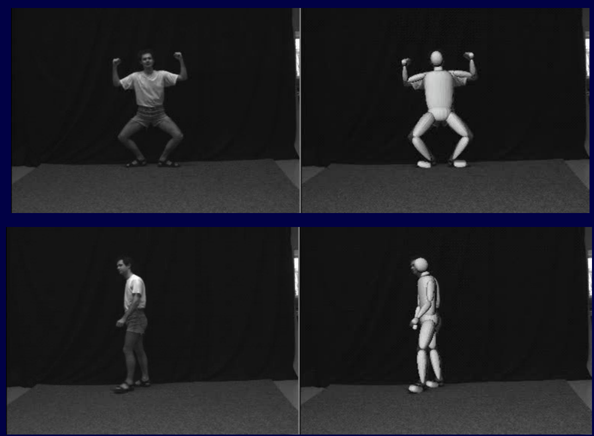
Statistical Separation of Minima



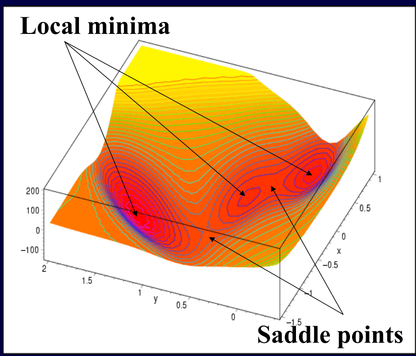
- Minima are usually at least $O(10^1)$ standard deviations away.

Direct Search for Nearby Minima

- Instead of sampling randomly, directly locate nearby cost basins by finding the 'mountain passes' that lead to them
  - i.e. find the saddle point at the top of the path
- Numerical methods for finding saddles :
  - modified Newton optimizers : eigenvector tracking, hypersurface sweeping
  - 'hyperdynamics' : MCMC sampling in a modified cost surface that focuses samples on saddles

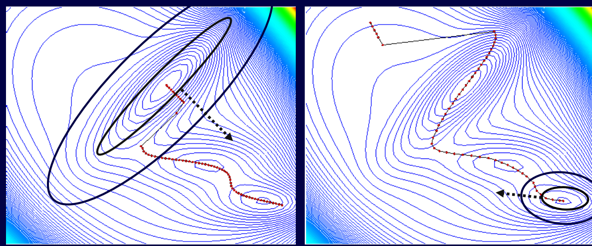


Direct Search for Nearby Minima

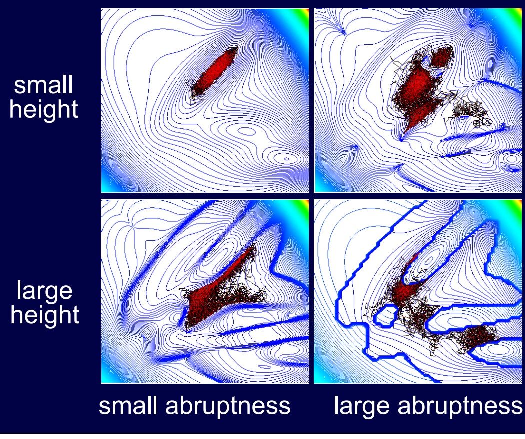


Hypersurface Sweeping

- Track cost minima on an expanding hypersurface
- Moving cost has a local maximum at a saddle point



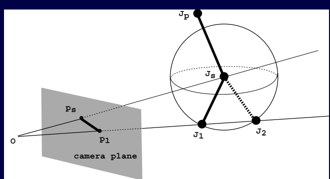
Hyperdynamics



Examples of Kinematic Ambiguities

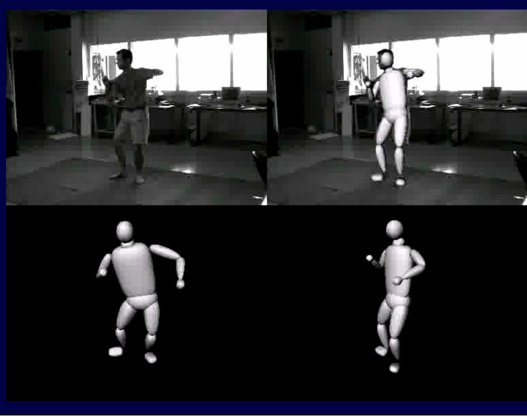
- Eigenvector tracking method
- Initialization cost function (hand specified image positions of joints)

Kinematic Jump Sampling



- Generate tree of all possible kinematic solutions
  - work outwards from root of kinematic tree, recursively evaluating forwards & backwards ‘flip’ for each body part
  - alternatively, sample by generating flips randomly
  - you can often treat each limb quasi-independently
- Yes, it really does find thousands of minima !
  - quite accurate too – no subsequent minimization is needed
  - random sampling is still needed to handle matching ambiguities

Jump Sampling in Action



Summary

- 3D articular human tracking from monocular video
- A hard problem owing to
  - complex model (many d.o.f., constraints, occlusions…)
  - ill-conditioning
  - many kinematic minima
  - model-image matching ambiguities
- Combine methods to overcome local minima
  - explicit kinematic jumps + sample for image ambiguities
- Current state of the art
  - relative depth accuracy is 10% or 10 cm at best
  - tracking for more than 5 – 10 seconds is still hard
  - still very slow – several minutes per frame

The End