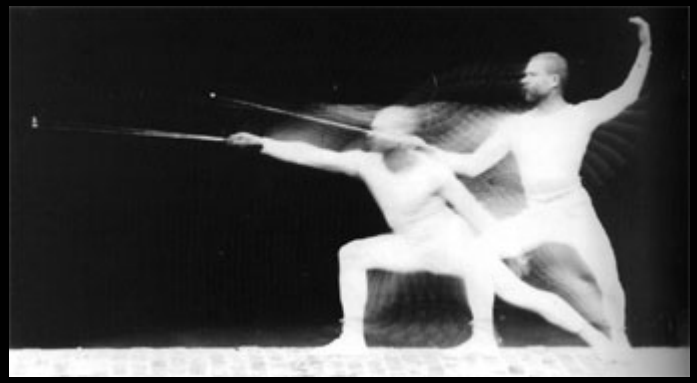
Optimization and Learning Algorithms for Visual Inference

3D Human Motion Reconstruction in Monocular Video

Cristian Sminchisescu



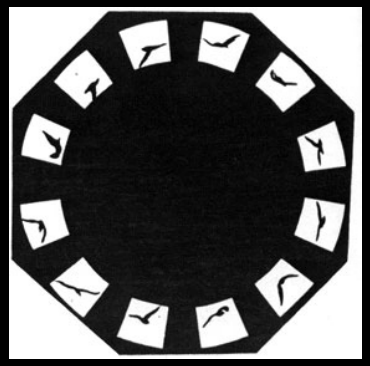
TTI-C / University of Toronto / Rutgers University
crismin@cs.toronto.edu
http://www.cs.toronto.edu/~crismin

Perception

- Inference from uncertain, ambiguous, incomplete and noisy data
- Combine measurement and prior knowledge

History of Analyzing Humans in Motion

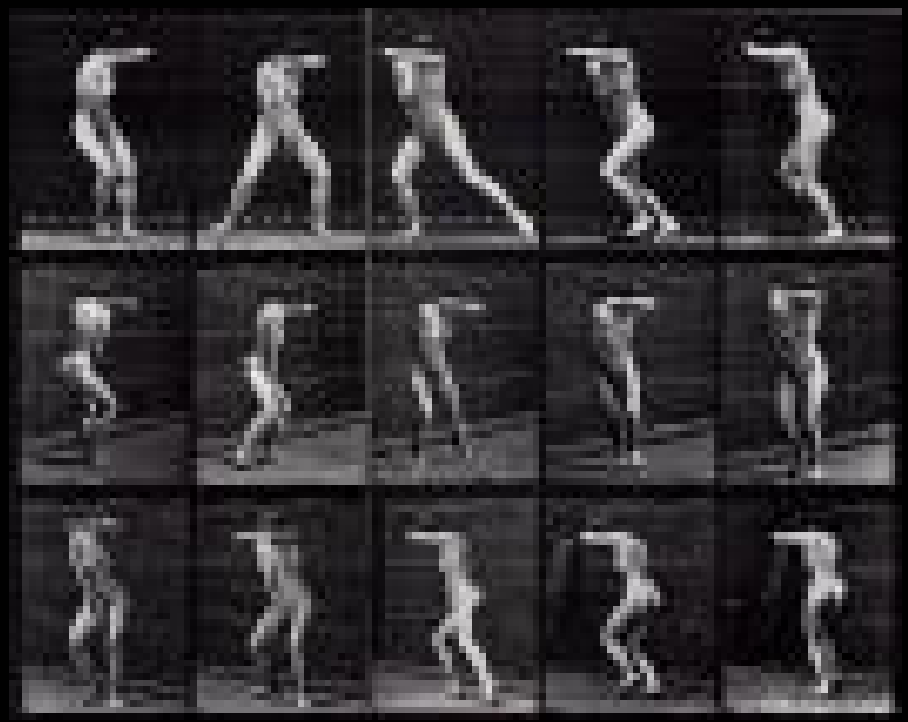
- Markers (*Etienne Jules Marey, 1882)*



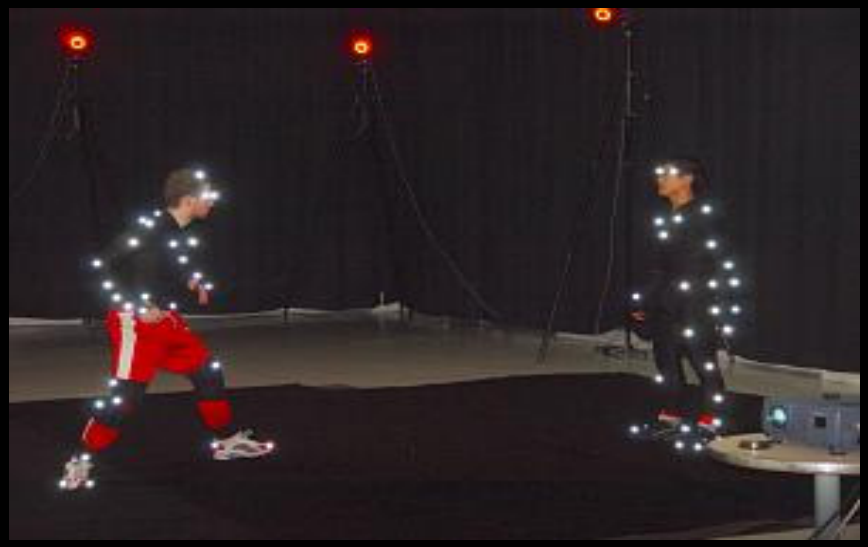
chronophotograph

- Multiple Cameras

(Eadweard Muybridge , 1884)



Human motion capture today
120 years and still fighting …



- VICON ~ 100,000 $
  - Excellent performance, *de-facto* standard for special effects, animation, *etc*
- But heavily instrumented
  - Multiple cameras
  - Markers in order to simplify the image correspondence
  - Special room, simple background

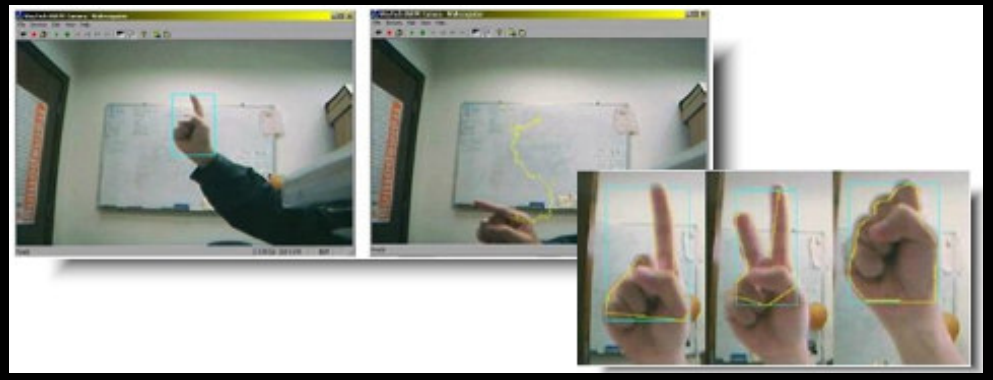
Major challenge: Move from the laboratory to the real world

Understanding people *« in vivo »*
Long-term problems to solve

- *Find* the people
- *Infer* their poses
- *Recognize* what they do
- *Recognize* what objects they use

Monocular motion capture: applications

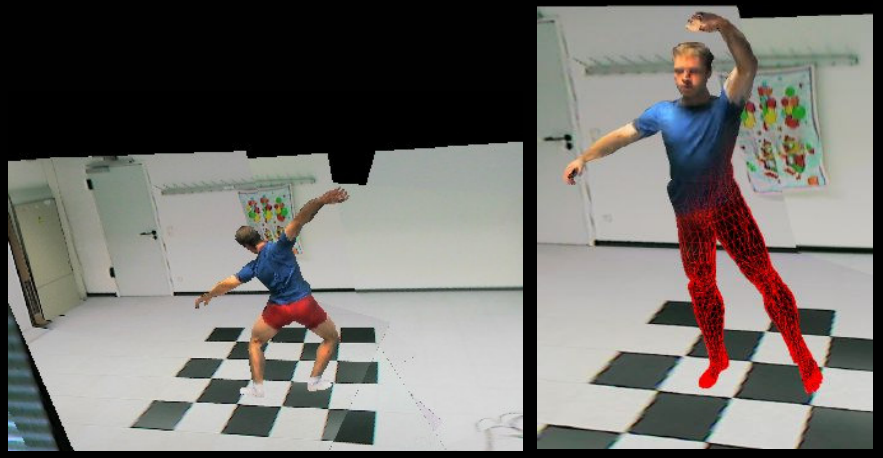
- Human-computer interfaces



- Video games, augmented reality

Monocular motion capture: applications

- Scene resynthesis, change of actor
- Photorealistic scene reconstruction



Seidel et al 04 (multiple cameras)

Monocular motion capture: applications

- Sports and rehabilitation medicine, elderly care
- Surveillance and protection systems

Last but not least

Great cognitive challenge: match the performance of the human eye

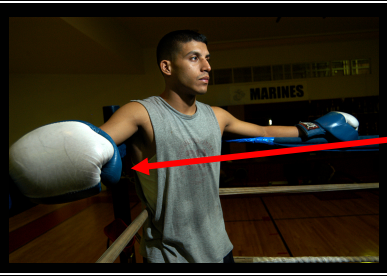
Monocular Motion Capture Difficulties

General poses

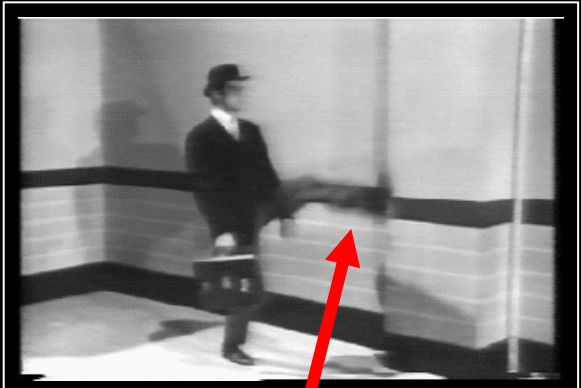
Self-occlusions

Difficult to segment the individual limbs

Different body sizes



Loss of 3D information in the monocular projection



Accidental allignments
Motion blur

Several people, occlusions

Reduced observability of body parts due to loose fitting clothing

Issues

- Model the complex appearance of humans
  - Integrate multiple image measurements (*e.g.* contours, intensity, silhouettes, *etc*) and learn their optimal weighting in order to reduce uncertainty
- Model the complex structural constraints of the body
  - Learn body proportions, as well as structured representations that encode the typical correlations between the motion of the body parts
- Represent and propagate uncertainty
  - Exploit the problem structure (*e.g.* locality, symmetries) during search
  - Integrate information over time in a compact and efficient way

The temporal recovery of the human pose is a problem of inference under uncertainty. An effective solution needs complex models and powerful inference and learning algorithms

Presentation Plan

- Introduction, history, applications
- State of the art for 2d and 3d, human detection, initialization
- 3D human modeling, generative and discriminative computations
- Generative Models
  - Parameterization, shape, constraints, priors
  - Observation likelihood and dynamics
  - Inference algorithms
  - Learning non-linear low-dimensional representations and parameters
- Conditional (discriminative) models
  - Probabilistic modeling of complex inverse mappings
  - Observation modeling
  - Discriminative density propagation
  - Inference in latent, kernel-induced non-linear state spaces
- Conclusions and perspectives

State of the Art (sparse sample)
Monocular Methods

- 2d global detectors
- 2d part-based methods
- 3d pose reconstruction and tracking

Potential focus of attention for 3d reconstruction algorithms

Constructing a Simple Person Detector

- Overcomplete filter bank + normalization + rectification
  - templates, wavelets, edge detectors
- Learn a decision boundary between people and non-people
  - Use either a generative model (Bayes classifier) or a good discriminator, e.g. SVM, RVM, AdaBoost
- Train on a large supervised training set (images + labels)
  - positives and random negatives
  - bootstrap by adding failures
- To detect, scan image at multiple positions, scales, orientations
  - Typically scale the image not the detection window

Support Vector Machine Detector

(Papagerogiu & Poggio, 1998)

Training set

descriptors

training

Support vector machine

test

Multi-scale search

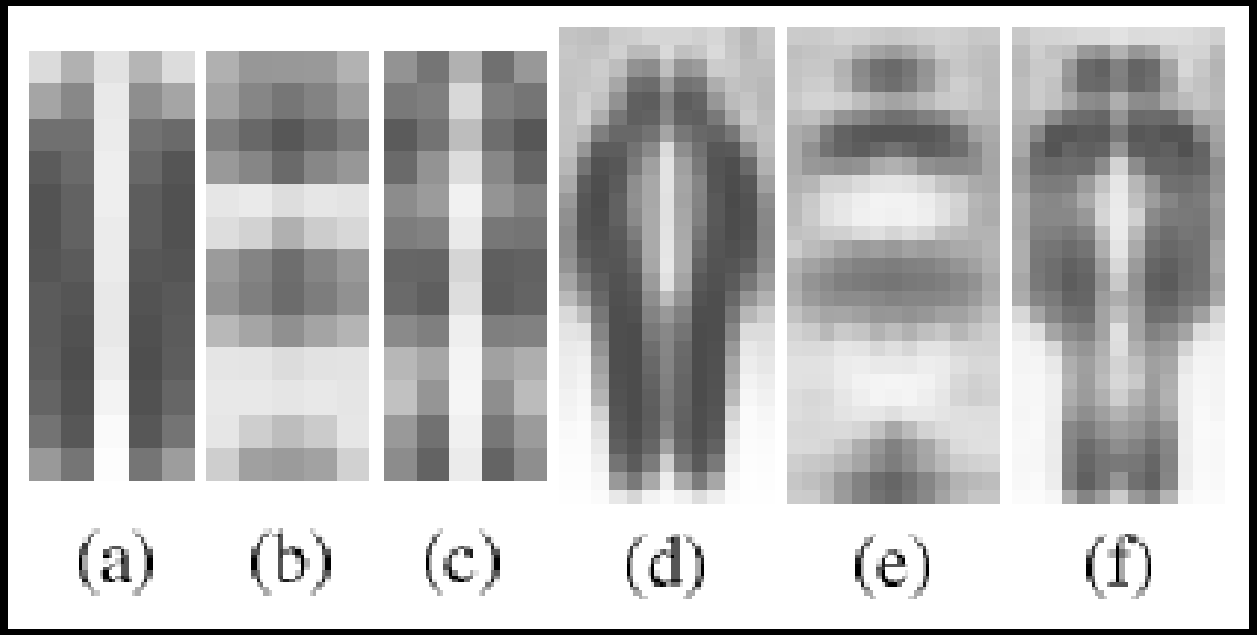
descriptors

Test image

results

Descriptor Relevance Determination

(Jean Goffinet, MSc. INPG 2001)



The machine learns a coarse silhouette template

Dynamic Pedestrian Detection

Viola, Jones and Snow, ICCV 2003

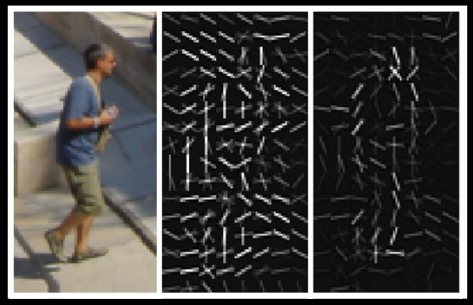
- Train using AdaBoost, about 45,000 possible features
- Efficient and reliable for distant detections (20x15), 4fps

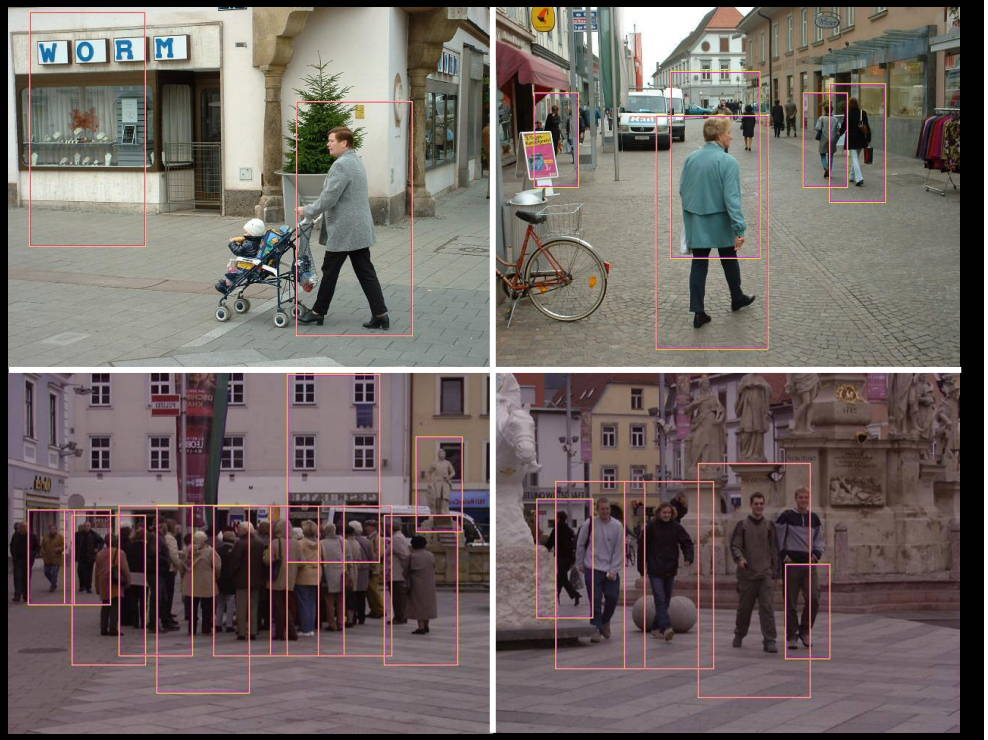
2d Global Detector

Dalal and Triggs, CVPR 2005

- 3-D Histogram of Oriented Gradients (HOG) as descriptors
- Linear SVM for runtime efficiency
- Tolerates different poses, clothing, lighting and background
- Currently works for fully visible upright persons

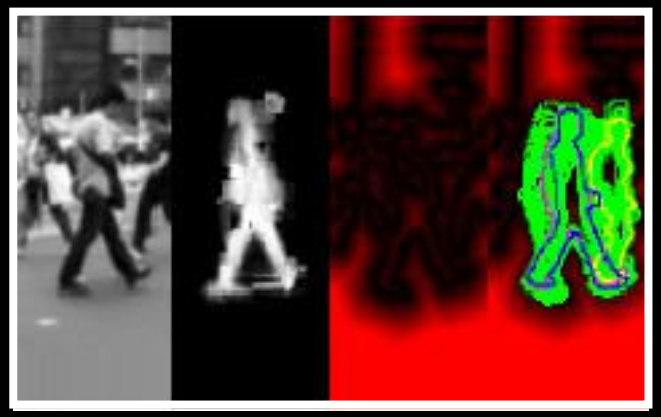
Importance weight responses

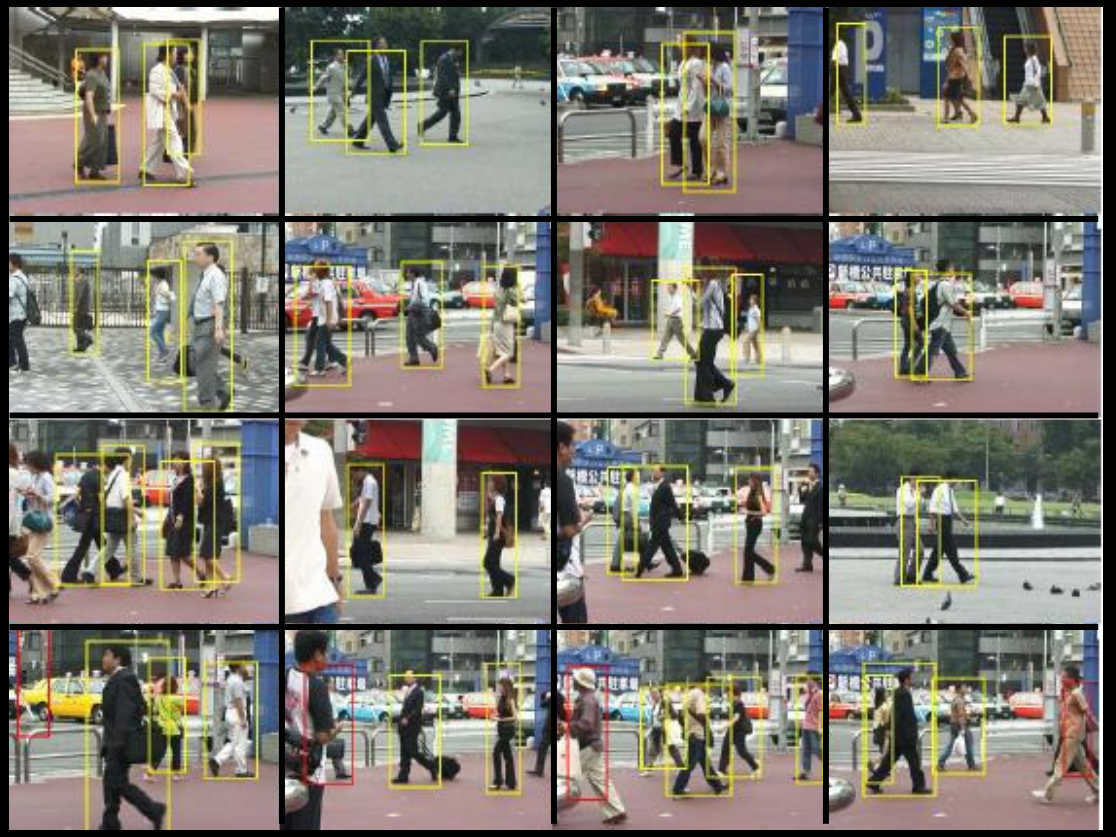




2D Global Detector

Leibe and Schiele, CVPR 2005





- System combining local and global cues based on probabilistic top-down segmentation
- Improved behavior during partial occlusions

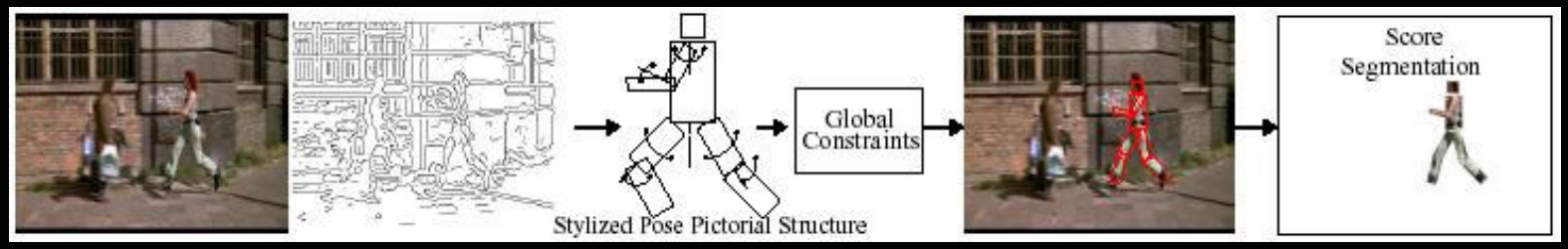
Detecting 2d Articulated Structures

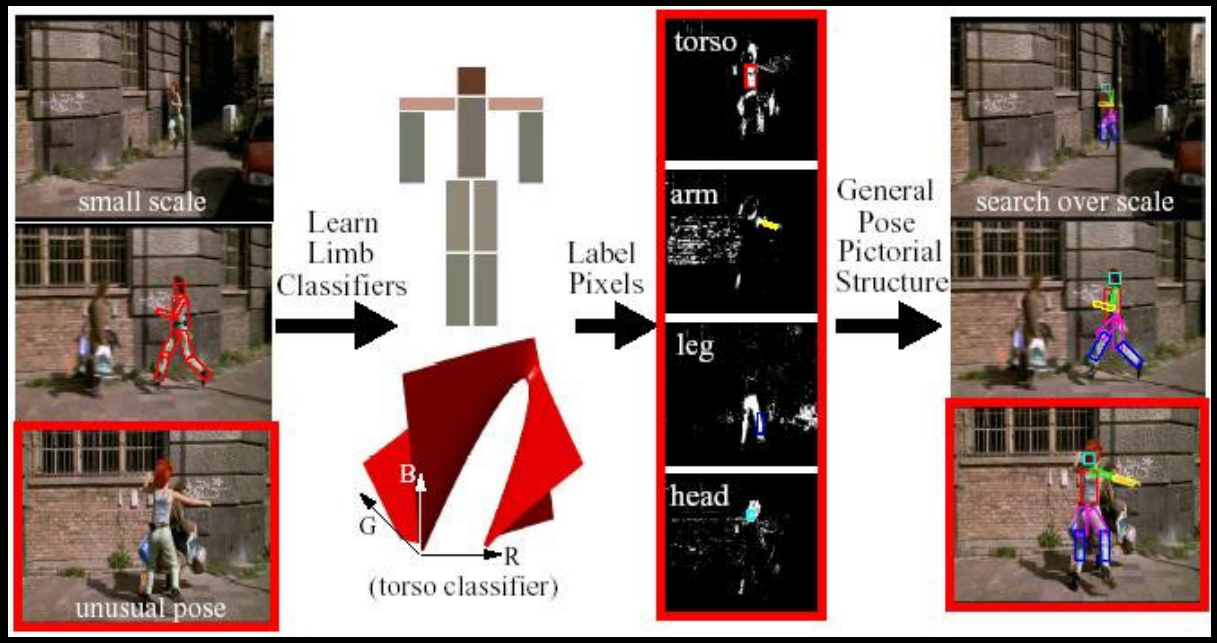
- Global detectors are useful for focusing attention
- Effective for pedestrians or distant detections
- May not scale well for more complex human poses
  - Hard to sample the high-d space of possible articulations
  - Need prohibitively large amounts of data, slow, *etc*

Instead, detect simple parts (faces, limbs) and glue them using human body assembly rules and consistency constraints

2d Part-Based Model Detection

Ramanan and Forsyth, CVPR 2005

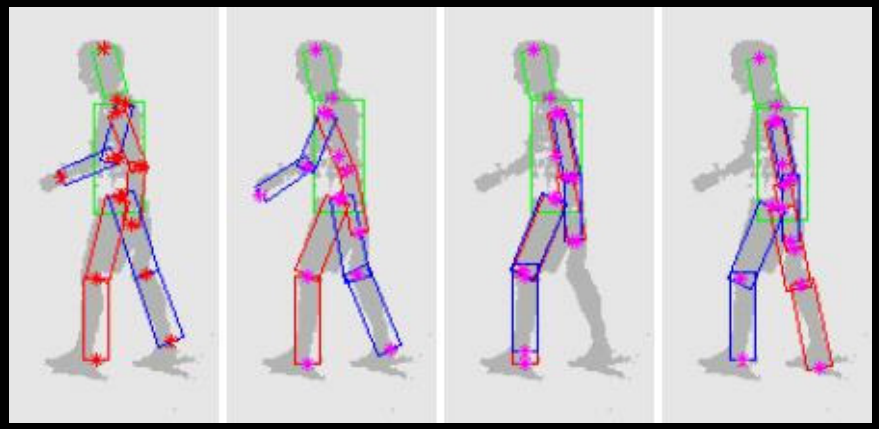




- Find a typical pose using a generic pictorial model (having dependence tree structure) with parts identified using a ribbon detector
  - Tree prior *cf. Chow & Liu IEEE Tran. Info.Theory 1968; Different models in: Meila ICML1999, Felzenszwalb & Huttenlocher CVPR 2000, Ioffte and Forsyth ICCV2001*
- Acquire appearance model and use it to detect unusual poses

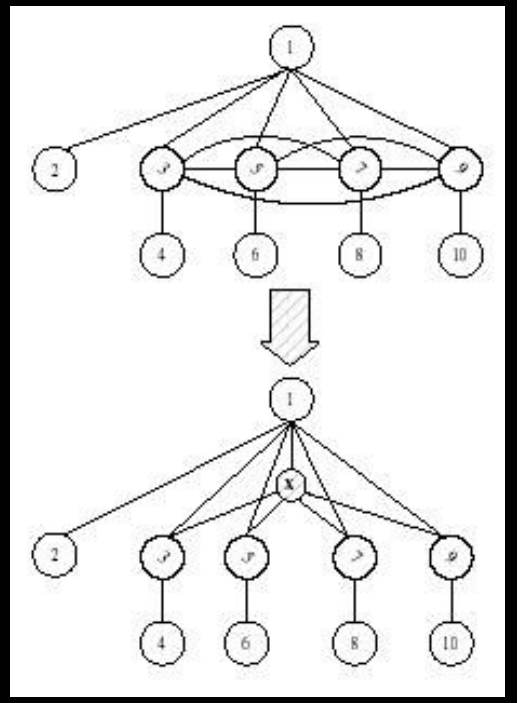
2d Part-Based Detection Common-Factor Models

Lan and Huttenlocher, ICCV 2005



Ground truth CFM PS LBP

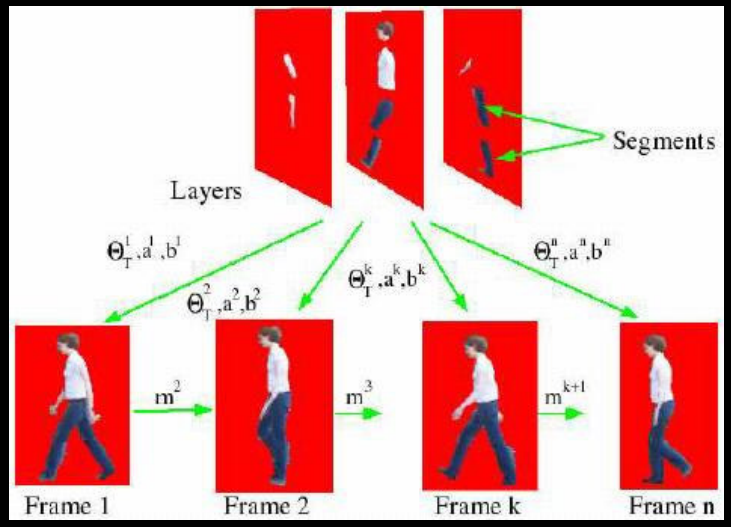
Break large cliques

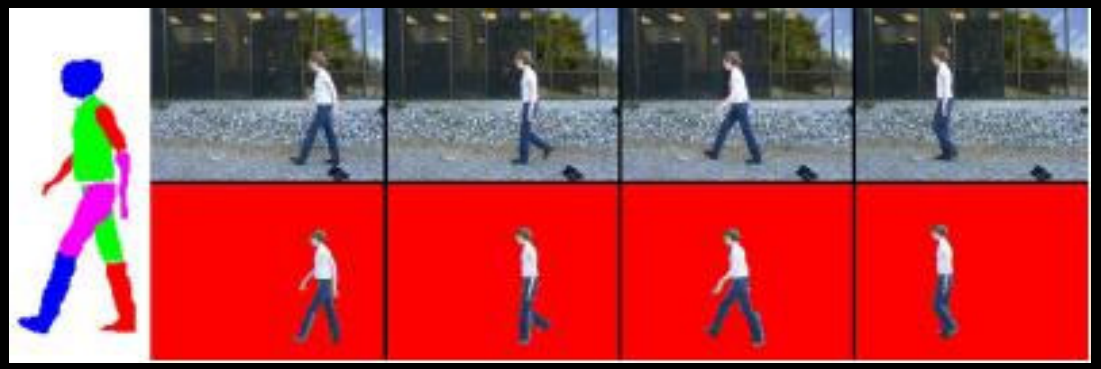


- Models correlations between limbs, enforces consistency
- Replaces a tree prior with a tractable hidden-variable one, based on Factor Analysis (see also a different model in *Frey et al, CVPR 2003*)
- Avoids expensive inference due to large cliques

Layered Motion Segmentations

Kumar, Torr and Zisserman, ICCV 2005

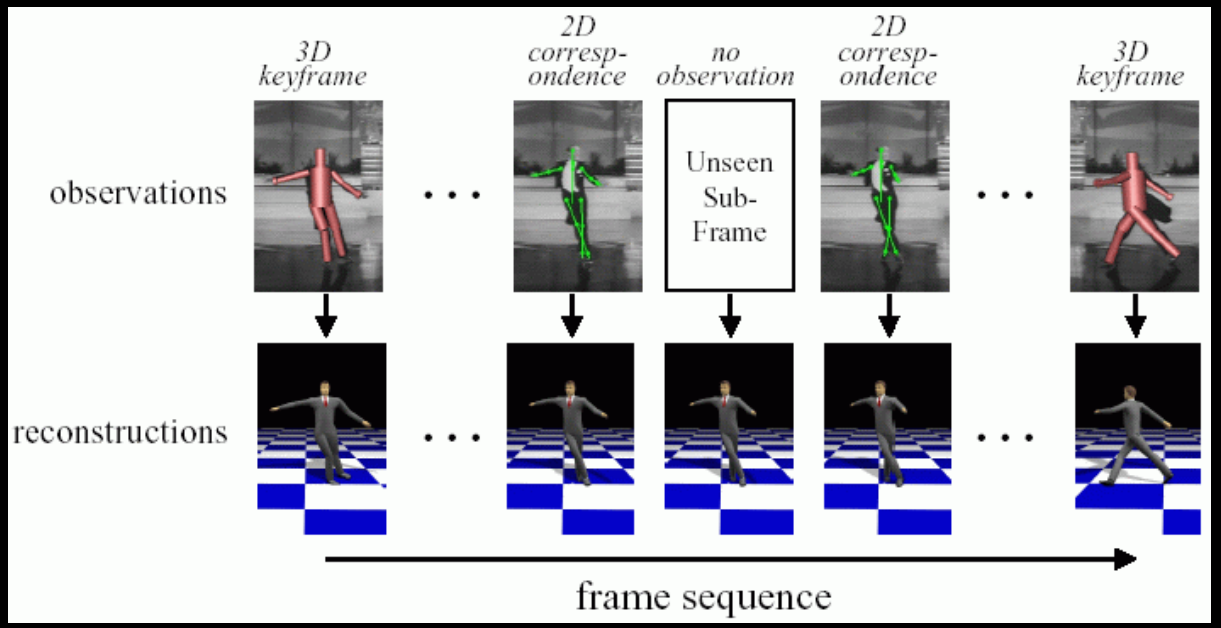




- Models image projection, lighting and motion blur
- Models spatial continuity, occlusions, and works over multiple frames (see a different model in *Jojic & Frey, CVPR 2001*)
- Estimates the number of segments, their mattes, layer assignment, appearance, lighting and transformation parameters for each segment
- Initialization using loopy BP, refinement using graph cuts

2d Kinematic Tracking
The Scaled Prismatic Model

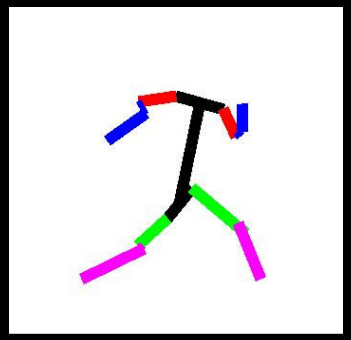
Cham and Rehg, CVPR 1999



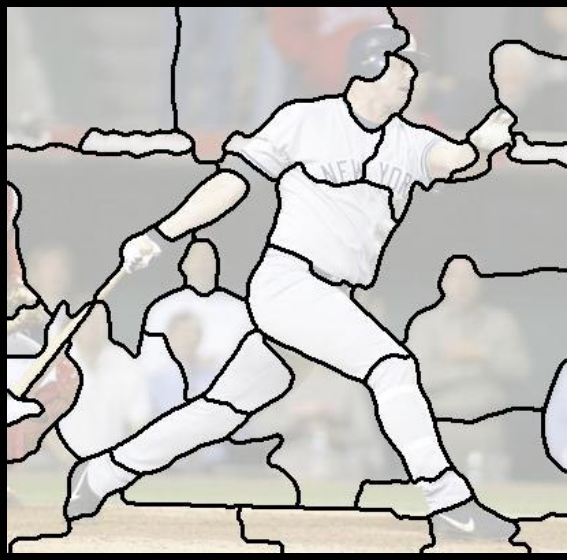
- Avoids 3d singularities and non-observabilities
  - but less intuitive reasoning about physical constraints
- Multiple Hypothesis Method: random sampling + local optimization
- Framework also used for interactive reconstruction based on user-provided 3d keyframe poses (*DiFranco, Cham and Rehg, CVPR 2001*)

2d Part-Based Detection

Mori, Ren, Efros and Malik, CVPR 2004



SEGMENTS



SUPERPIXELS

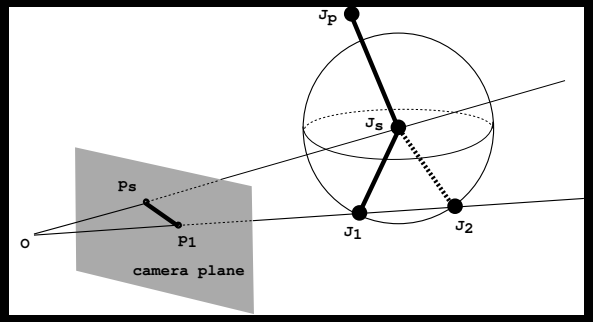
- Window-scanning (*e.g.* face detection)
- Bottom-up, detect half-limbs and torsos
- Top-down, assemble parts into human figure.
  - These can be matched to exemplars with known 2d joint positions using shape contexts;
  - 3d can be extracted interactively (*cf. Lee & Chen'85, Taylor'00, Mori & Malik'02*)

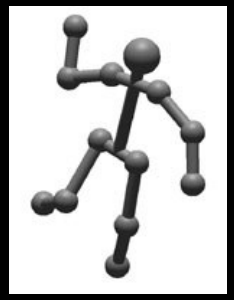
3D

3d from 2d Joint Positions

*Lee and Chen, CVGIP **1985***

- Characterizes the space of solutions, assuming
  - 2d joint positions + limb length
  - internal camera parameters
- Builds an interpretation tree of projection-consistent hypotheses (3d joint positions)
  - obtained by forward-backward flips in-depth
  - $O(2^{\#\text{ of body parts}})$ solutions
  - In principle, can prune some by physical reasoning
  - But no procedure to compute joint angles, hence difficult to reason about physical constraints
- Not an automatic 3d reconstruction method
  - select the true solution (out of many) manually
- Adapted for orthographic cameras *(Taylor 2000)*





Taylor, CVIU 2000

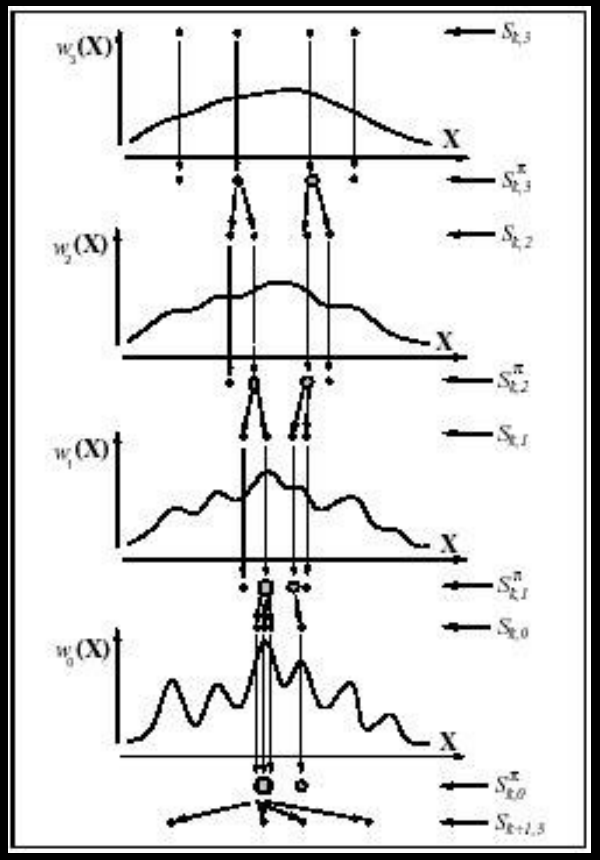
Generative 3D Reconstruction
Annealed Particle Filter

(Deutscher, Blake and Reid, CVPR 2000)

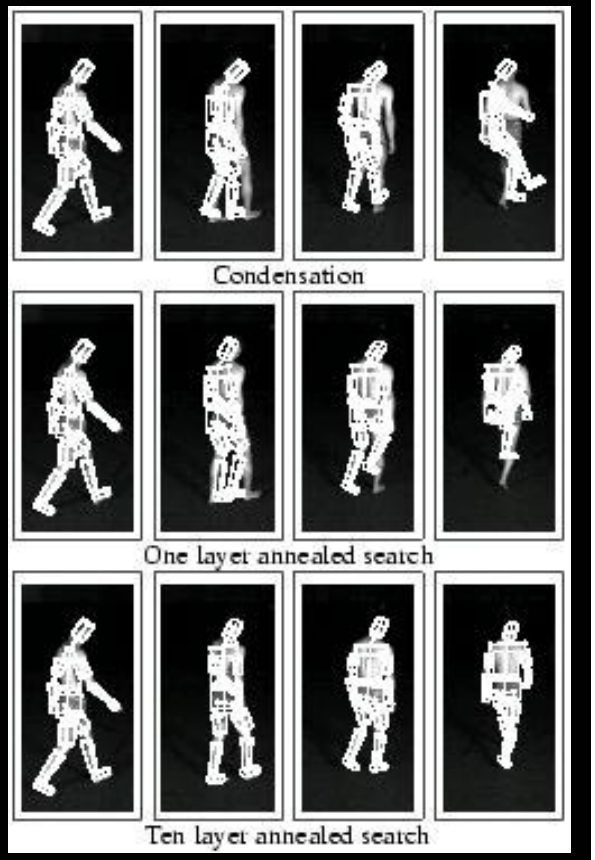
Careful design

- Dynamics
- Observation likelihood
  - edge + silhouettes
- Annealing-based search procedure, improves over particle filtering

- Simple background and clothing



monocular

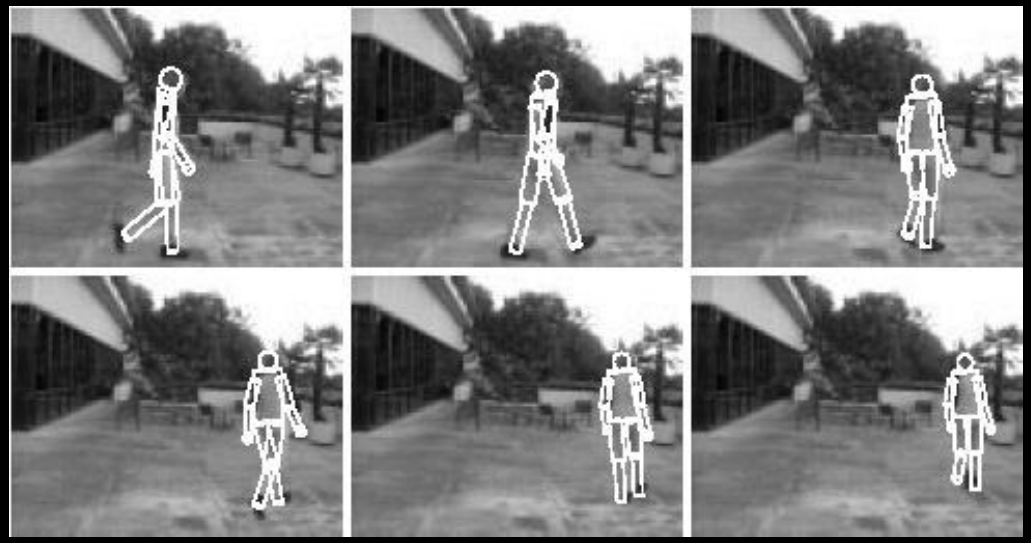


Improved results (complex motions) when multiple cameras (3-6) were used

Generative 3D Reconstruction

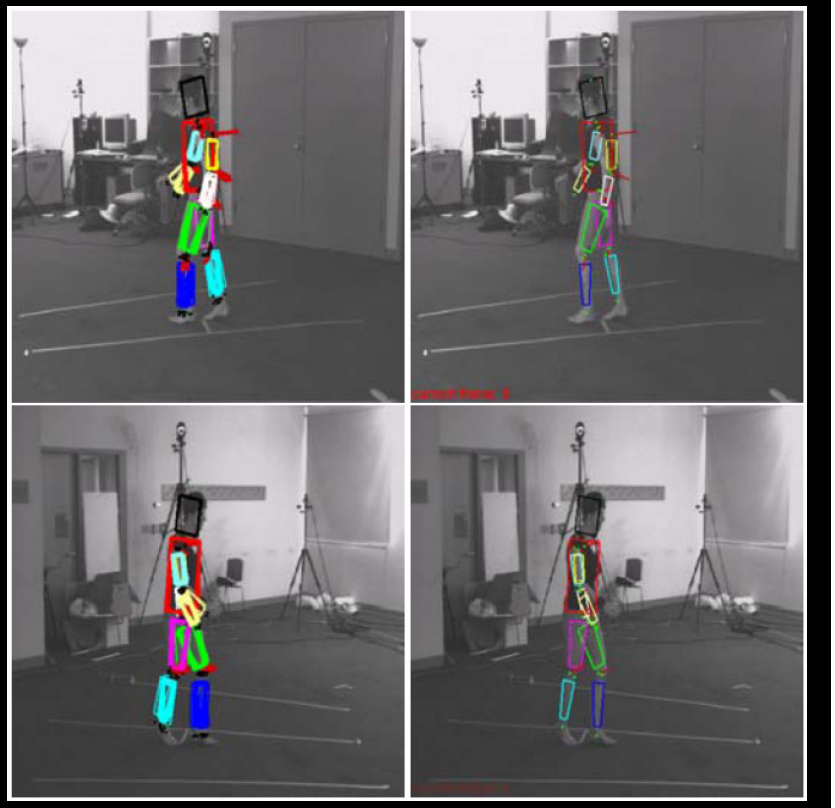
Sidenbladh, Black and Fleet, ECCV 2000; Sidenbladh & Black, ICCV 2001, Sienbladh, Black and Sigal, ECCV 2002; Sigal et al, CVPR 2004

Monocular



- Condensation-based filter
- Dynamical models
  - walking, snippets
- Careful learning of observation likelihood distributions *(more later)*

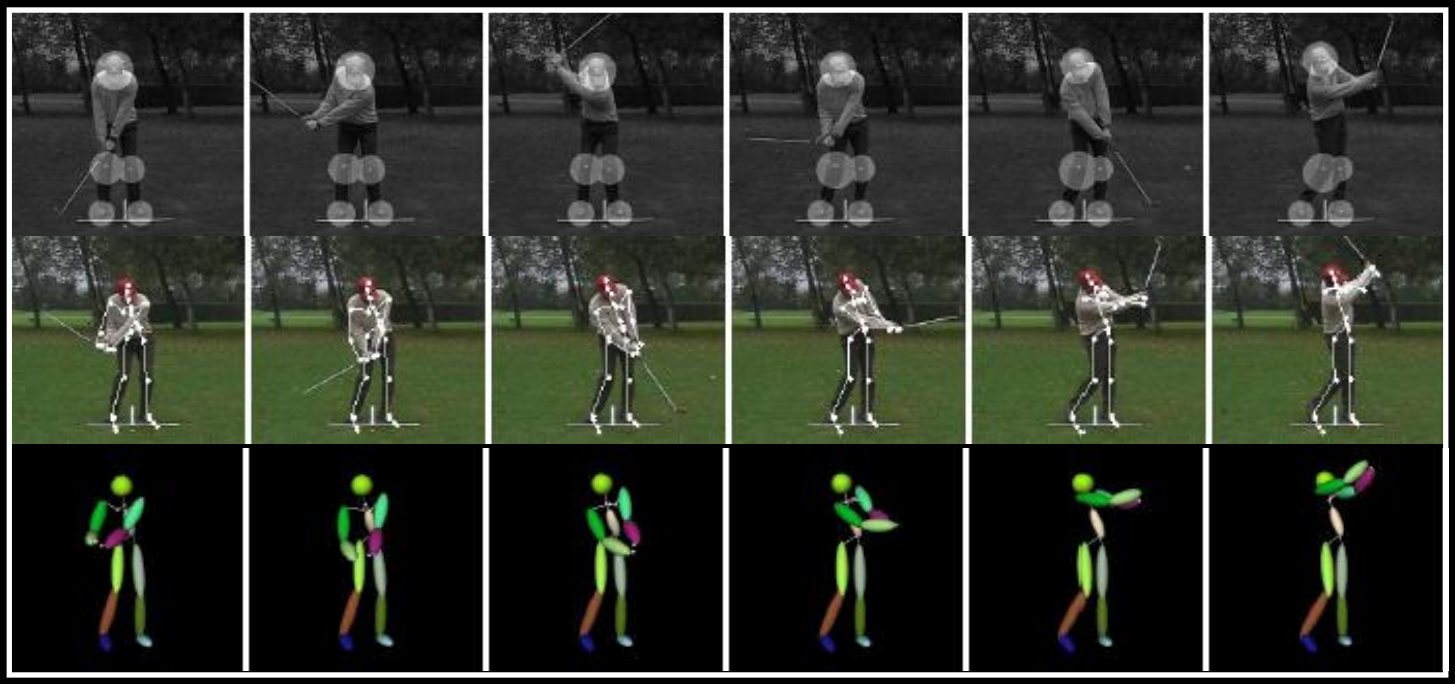
Multi-camera



- Non-parametric belief propagation, initialization by limb detection and triangulation

3D Model-Based Reconstruction

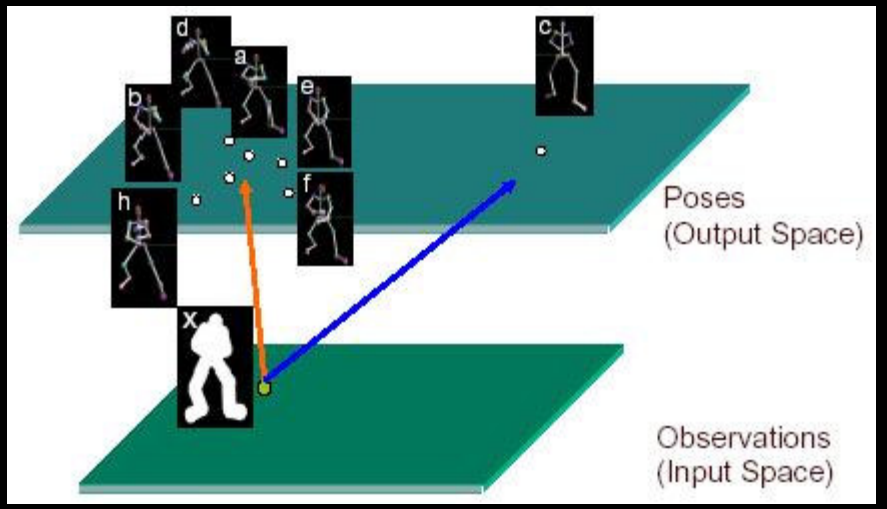
(Urtasun, Fleet, Hertzmann and Fua, ICCV 2005)

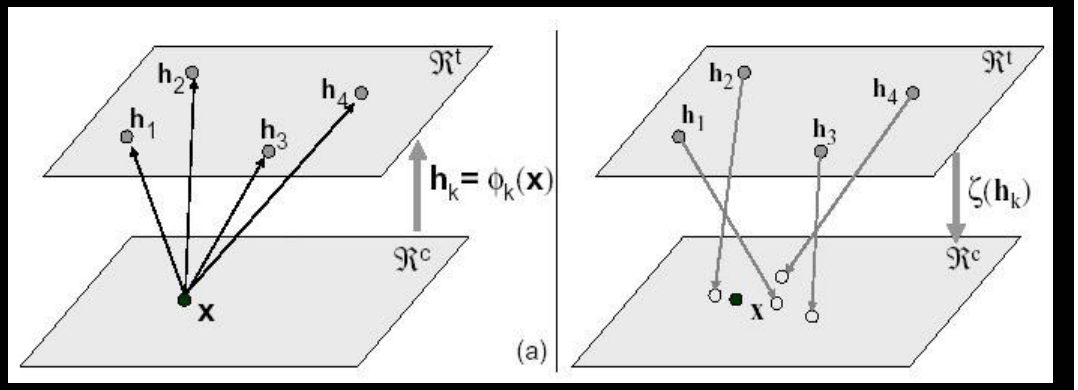


- Track human joints using the *WSL* tracker (*Jepson et al'01*)
- Optimize model joint re-projection error in a low-dimensional space obtained using probabilistic PCA *(Lawrence'04)*
  - *Effective low-dimensional model, but local geometry not necessarily preserved (see Lawrence & Candela's ongoing work on using backconstraints; see also Sminchisescu & Jepson, ICML 2004, later in the talk)*

Discriminative 3d:
Specialized Mappings Architecture
Rosales and Sclaroff, ICCV 2001

- Static 3D human pose estimation from silhouettes (Hu moments)
- Approximates the observation-pose mapping from training data
  - Mixture of neural network predictors
  - Models the joint distribution
- Uses the forward model (graphics rendering) to verify solutions

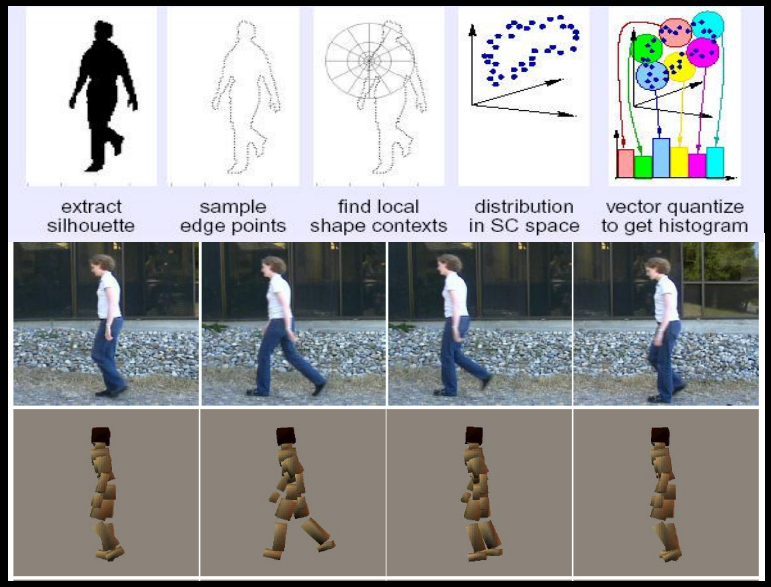


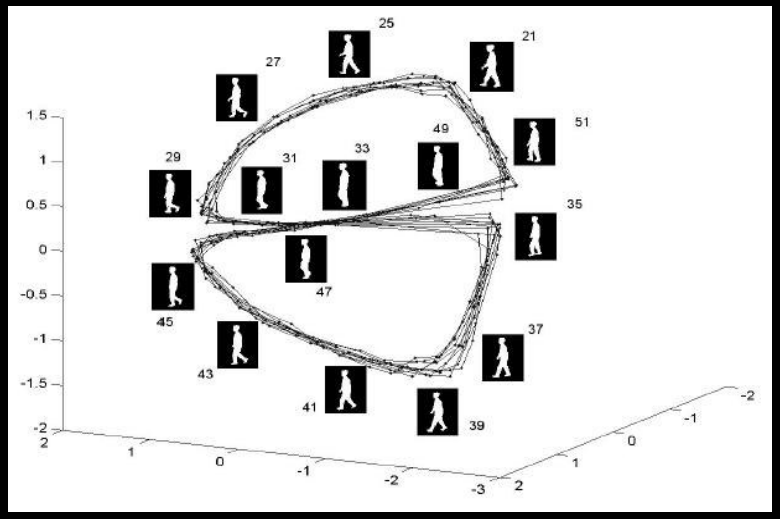


Discriminative 3d: Regression Methods

Aggarwal and Triggs, CVPR 2004, Elgammal & Lee, CVPR 2004

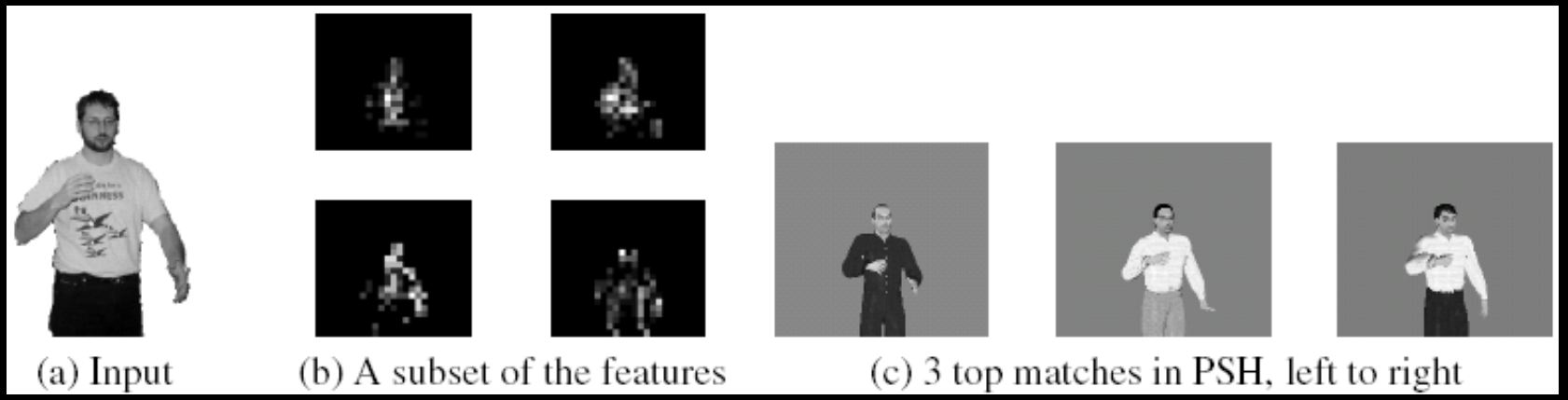
- *(A&T)* 3d pose recovery by non-linear regression against silhouette observations represented as shape context histograms
  - Emphasis on sparse, efficient predictions, good generalization
- *(A&T)* Careful study of dynamical regression-based predictors for walking (ICML'04) and extensions to mixture of regressors (HCI'05)
- *(E&L)* pose from silhouette regression where the dimensionality of the input is reduced using non-linear embedding.
  - Latent (input) to joint angle (output) state space map based on RBF networks





Discriminative 3d: Nearest Neighbor Parameter Sensitive Hashing (PSH)

Shakhnarovich, Viola and Darell, ICCV 2003



- Relies on database of (observation, state) pairs rendered artificially
  - Locates samples that have observation components similar to the current image data (nearest neighbors) and use their state as putative estimates
- Extension to multiple cameras and tracking by non-linear model optimization (PSH used for initialization *Demirdjan et al, ICCV05*)
  - Foreground / background segmentation from stereo

Overview Monocular 3D Research

- **Generative**
  - Priors on short motion pieces (snippets): *Howe, Leventon & Freeman'99*
  - HMM, silhouette sequence, Viterbi inference: *Brand'99*
  - Annealing: *Deutscher, Blake & Reid '00*
  - Non-linear optimization: *Wachter & Nagel '99, Urtasun et al'04'05*
  - Non-linear optimization, Laplace approximations, MCMC: *Choo&Fleet'01*, *Sminchisescu&Tiggs'01-'03; Sminchisescu & Jepson'04*
  - Importance sampling, NBP: *Sidenbladh, Black, Fleet'00; Seidenbladh&Black'01; Sidenbladh et al.'02; Sudderth et al'04; Lee&Cohen'04*

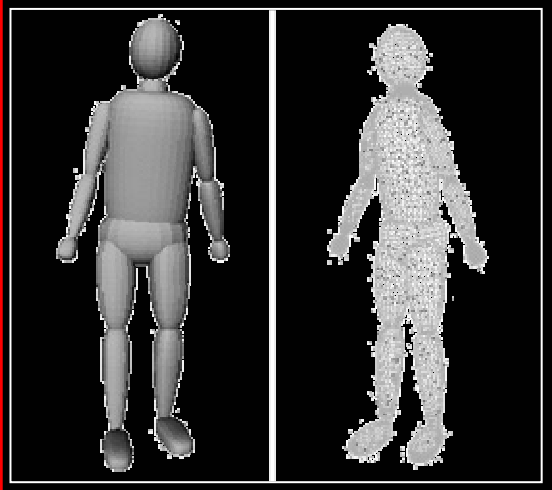
- **Discriminative**
  - Static silhouette features, specialized mappings: *Rosales & Sclaroff'00*
  - Approximate nearest neighbor search: *Shahnarovich, Viola & Darell '03*
  - Semi-automatic: *Lee&Chen'85, Taylor'00, Mori et al'03'04*
  - Classification + interpolation: *Tomasi et al '03*
  - Regression, mixtures: *Aggarwal & Triggs ' 04 '05, Elgammal et al.'04*
  - Discriminative density propagation: *Sminchisescu, Kanaujia, Li, Metaxas'05*
- **Geometric** *Tresadern & Reid'05, Yan & Pollefeys'05, Sinclair et al'97*

Presentation Plan

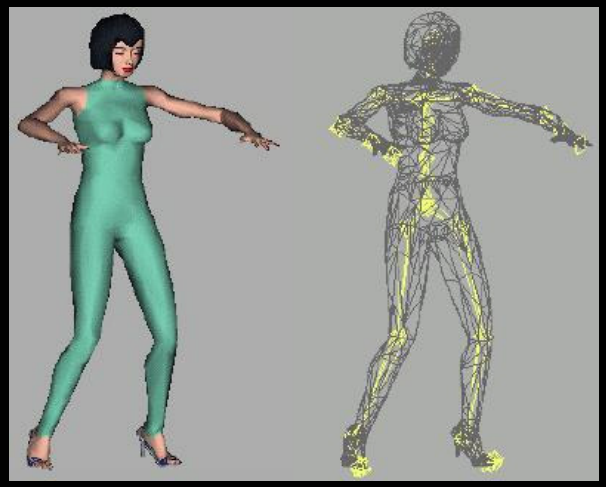
- Introduction, history, applications
- State of the art for 2d and 3d, human detection, initialization
- 3D human modeling, generative and discriminative computations
- Generative Models
  - Parameterization, shape, constraints, priors
  - Observation likelihood and dynamics
  - Inference algorithms
  - Learning non-linear low-dimensional representations and parameters
- Conditional (discriminative) models
  - Probabilistic modeling of complex inverse mappings
  - Observation modeling
  - Discriminative density propagation
  - Inference in latent, kernel-induced non-linear state spaces
- Conclusions and perspectives

Levels of 3d Modeling

This talk



- Coarse body model
- 30 - 35 d.o.f
- Simple appearance (implicit texture map)

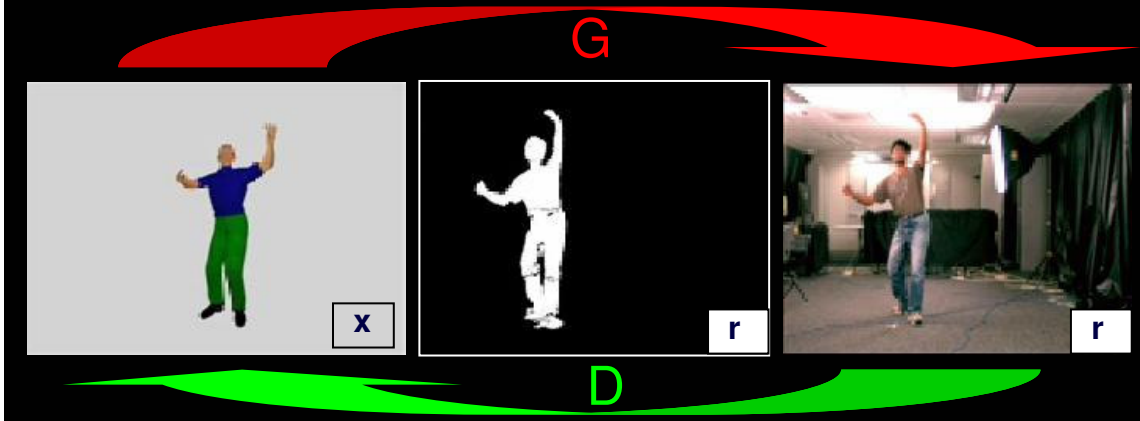


- Complex body model
- 50 - 60 d.o.f
- Simple appearance
- (edge histograms)

Photo Synthetic

- Complex body model
- ? (hundreds) d.o.f
- Sophisticated modeling of clothing and lighting

Generative vs. Discriminative Modelling



x is the model state

r are image observations

Goal: $p_\theta(\mathbf{x}|\mathbf{r})$

θ are parameters to learn given training set of (**r**,**x**) pairs

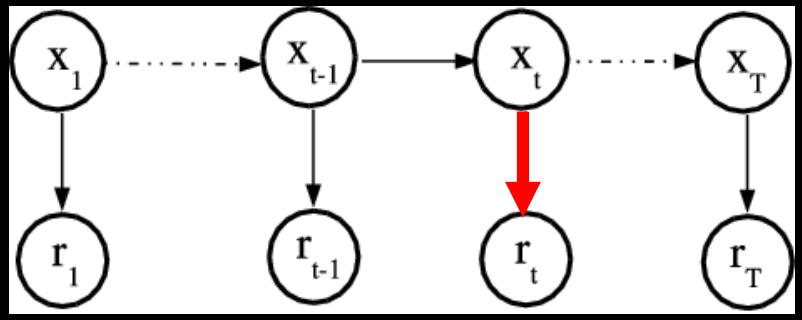
$$p_\theta(\mathbf{x}|\mathbf{r}) \;\alpha\; p_\theta(\mathbf{r}|\mathbf{x}) \cdot p(\mathbf{x})$$

- Learning to `invert' perspective projection and kinematics is difficult and produces multiple solutions
  - *Multivalued mappings* ≡ *multimodal conditional state distributions*
- Probabilistic temporal framework lacking until now
  - What distributions to model?
  - Which propagation rules to use?

- Learn
  - State representations and priors
  - Observation likelihood; but difficult to model human appearance
  - Temporal dynamics
- Sound probabilistic framework
  - Mixture or particle filters
  - State inference is expensive, need effective optimization

Chains for Temporal Inference

- Generative (top-down), Kalman filtering, CONDENSATION

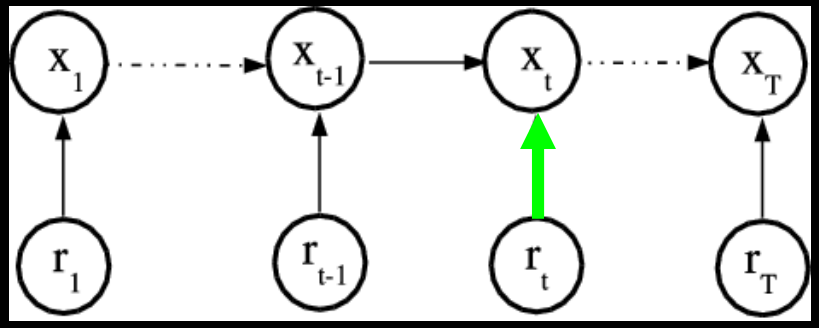


← *Models* the observation

$$p(\mathbf{x}_t|\mathbf{R}_t) \propto \frac{p(\mathbf{x}_t|\mathbf{r}_t)}{p(\mathbf{x}_t)} \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$$

$$p(\mathbf{x}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1})$$

- Discriminative (bottom-up) (*Sminchisescu et al, CVPR 2005*)

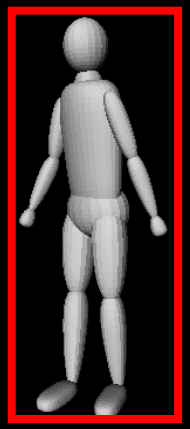


← *Conditions on* the observation

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$$

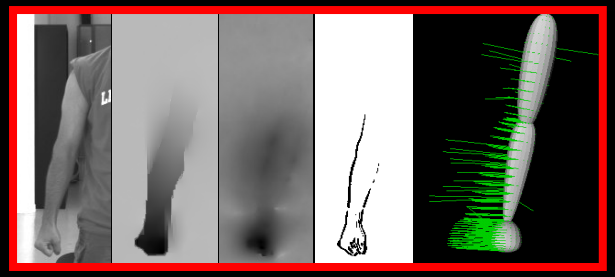
Generative Modeling

1

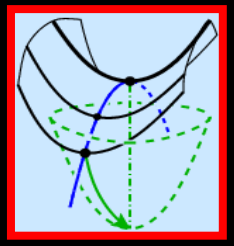


$$p(\mathrm{x}^{\mathrm{H}} \mid \mathrm{o}) \propto p(\mathrm{o} \mid \mathrm{x}^{\mathrm{H}}) \cdot p_H(\mathrm{x}^{\mathrm{H}})$$

2



3



1. Generative Human Model
   - Complex, kinematics, geometry, photometry, physical priors
   - Predicts images or descriptors
2. Observation likelihood function (matching cost)
   - Associates model predictions to image features
   - Robust, probabilistically motivated
3. Static state estimation and tracking by probabilistic inference / optimization
   - Discovers well supported configurations of matching cost

Human Body Model

Explicit 3D model allows high-level interpretation

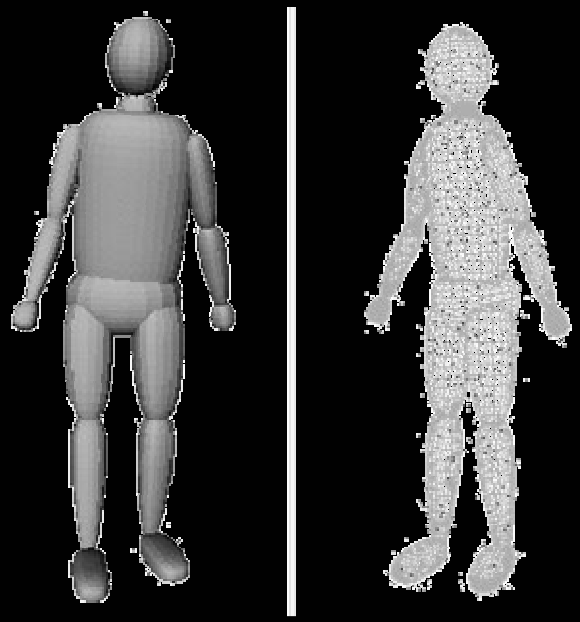
- 35 d.o.f. articular ‘skeleton’ x^H

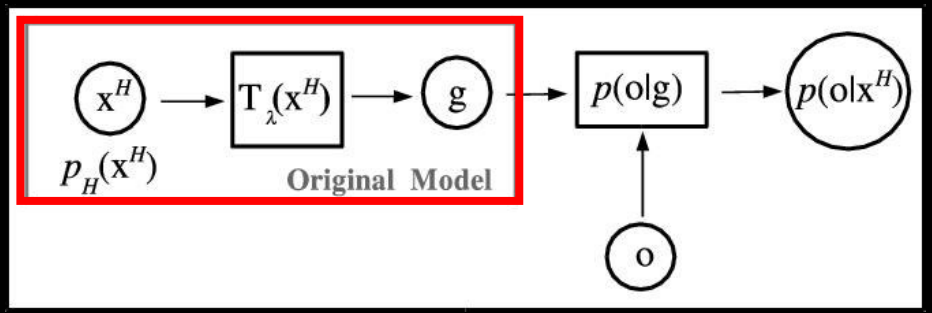
- ‘Flesh’ of superquadric ellipsoids λ
  - tapering & bending deformations

- Points on `skin’ mapped through
  - Kinematic chain
  - Camera matrix
  - Occlusion

  $\mathrm{T}_\lambda(\mathrm{x}^H)$

- Priors $p_H(\mathrm{x}^H)$





State Space Priors

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

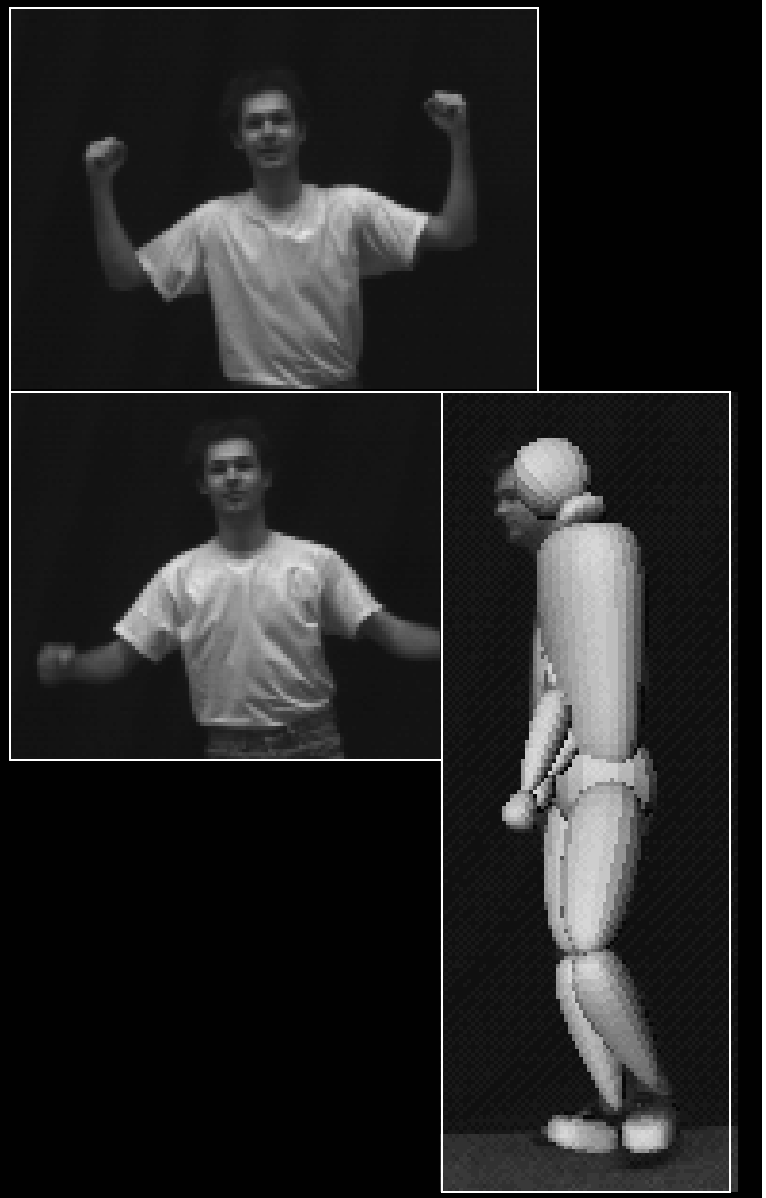
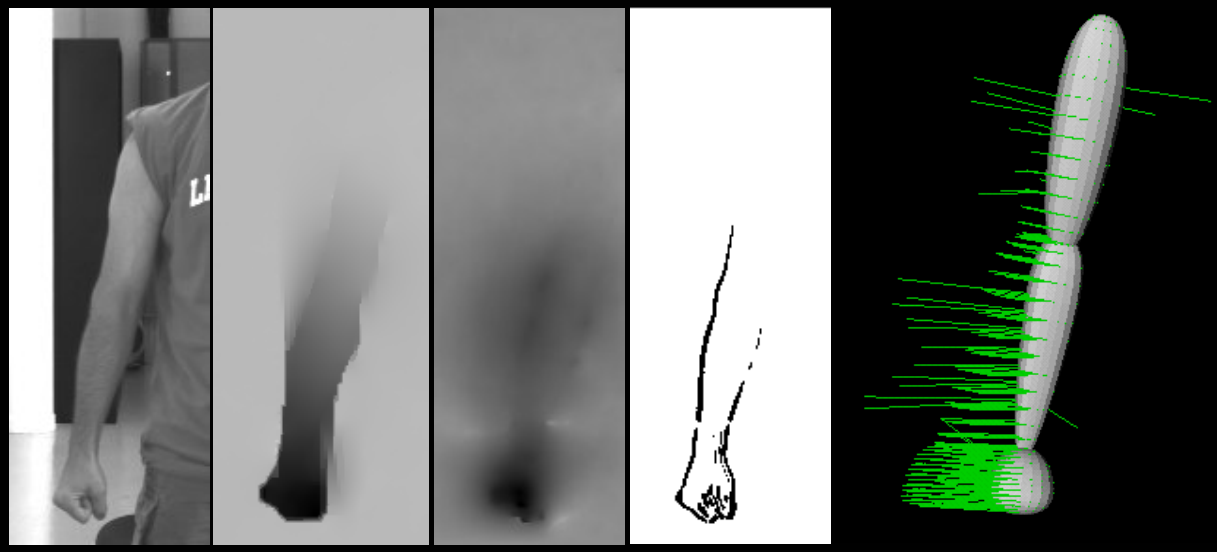


Image Features, Integrated Robustly

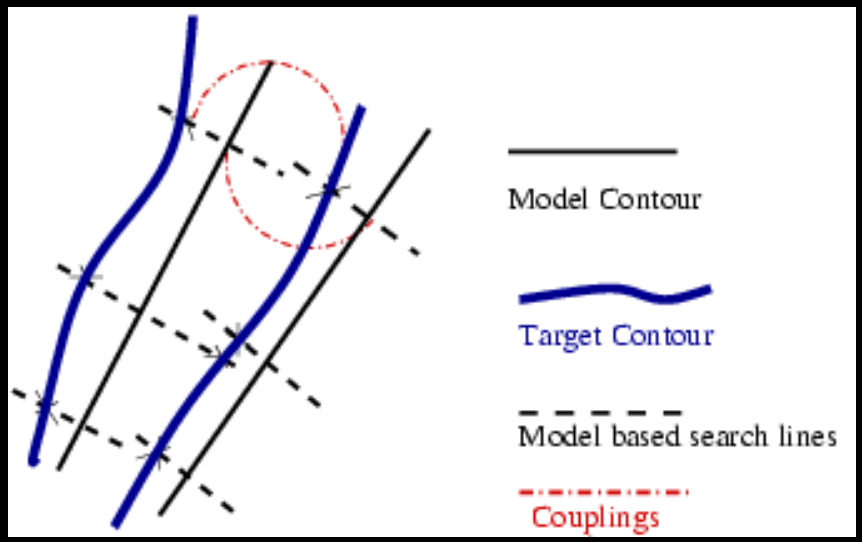
These form the observation likelihood, a continuous function: a weighted combination of several types of image observations, collected at predicted model elements



1. Intensity

- Model `dressed' with the image texture under its projection (visible parts) in the previous time step & hypothesis
- Measure 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)
- Accounts for symmmetries in the model (non-independence)
  - partially removes ambiguities resulting from independent, local model-data matching

3.Silhouettes

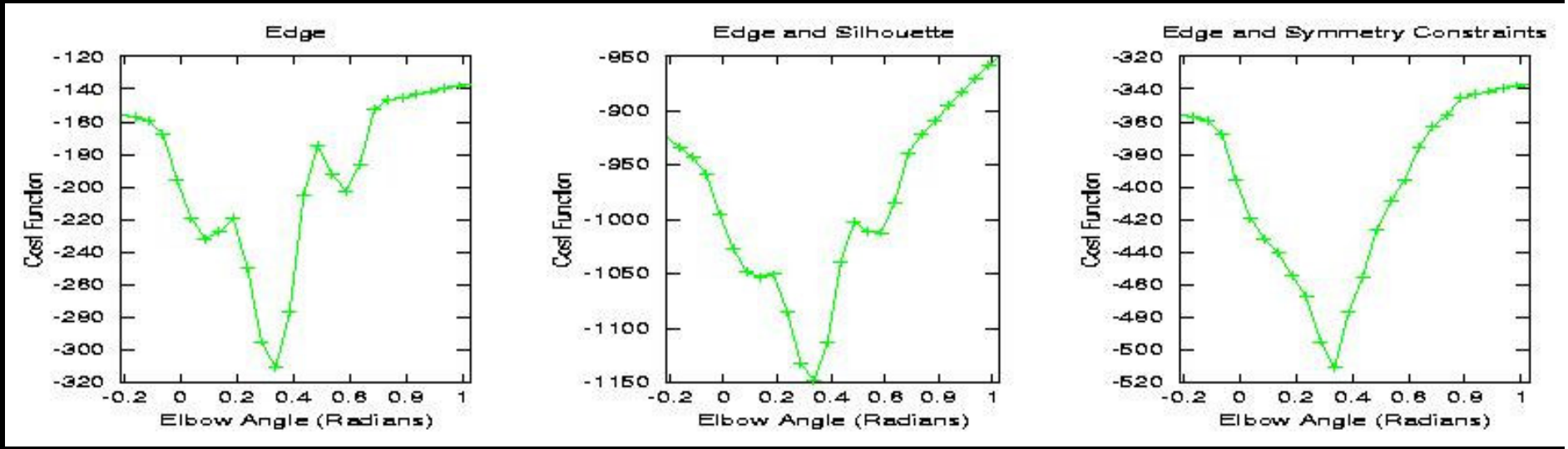
Use an attraction-explanation pair of cost terms

inconsistent *consistent*

- Push the model inside image silhouette
  - use distance level functions
- To avoid inconsistency, demand the model explains the image
  - maximize the model / image silhouette overlap

The Importance of Correct Modeling

Sminchisescu, F&G 2001



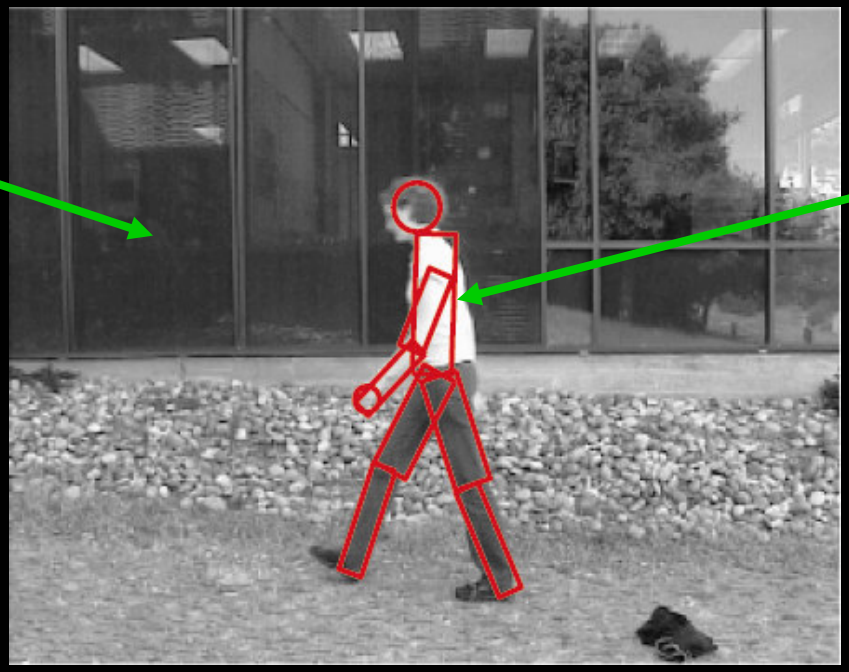
- Different cues and weightings change the model energy landscape

Explain the Image

(original slide courtesy of Michael Black, adapted)

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

Generic, unknown, background



Foreground person

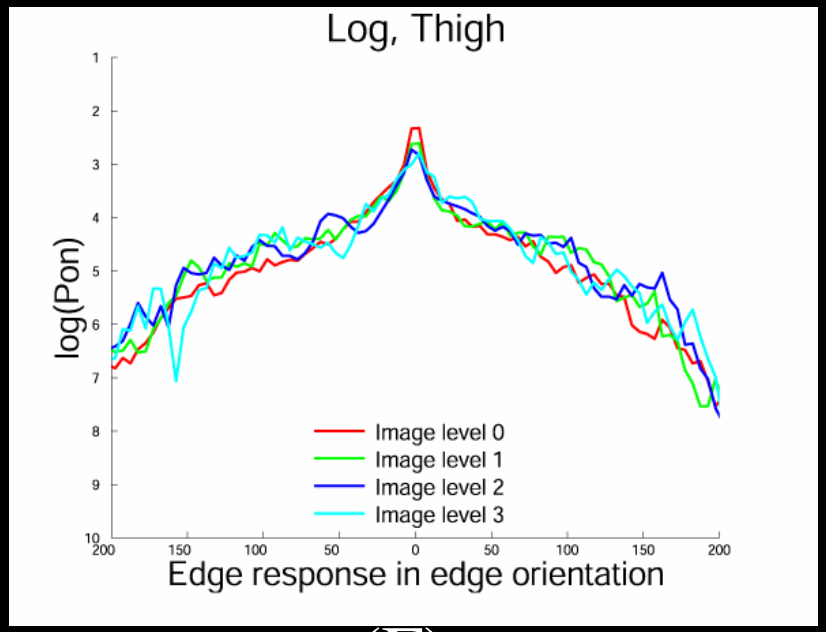
See:

Geman and Jednyak, PAMI '96
Sullivan et al, ICCV 1999;
McCormick and Isard, ICCV'01;
Sidenbladh and Black ICCV'01, IJCV03

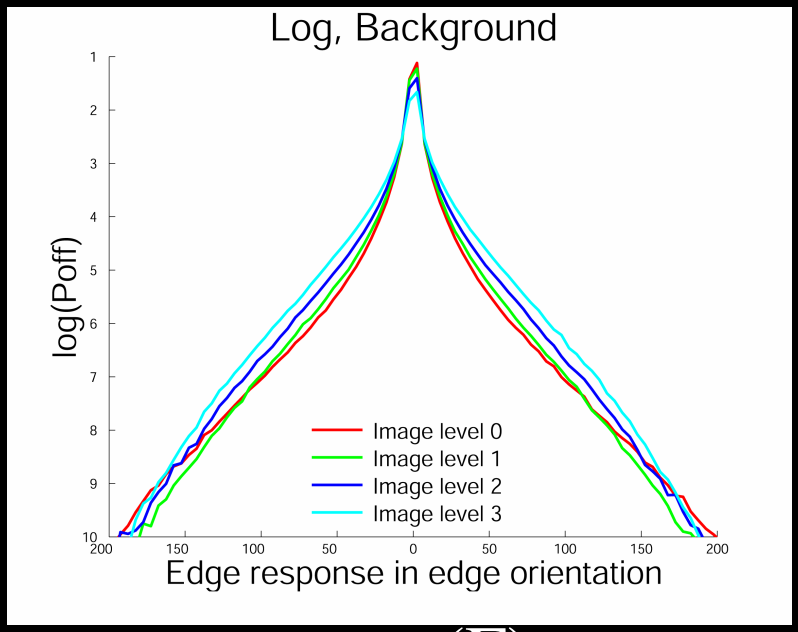
Foreground should explain what the background can't.

Empirical Distribution of Edge Filter Responses

(original slide courtesy of Michael Black)



$p_{on}(\mathrm{F})$



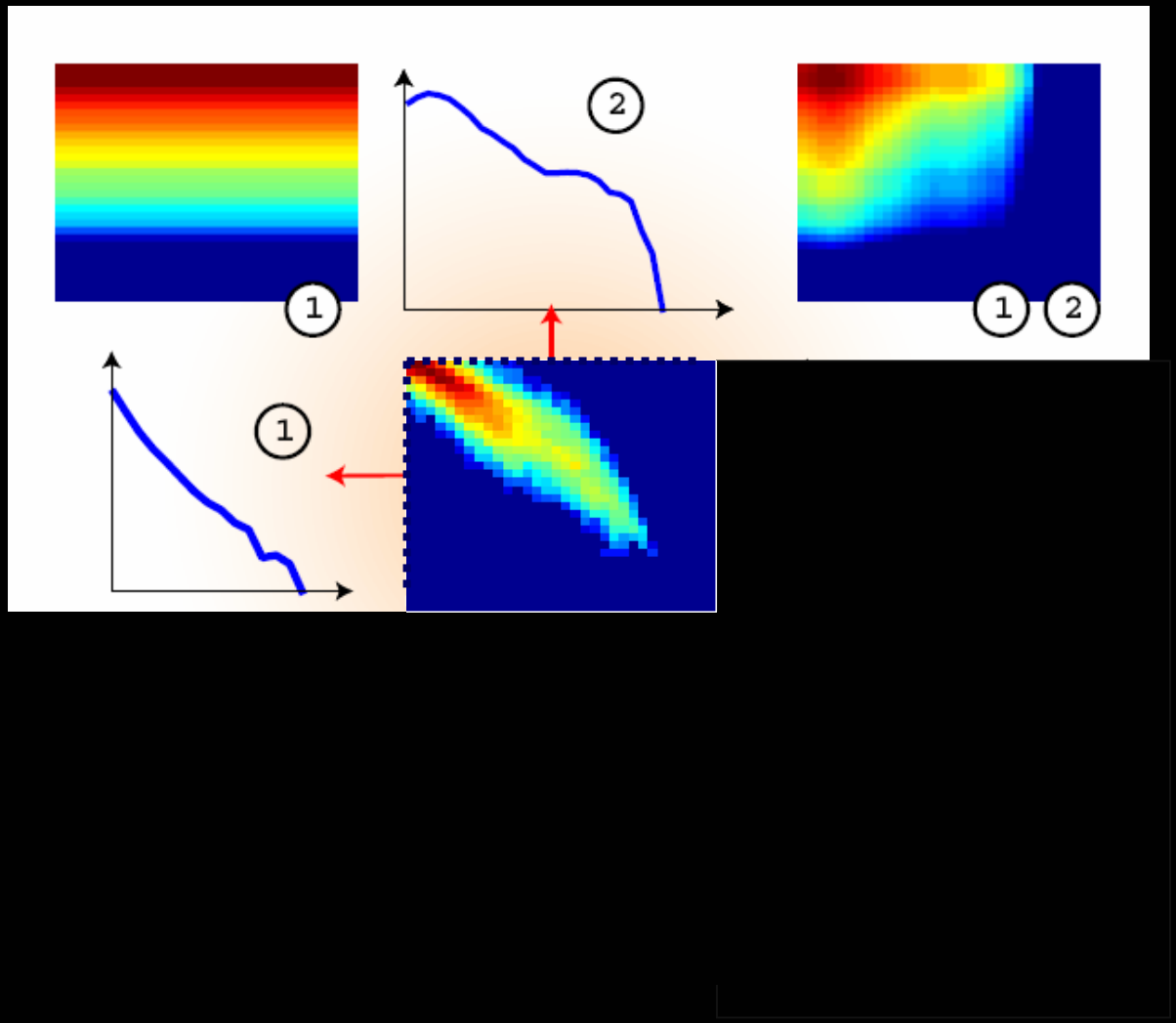
$p_{off}(\mathrm{F})$

Likelihood ratio, p_{on} / p_{off} , used for edge detection

Geman & Jednyak and Konishi, Yuille, & Coughlan

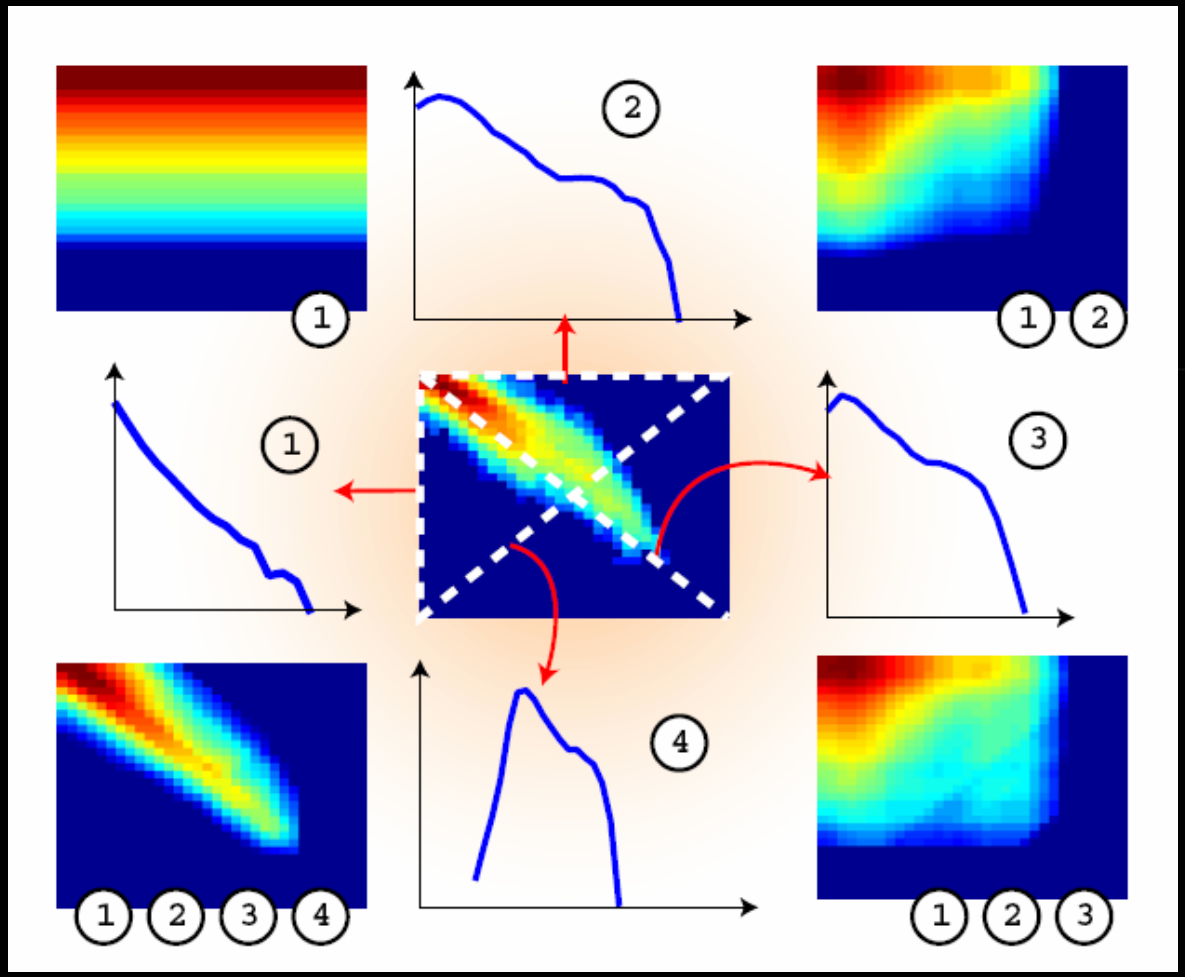
Learning Dependencies

(original slide courtesy of Michael Black); Roth, Sigal and Black, CVPR'04



Learning Dependencies

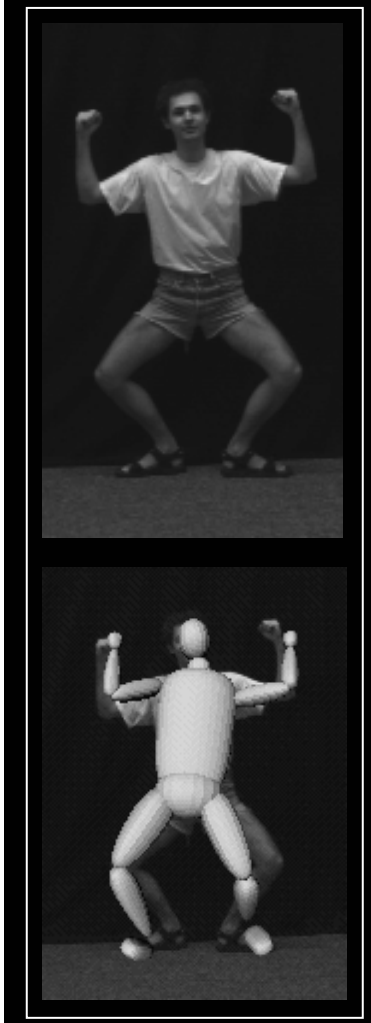
(original slide courtesy of Michael Black); Roth, Sigal and Black, CVPR'04

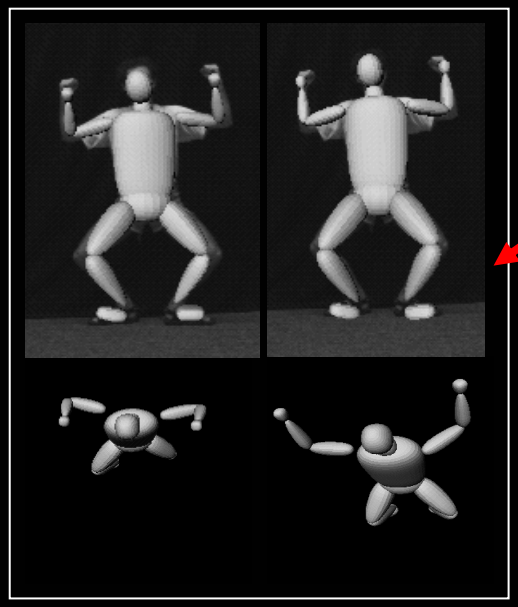


Filter responses are not conditionally independent

Leaning by Maximum Entropy

Difficulties for Generative Inference





Depth ambiguities

Occlusions

(missing data)

Left arm

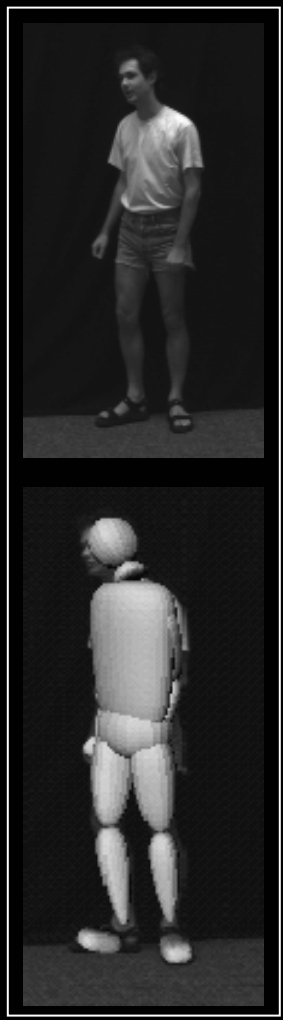


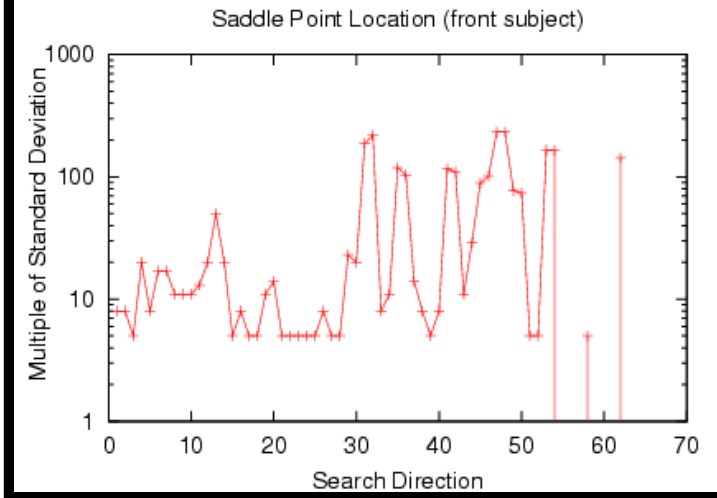
Image matching ambiguities

Left / right leg ?

Preservation of physical constraints

Observation Likelihood Properties (Model / Image Matching Cost)

- High dimension
  - at least 30 – 35 d.o.f.
- Very ill-conditioned
  - depth d.o.f. often nearly unobservable
- Many local optima
  - depth ambiguitiy *X* image ambiguity
- Optima are usually well separated
  - Merge and bifurcate during tracking
  - Passage through singular / critical configurations – frontoparallel limbs



Optima separation

$O(10^1)$ standard deviations

Modeling Dynamics $p(x_t|x_{t-1})$

Brief State-of-the Art

- Dedicated models of cyclic motion (*e.g.* walking), auto-regressive processes
  - *Blake et al'99'00, Sidenbladh et al'00*
- Bayesian Regression (RVM, GP)
  - *Aggarwal & Triggs'04, Sminchisescu et al'05, Wang et al'05*
- Multi-class dynamics (multiple ARPs)
  - *Blake et al'98, Pavlovic et al'01*
- Nearest neighbor search for motion pieces
  - *cf.* work in texture synthesis*, Sidenbladh et al'02*

Modeling Dynamics $p(x_t|x_{t-1})$

- If accurate, dynamic models are effective tracking stabilizers
  - If not, they may be harmful (see also *Balan et al, PETS@ICCV'05*)
  - Essentially a class of search heuristics
- Key Problems
  - How to deal with variability (*i.e.* individual motion styles)
  - How to efficiently model contextual branching
    - *e.g.* as opposed to a unimodal (Gaussian) prediction
    - *e.g.* as opposed to switching regimes with fixed relative probability
    - Need good models for complex multimodal *conditional* distributions
      - More on this later (*e.g.* conditional Bayesian mixture of experts)
  - What to do when the motion is not known in advance
    - Common practice to use white, unstructured noise

Presentation Plan

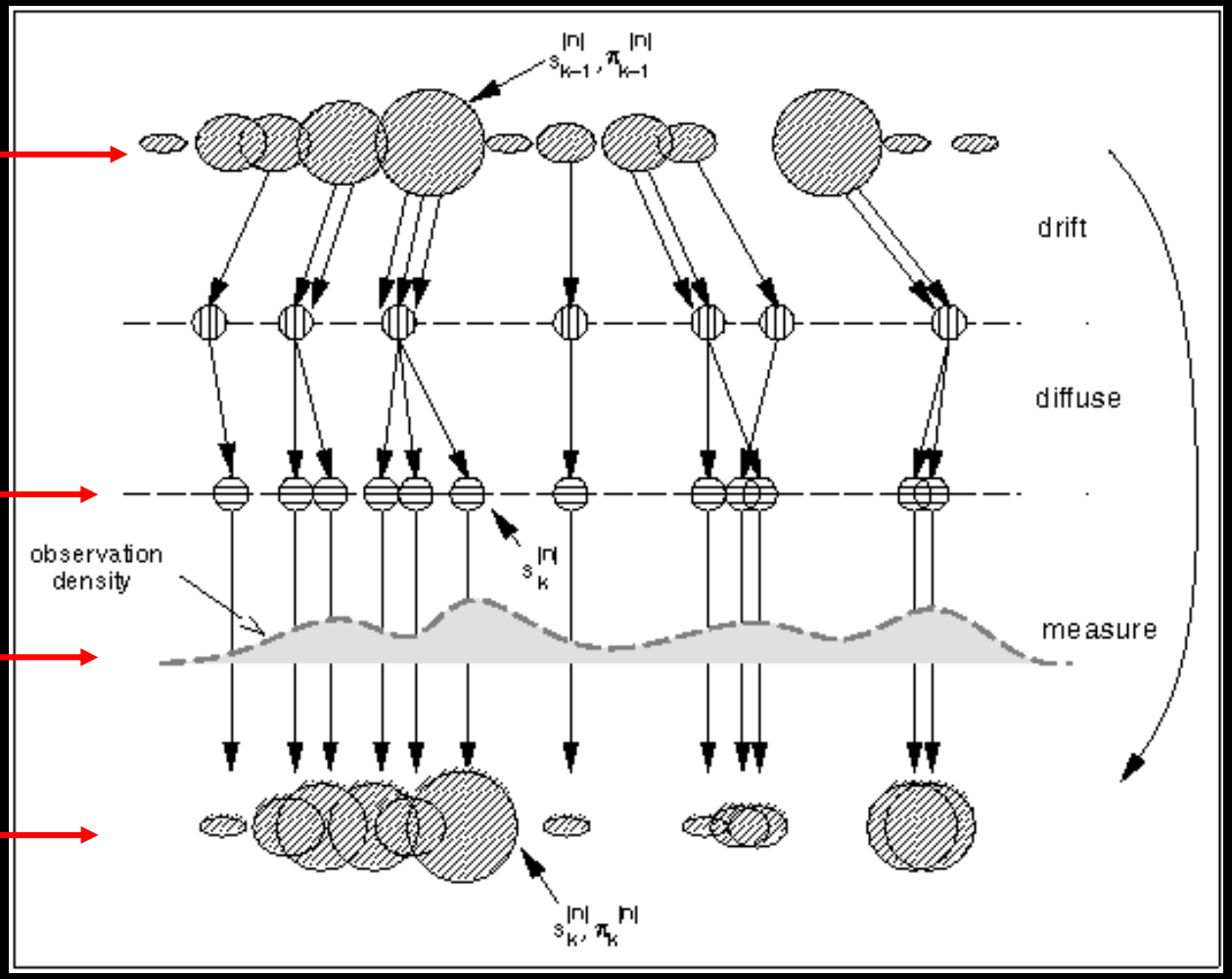
- Introduction, history, applications
- State of the art for 2d and 3d, human detection, initialization
- 3D human modeling, generative and discriminative computations
- Generative Models
  - Parameterization, shape, constraints, priors
  - Observation likelihood and dynamics
  - Inference algorithms
  - Learning non-linear low-dimensional representations and parameters
- Conditional (discriminative) models
  - Probabilistic modeling of complex inverse mappings
  - Observation modeling
  - Discriminative density propagation
  - Inference in latent, kernel-induced non-linear state spaces
- Conclusions and perspectives

Temporal Inference (Density Propagation)

- x_t state at time t
- $O_t = (o_1, o_2, \ldots o_t)$

observations up to time t

$p(x_t \mid O_t)$



time t

$p(x_{t+1} \mid O_t)$

$p(o_{t+1} \mid x_{t+1})$

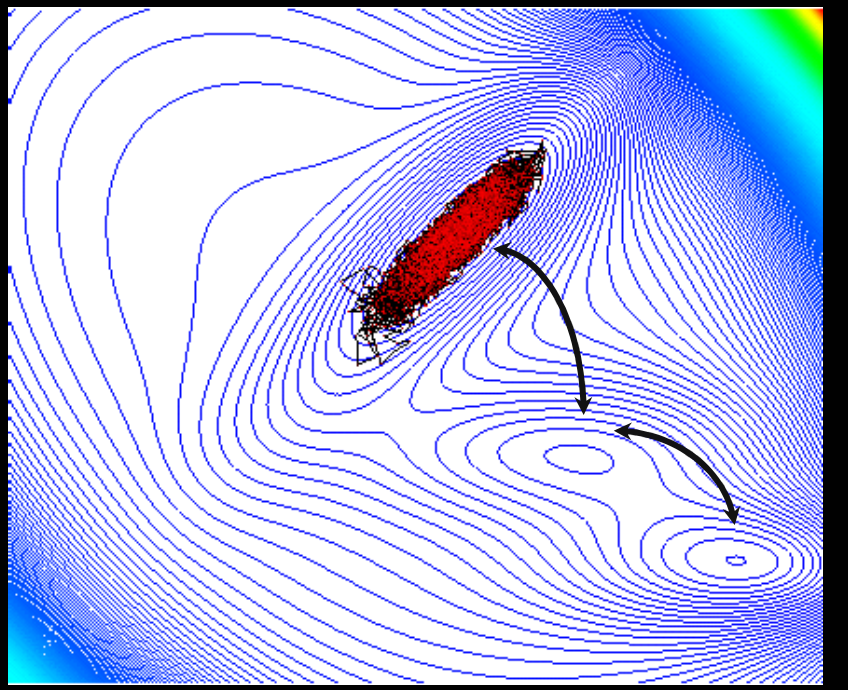
$p(x_{t+1} \mid O_{t+1})$

time t+1

cf. CONDENSATION, *Isard and Blake, 1996*

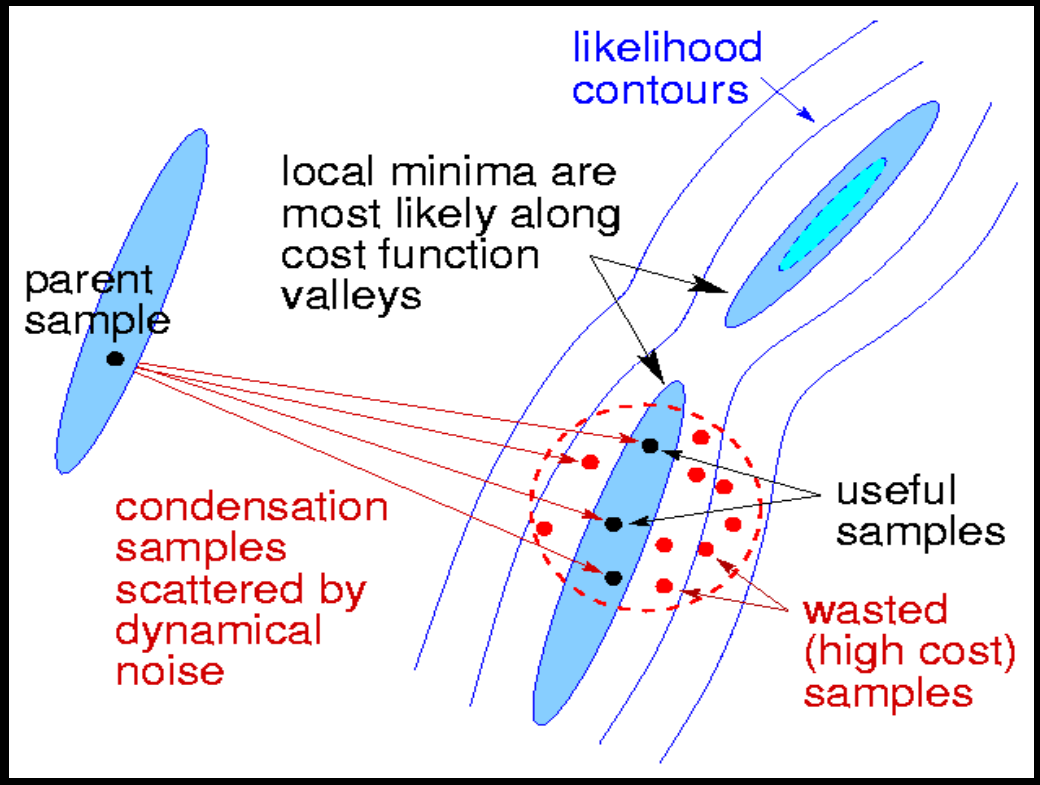
Sampling the Observation Likelihood
Problems: Trapping in Local Optima

- Adjacent likelihood peaks are typically separated by *many* standard deviations
  - *e.g.* reflective ambiguities, false model-image correspondences
- The state density is too localized to be a good search hypothesis generator
  - Samples almost never reach nearby peaks



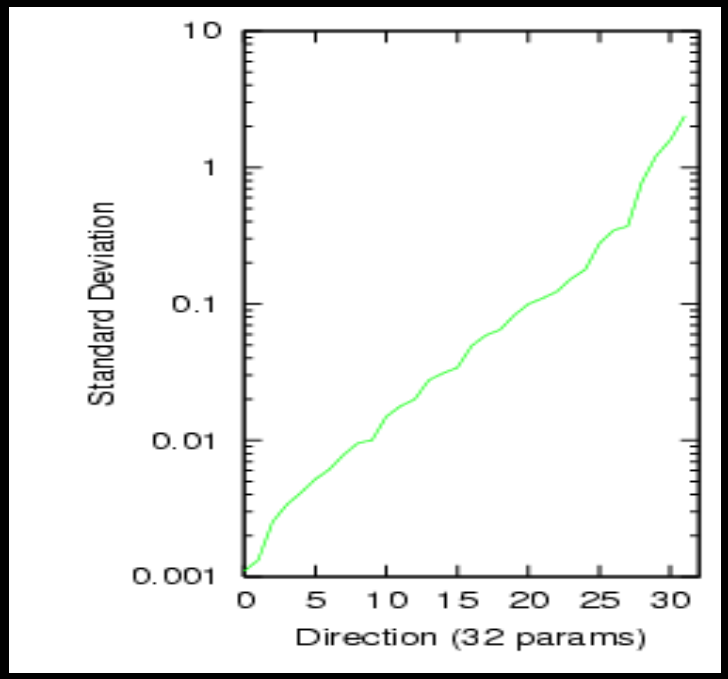
Devote some of the samples to more global search

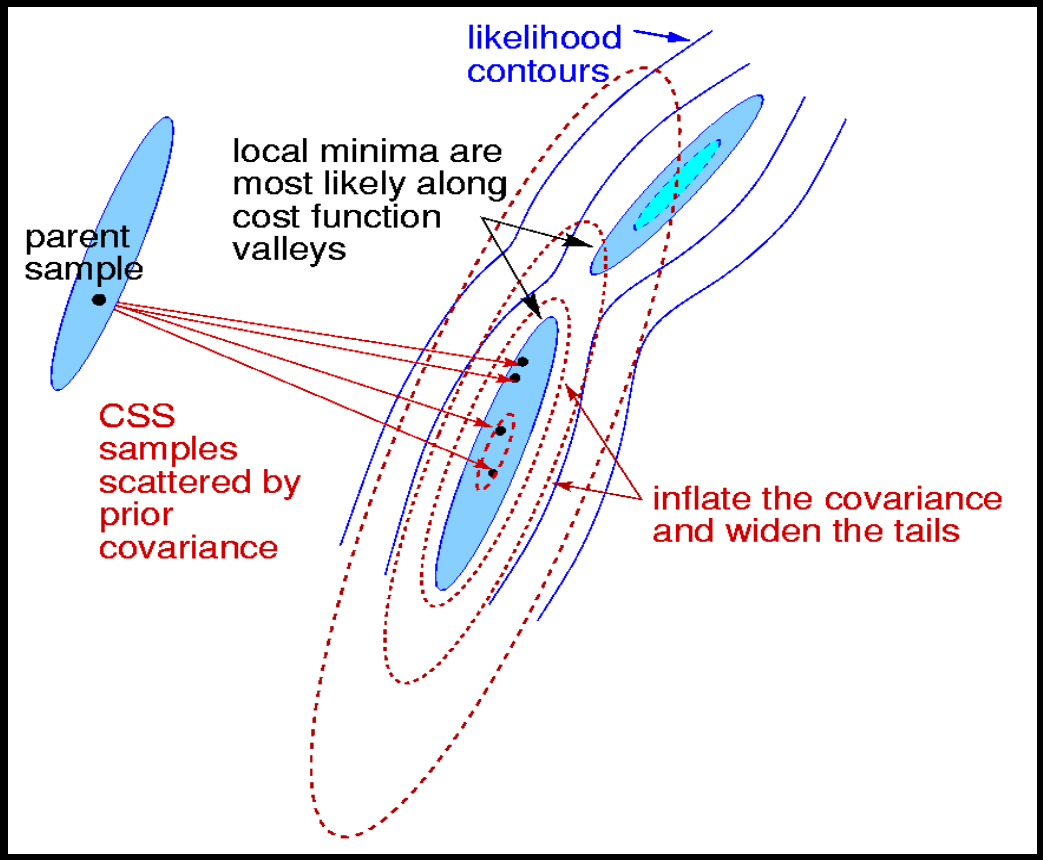
Why Uniformly Boosted Dynamics Wastes Sampling Resources



- For ill-conditioned / high dimensional problems any nearly ‘uniform’ noise causes either
  - insufficient scatter if it is small
  - *massive sample wastage* if it is large

Covariance Scaled Sampling



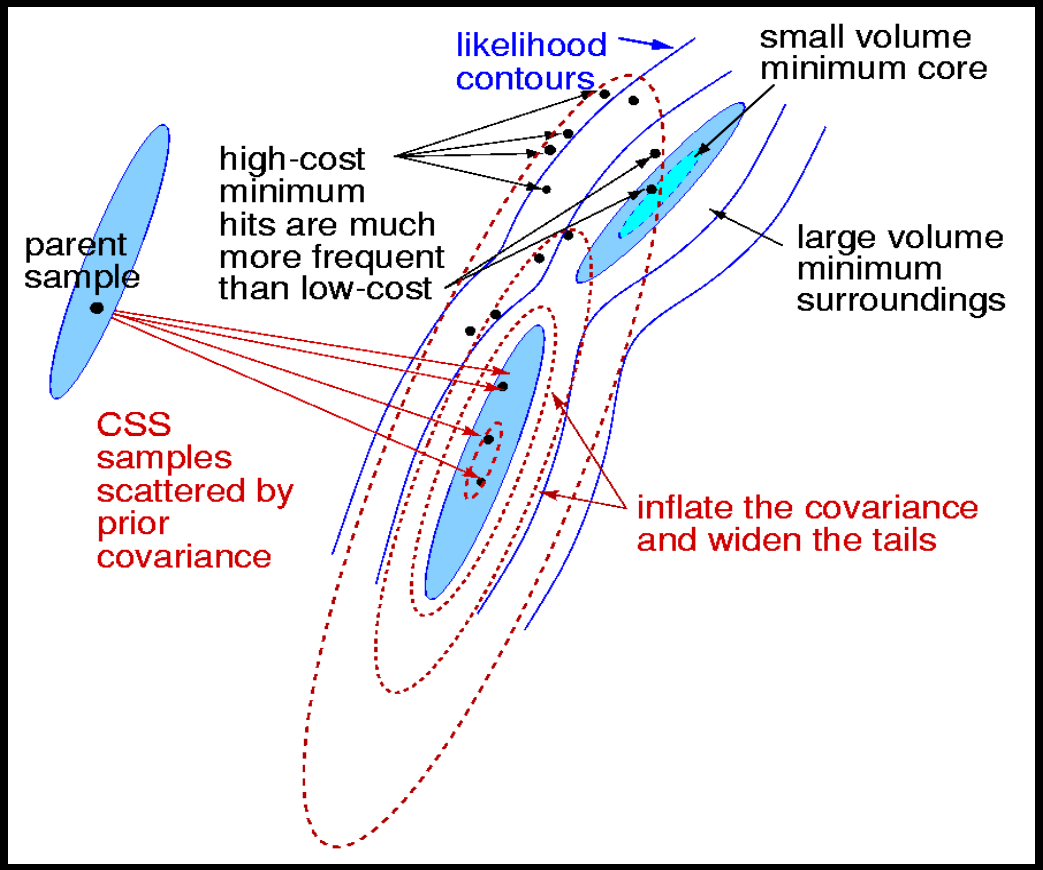


$\sigma_{max}/\sigma_{min} \sim 2000$

- Avoids large volume wastage factors
- Condensation vs. CSS volume factor ~ 10^{54}
  - For comparable sample spread

Sampling is not enough

- In high dimensions volume increases very rapidly with radius
- High cost samples are not resampled, so …



To find minima efficiently, it's necessary to optimize locally after sampling

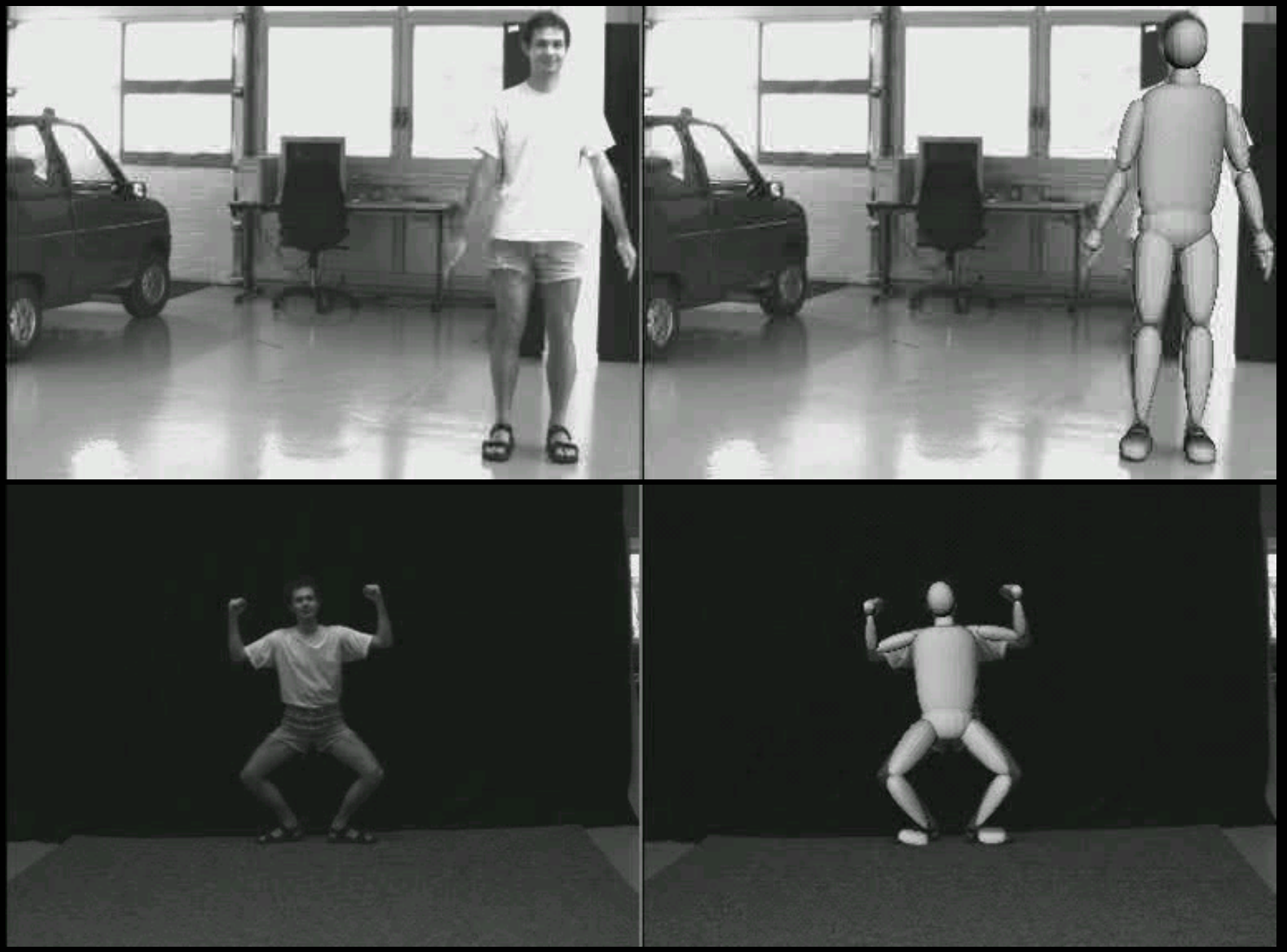
Covariance Scaled Sampling Algorithm

Sminchisescu & Triggs, CVPR 2001, IJRR 2003

A density propagation method combining

- Mode + Covariance mixture representation
- Wide tailed sampling to reduce trapping
- Covariance scaling to reduce volume wastage
- Local optimization (robust, constraint consistent) to reduce needle-in-haystack effect

Covariance Scaled Sampling <v>



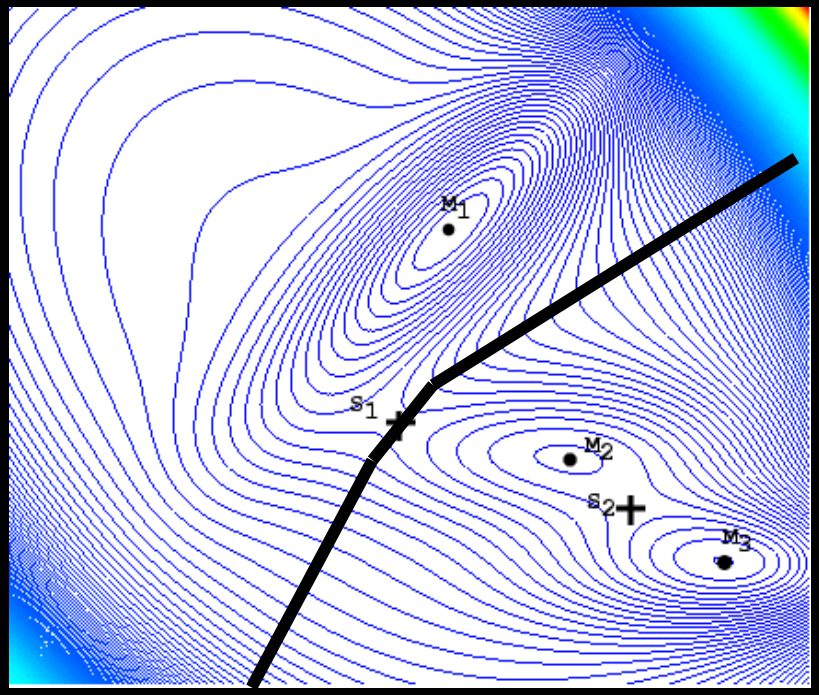
Sampling Adaptation

- The CSS sampling criteria is yet local
- Regime adapted only based on covariance at optima
  - The energy landscape changes further away
  - How wide to spread ?

Are there longer-range forms of adaptation?

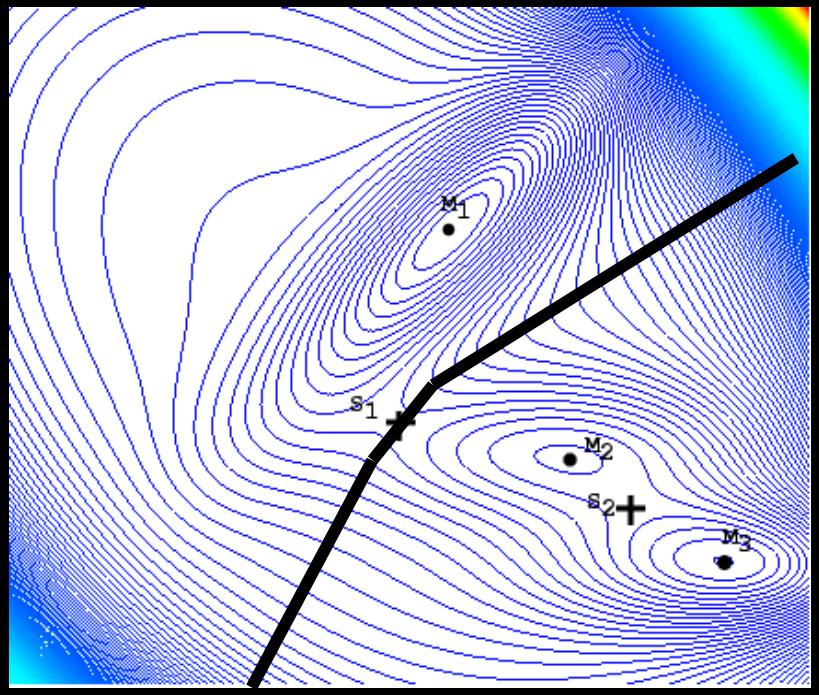
Can we better focus the samples?

The Dividing Surface (DS)



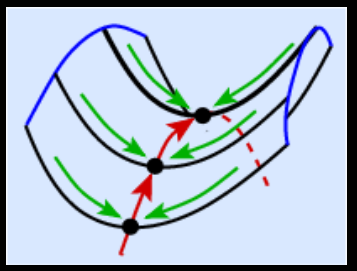
- For a minimum M_1
  - n 1dimensional surface separating M_1 from its neighbors
  - Essentially *non locally defined*
  - Steepest descent trajectories converge to *other* minima than M_1
  - May include positive curvature regions, various saddles…

The Dividing Surface (contd.)



- Sampling involves
  - Long periods of exploration of one minimum
  - Infrequent transitions to other minima through ***low DS regions***
- The flux through DS defines the inter-minima transition rates
- But the low DS regions of high-flux are the **transition states** !

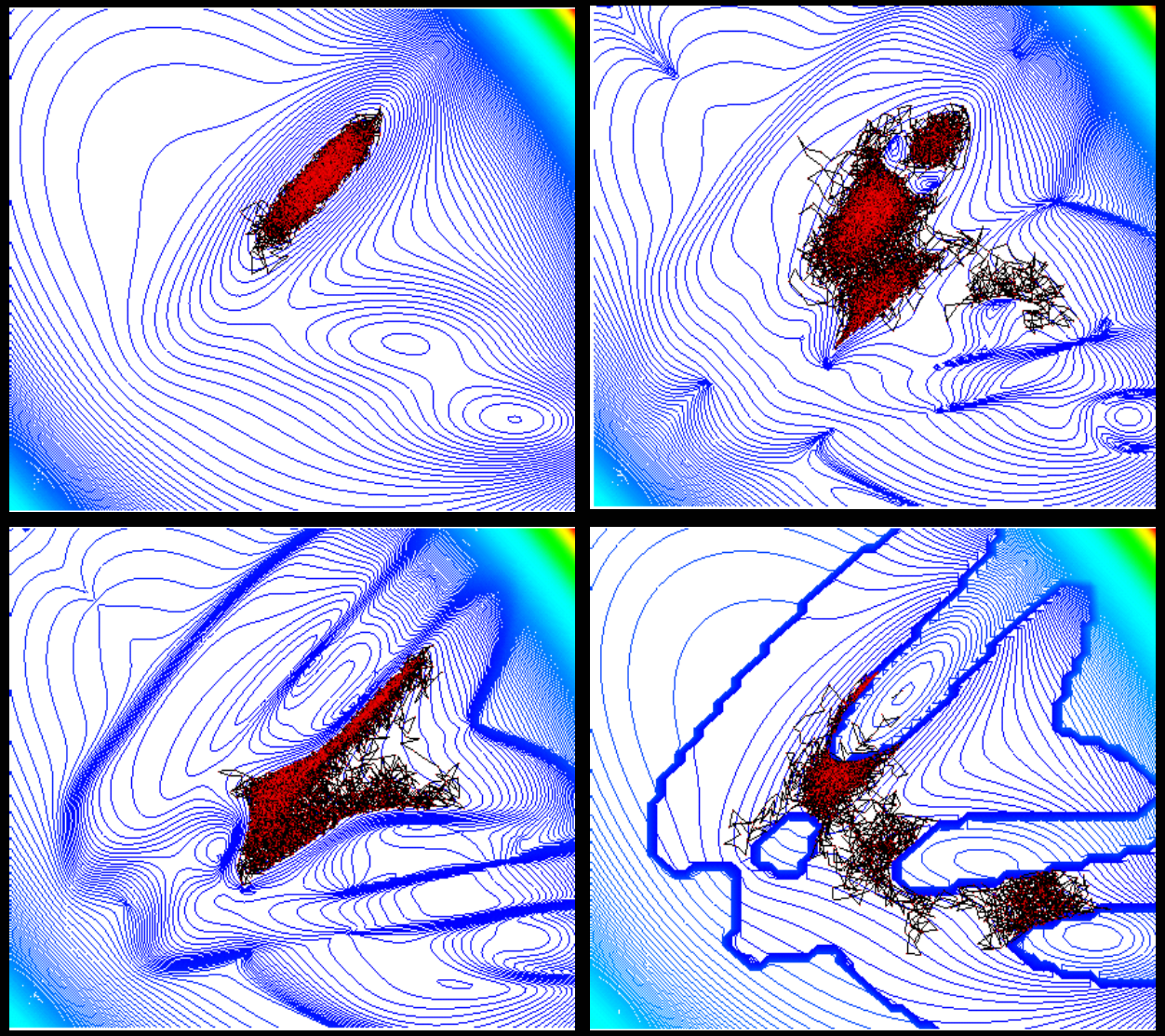
Importance of Transition Neighborhoods



- ***Transition States (TS)*** are
  - Co-dimension 1 saddle points
  - Zero gradient, 1 negative, (n-1) positive curvatures
  - Cols rather than mountain tops
- Low cost saddles (TS) lead to low cost minima
- Provide useful *local* approximation to the DS
  - f = cost function, g = gradient, (e_1, V_1) = smallest Hessian (eigenvalue, eigenvector)

$$g_{p1} = V_1^T g = 0 \quad , \quad e_1 < 0$$

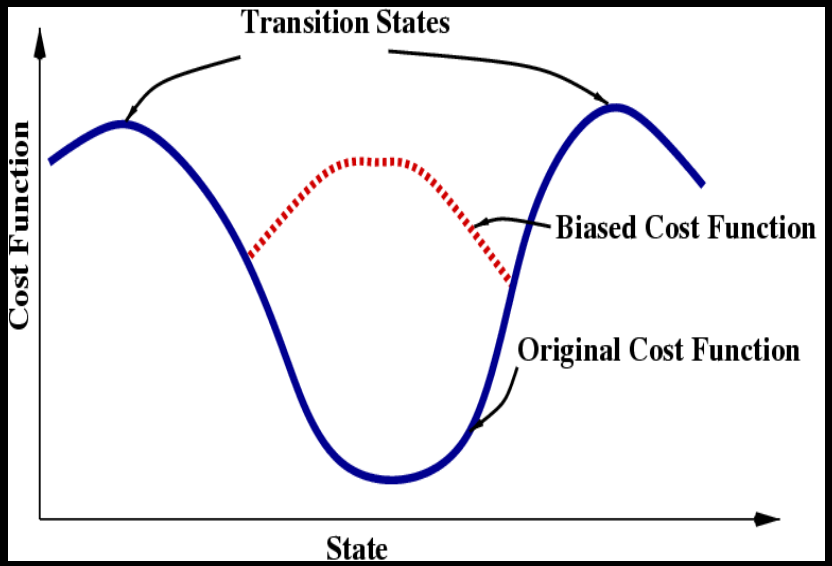
Sampling a modified potential *P(h,d)*



The Adaptive Bias Potential

- Add a bias $f_b(x)$ to the cost

$$f_b = \frac{h}{2}\left[1 + \frac{e_1}{\sqrt{e_1^2 + g_1^2 / d^2}}\right]$$



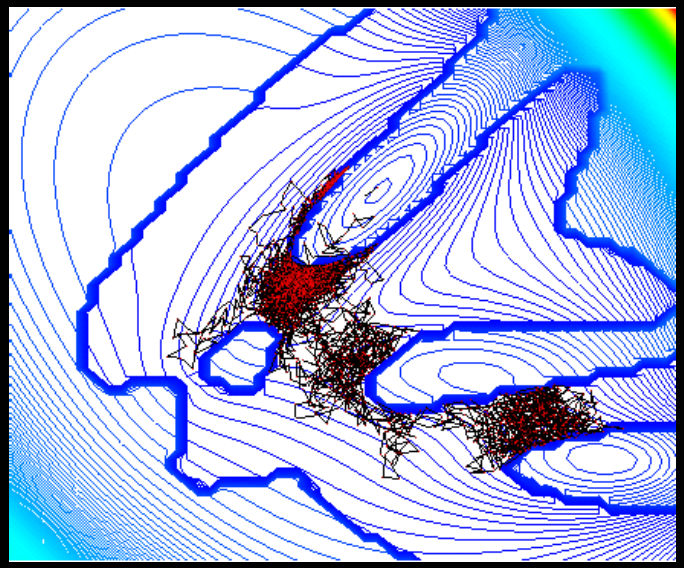
  - h is height of bias well
  - d is a length scale (≈ distance to nearest minimum)
  - e_1 is smallest Hessian eigenvalue
  - g_1 is cost gradient in first eigendirection

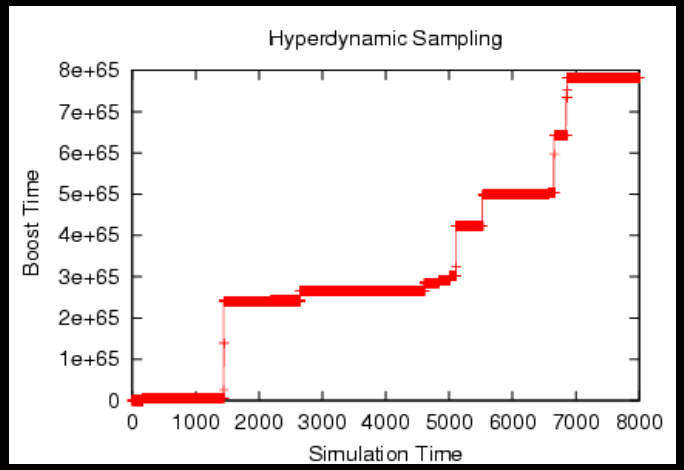
- $f_b(x)=0$ at a saddle ($e_1 < 0$, $g_1 = 0$)

Hyperdynamics

Sminchisescu & Triggs, ECCV 2002, IVCJ 2005

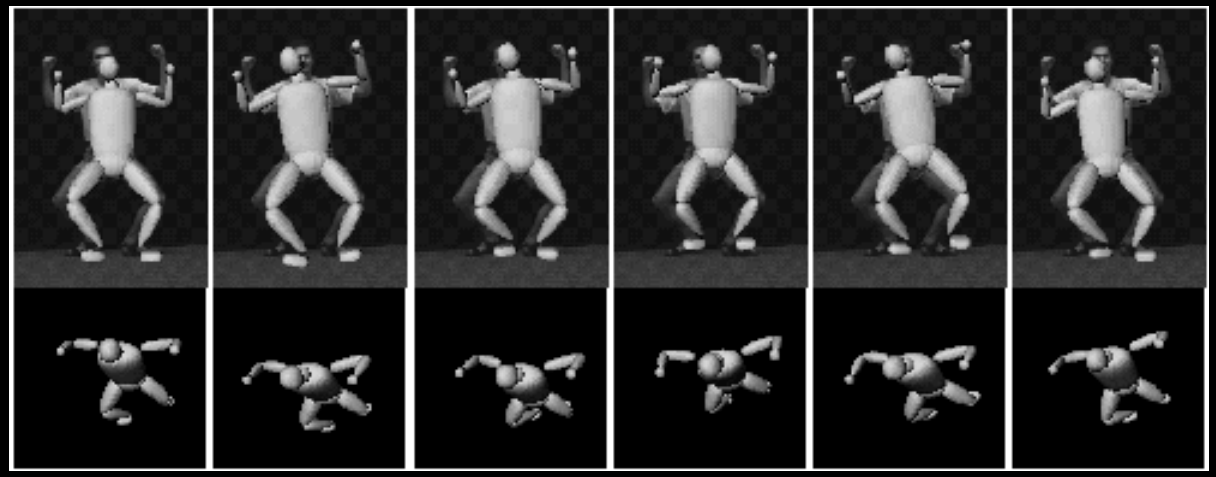
- A generalized adaptive search broadening method
- Focuses samples near transition states by ***adaptively*** raising the cost
  - in the cores of local minima
  - in high gradient regions
- Exponentially increased probability of reaching a transition state



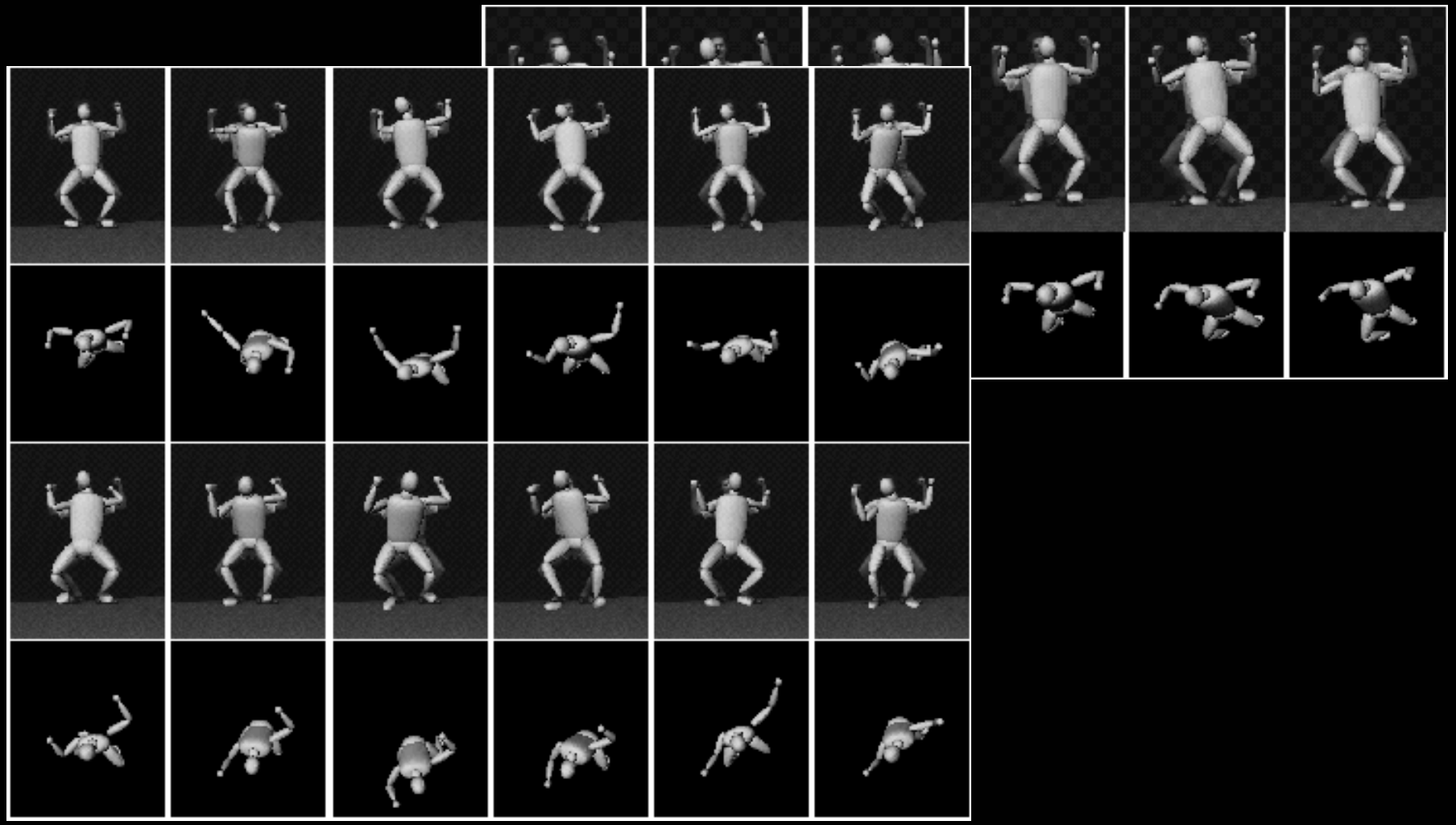


Pure MCMC vs. Hyperdynamics

Pure MCMC



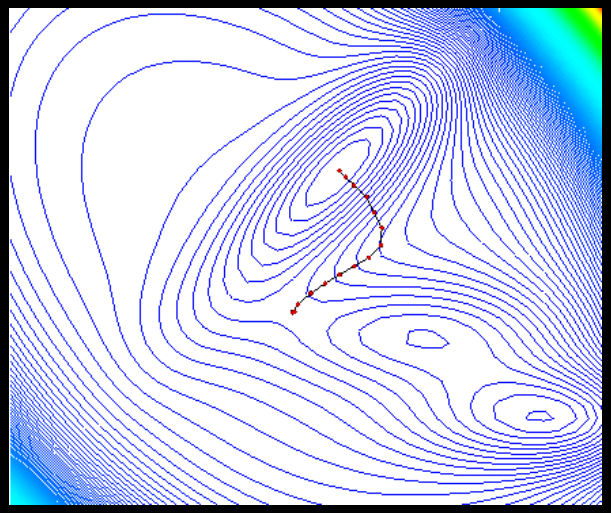
Pure MCMC vs. Hyperdynamics

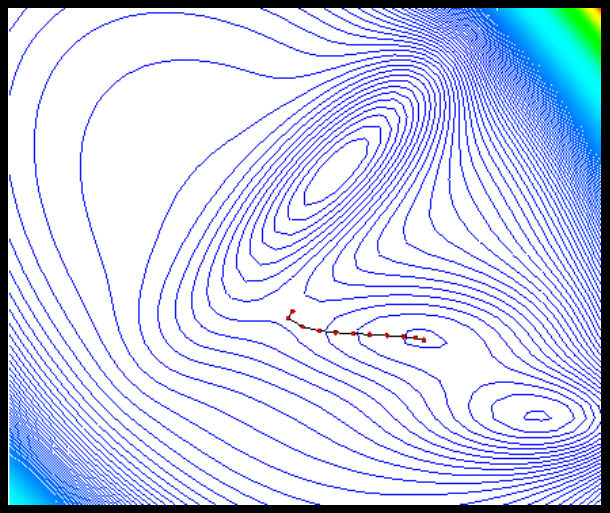


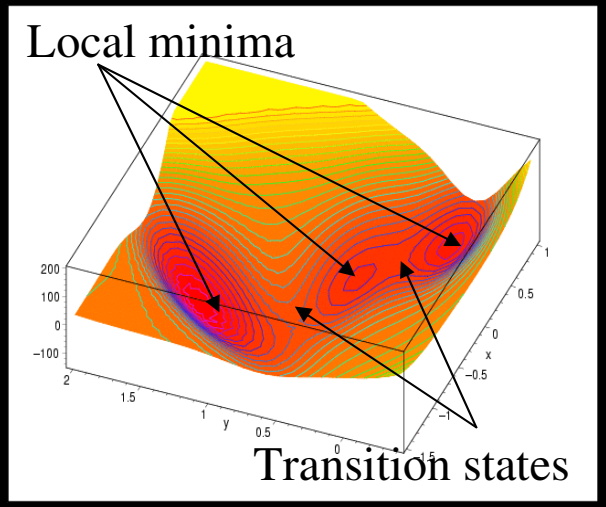
- Hyperdynamics mixes better and explores multiple minima (8000 simulation steps)

How can we find nearby minima deterministically ?

Sminchisescu & Triggs, ECCV 2002, IJCV 2005



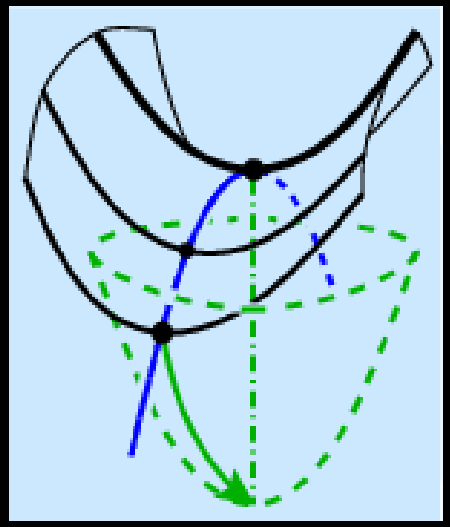




From any ***Transition state (saddle point),*** we can usually slide downhill to an adjacent minimum

- Low cost saddles lead to low cost minima

Local Minimization versus Saddle Point Search



- Minimization
  - many local descent based algorithms
  - measure progress by decrease in function values
  - local ‘sufficient decrease’ ensures ‘global convergence’ to some local minimum
- Local Saddle Point Search
  - ascend, using modified descent algorithms
  - no universal progress criterion
  - initialization needed (e.g. ascent direction)

Newton Minimization

- Pure Newton iteration

$$0 = g(x + \delta x) \approx g(x) + H\,\delta x \quad \Rightarrow \quad \delta x = -H^{-1} g$$

  - efficient near minimum, but globally unreliable
  - may diverge, converge to any type of stationary point, etc

- Damped Newton iteration
  - globalize convergence by adding 'damping' matrix D

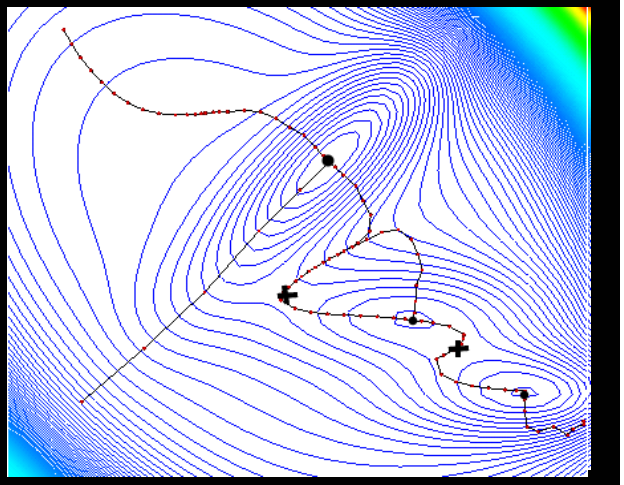
$$\delta x = -(H + \lambda D)^{-1} g$$

  - in an eigenbasis where D = I

$$\bar{\delta} x(\lambda) = \left(\frac{\bar{g}_1}{\lambda_1 + \lambda}, \ldots, \frac{\bar{g}_n}{\lambda_n + \lambda} \right)^T \quad , \quad \lambda \geq -\min(0, \lambda_1, \ldots, \lambda_n)$$

Eigenvector Tracking

- Choose an ascent eigenvector *'k'* and track it as you progress
- Move uphill along '*k*' and downhill along all other eigenvectors

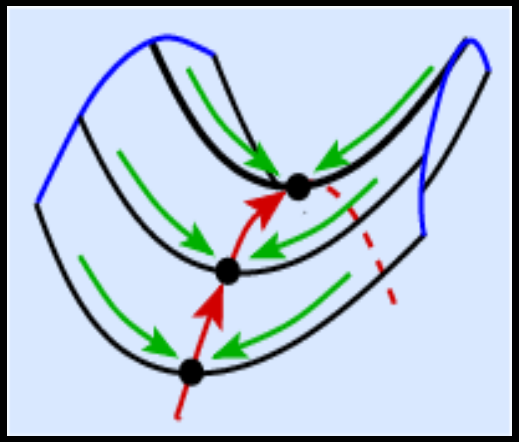


- Uses modified damping signs to minimize an equivalent local virtual cost model $\sigma_k = -1, \sigma_{i \neq k} = +1$

$$\bar{\delta}x(\lambda) = \left(\frac{\bar{g}_1}{\lambda_1 + \sigma_1 \lambda}, ..., \frac{\bar{g}_n}{\lambda_n + \sigma_n \lambda} \right)^T \quad , \quad \lambda \in (-\min(0, \sigma_1 \lambda_1, ..., \sigma_n \lambda_n), +\infty)$$

- But tracking 'the same' eigenvector is hard
  - when eigenvalues cross, their eigenvectors slew around rapidly...

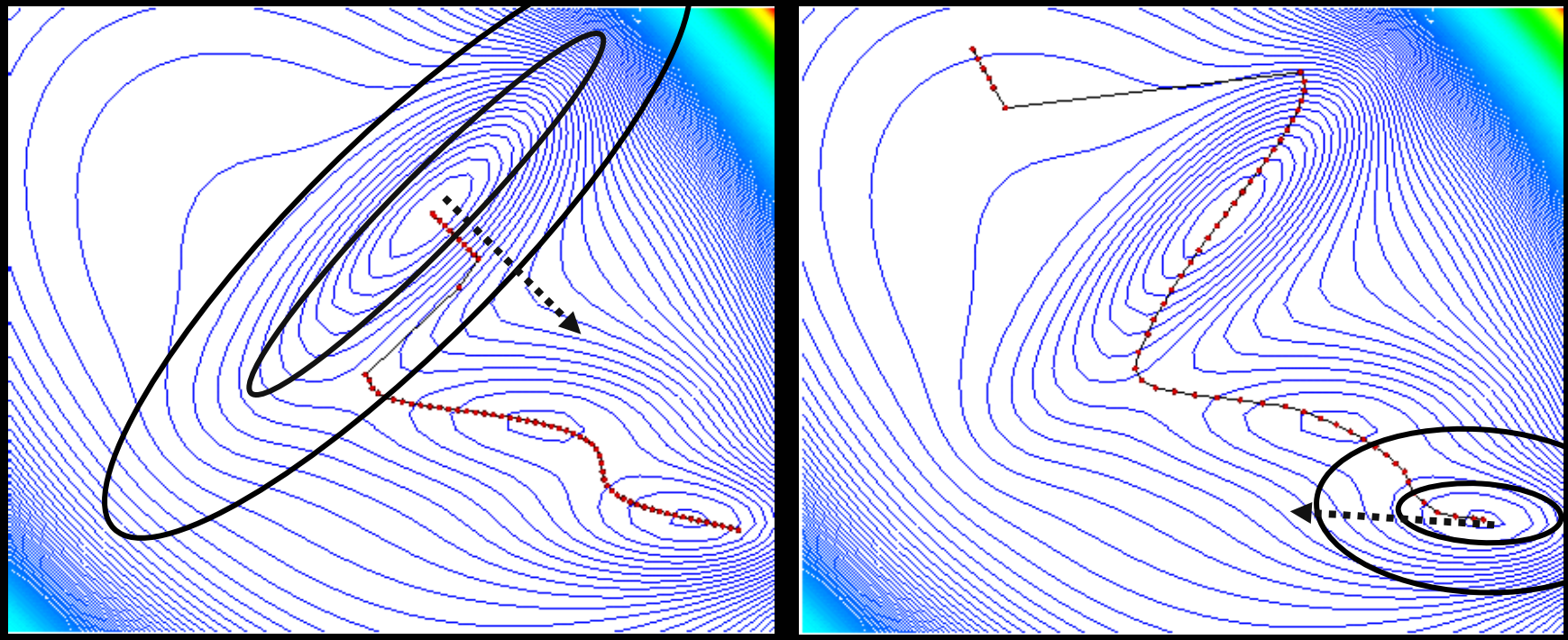
Hypersurface Sweeping



- (reminder) A codimension 1 saddle is
  - a local maximum in one direction
  - a local minimum in the other *n-1*
- Sweep space with a moving hypersurface
  - hyperplane, expanding hyper-ellipsoid…
- Find & track a local minimum on cost surface
  - optimize over hypersurface
- Find local maxima of track
- Hypersurface vs. cost isosurface relative curvatures control search diversity

Hypersurface Sweeping

- Long trajectories started in one minimum
  - Hypersurface ≈ cost isorsurface, but flattened orthogonally w.r.t. the search direction
  - Find other minima and saddles
  - For illustration didn't stop after each saddle detection



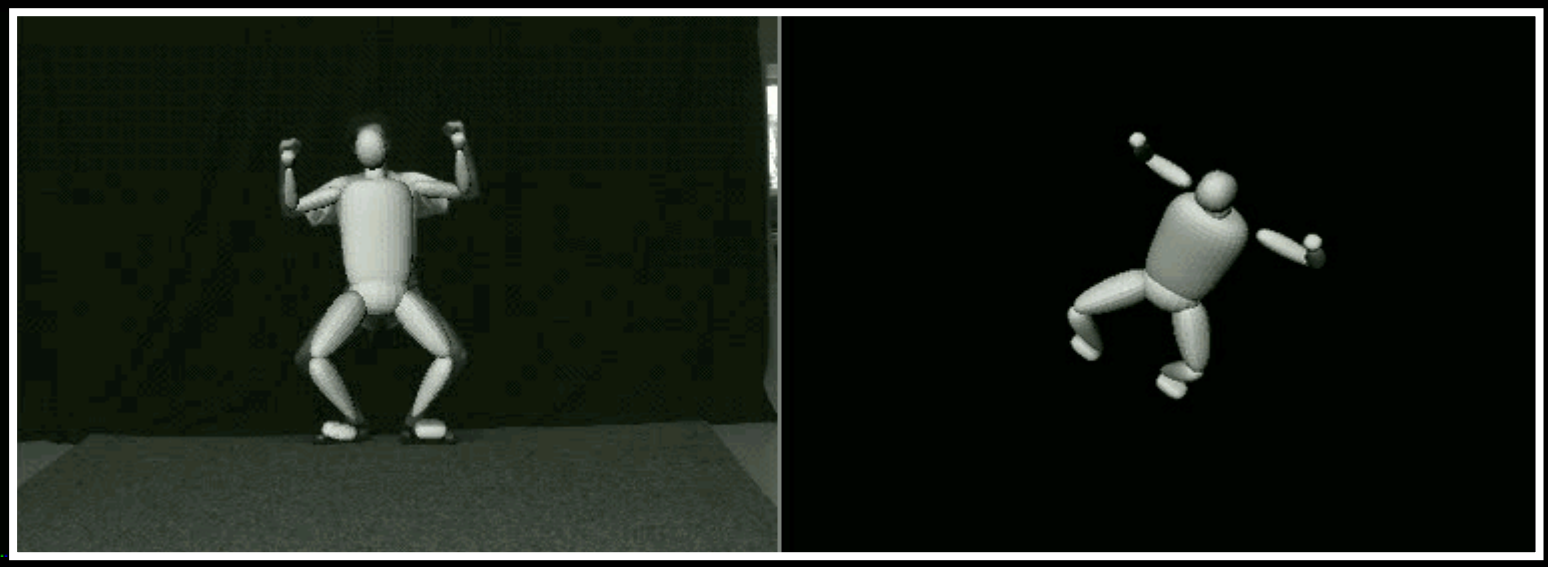
Minima Caused by Incorrect Edge Assignments

Intensity + edges

Edges only

Why is 3D-from-monocular hard? <v>

Reflective, Kinematic, Pose Ambiguities



- Monocular static pose optima (Eigenvector Tracking)
  - ~ $2^{\text{Nr of Joints}}$, some pruned by physical constraints
  - Temporally persistent

Quantitative Performance

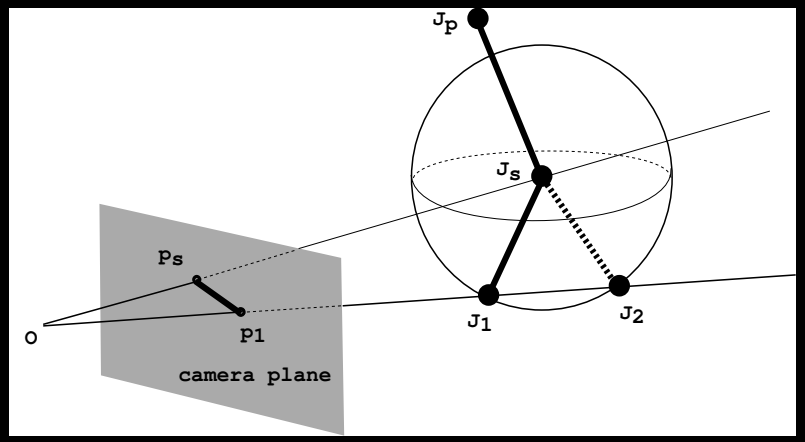
Method	Number of Detected Minima	Median State Distance	Median Standard Deviations	Median Cost
ET	55%	4.45	79.6	3.39
HS	60%	5.01	91.6	3.87

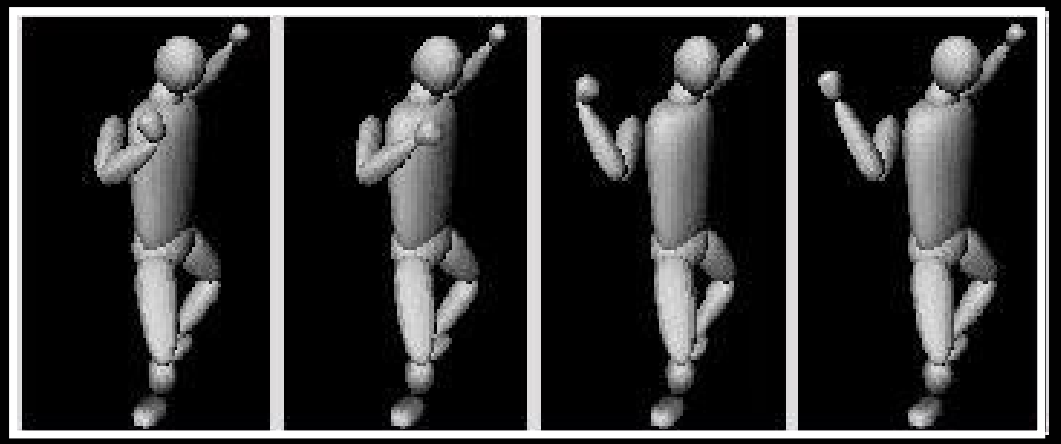
- Evaluate the search for static pose estimation optima
- Initialize in different optima and search along different principal curvature directions
- Good localization of distant solutions

Exploiting Problem Structure (Symmetries) During Search

Kinematic Jump Sampling (KJS)

Sminchisescu & Triggs, CVPR 2003

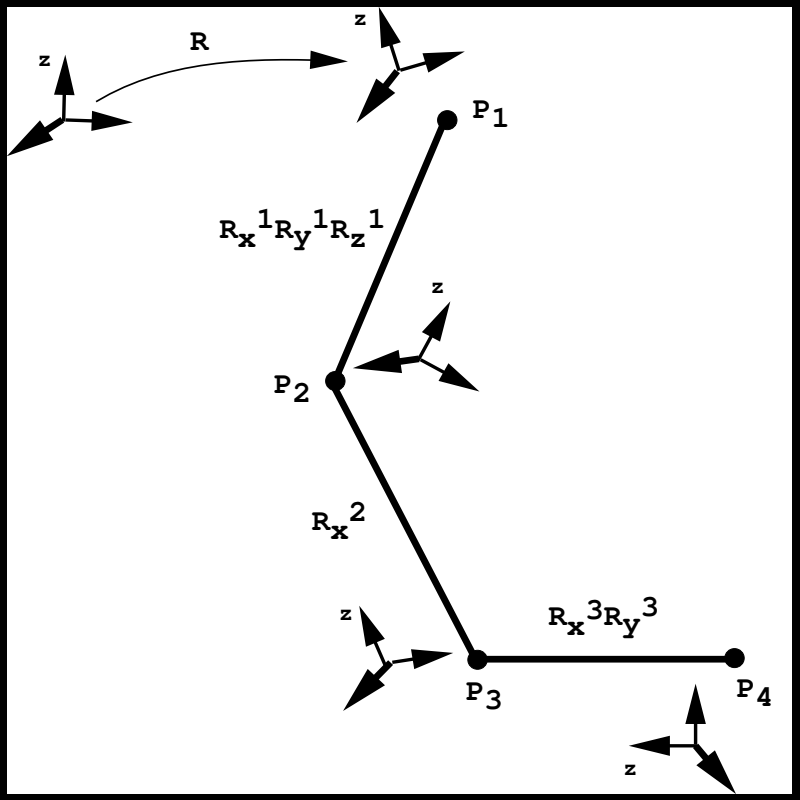




- For any given model-camera configuration, we can *explicitly* build **the interpretation tree of alternative kinematic solutions** with **identical** joint projections
  - work outwards from root of kinematic tree, recursively evaluating forward/backward 'flips' for each body part
    - Alternatively, sample by generating flips randomly
    - CSS / ET / HS still needed to handle matching ambiguities

Efficient Inverse Kinematics

- The inverse kinematics is simple, efficient to solve
  - Constrained by many observations (3D articulation centers)
  - The quasi-spherical articulation of the body
  - **Mostly in closed form**

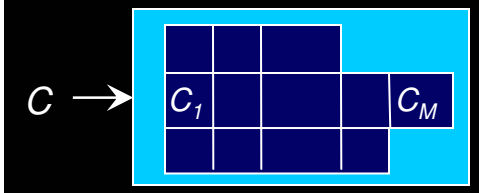


- The iterative solution is also very competitive
  - Optimize over model-hypothesized 3D joint assignments
  - 1 local optimization work per new minimum found

➢ An adaptive diffusion method (CSS) is necessary for correspondence ambiguities

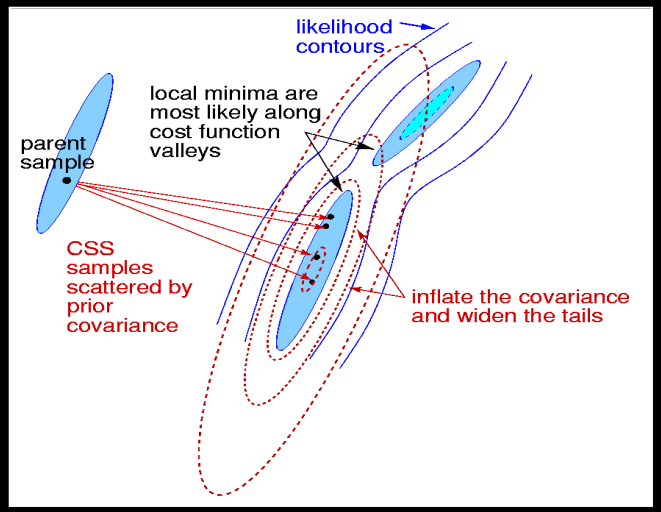
The KJS Algorithm

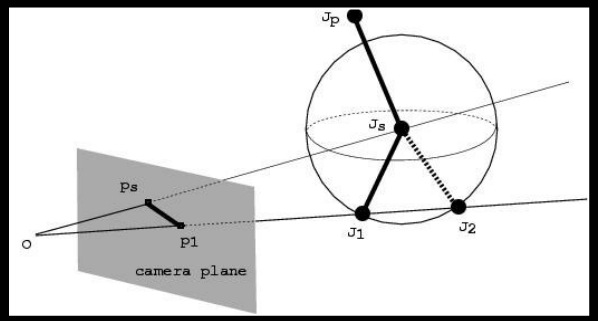
Candidate Sampling Chains



C=SelectSamplingChain(m_i)

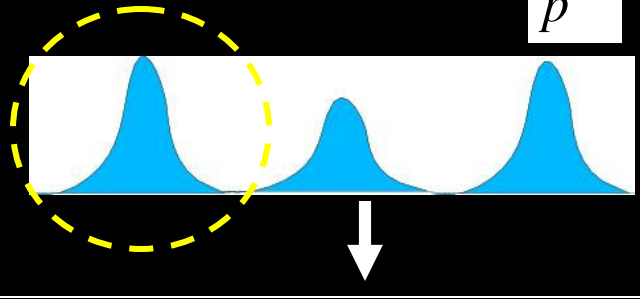
$$\text{vote}(C) = \sum_{i=1}^{M} \sum_{j=1}^{N} \sigma_j \mathrm{v}_j[C_i]$$





$m_i = (\mu_i, \Sigma_i)$

p^{t-1}

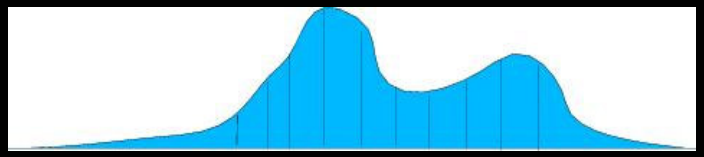


s=CovarianceScaledSampling(m_i)

T=BuildInterpretationTree (s,C)

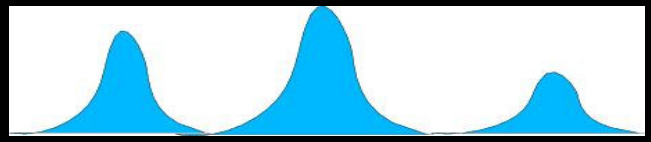
E=InverseKinematics(T)

E



Prune and locally optimize *E*

p^t



Jump Sampling in Action <v>

Quantitative Search Statistics

Method	Scale	Number of Minima	Median Parameter Distance		Median Standard Deviation		Median Cost	
			NO OPT	OPT	NO OPT	OPT	NO OPT	OPT
KJS1	-	1024	2.9345	2.8378	92.8345	93.9628	0.0998	0.0212
KJS2	-	1466	3.2568	2.2986	83.4798	82.5709	0.1045	0.0203
CSS	1	8	1.1481	2.5524	10.9351	47.6042	116.9512	8.4968
CSS	4	59	3.2123	2.9474	35.2918	55.3163	1995.1232	6.9810
CSS	8	180	4.9694	3.3466	75.1119	109.8131	16200.8134	7.0986
CSS	16	667	6.4242	6.7209	177.1111	465.8892	45444.1223	8.6958
CSS	1/HT	580	5.0536	6.9362	106.6311	517.3872	15247.7134	8.7242
SS	1	0	0.1993	-	24.5274	-	273.5091	-
SS	4	11	0.7673	2.0492	96.1519	39.0745	4291.1211	6.2801
SS	8	42	1.4726	2.5488	188.1571	56.8268	16856.1211	6.9648
SS	16	135	2.7195	2.8494	367.7461	87.8533	63591.4211	8.6958
SS	1/HT	232	2.1861	6.5474	178.6471	535.9991	18173.1121	17.8807

- Initialize in one minimum, different sampling regimes
- Improved minima localization by KJS
  - Local optimization often not necessary

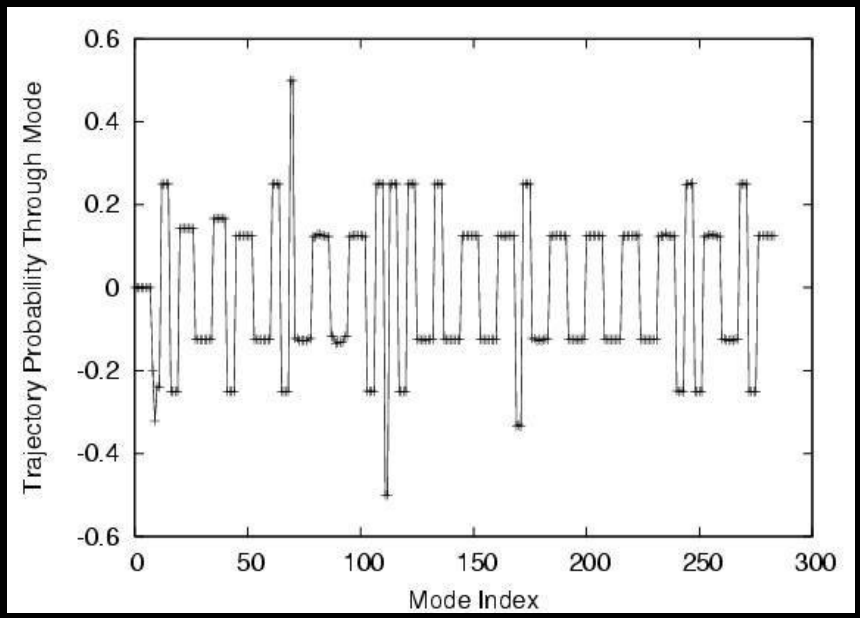
How many trajectories are here?

- KJS efficiently provides body pose hypotheses each time step
  - this multimodality is temporally persistent and not transient
- Often we select (and show) the most probable trajectory
  - this combines both good image fit and smooth dynamics

Are there other trajectories that are qualitatively different yet equally probable? YES, there are …

Why is 3D-from-monocular hard? <v>

Multimodal trajectory distribution



- Observation likelihood mixture has (here) up to 8 modes per timestep
  - flip sign in-between timesteps for visualization (right plot)
- Fewer modes where the uncertainty diminishes (*e.g.* index 100-150)
  - Regions where the arms are in front of the face (physical priors limit uncertainty)

Approximate the Trajectory Distribution

$$P(\mathrm{X},\mathrm{R}) = p(x_1)\,p(r_1 \mid x_1)\prod_{t=2}^{T} p(x_t \mid x_{t-1})\,p(r_t \mid x_t)$$

- $p(x_1)$ is the state space prior
- $\mathrm{X} = (x_1,...,x_T)$ is the joint state (trajectory) vector
- $\mathrm{R} = (r_1,...,r_T)$ is the joint observation vector
- $\mathrm{R}_t = (r_1,...,r_t)$ denotes the observation vector up to time t
- $p(x_t \mid x_{t-1})$ is the non-linear dynamic transition rule
- $p(r_t \mid x_t)$ is the non-linear non-Gaussian observation model

We search a tractable approximation q^θ to P that mimizes the relative entropy:

$$\mathrm{D}(q^\theta \,\|\, P) = \int_{\mathrm{X}} q^\theta(\mathrm{X}) \log \frac{q^\theta(\mathrm{X})}{P(\mathrm{X} \mid \mathrm{R})}$$

Equivalent to minimizing the variational free energy F:

$$F(q^\theta, P) = \mathrm{D}(q^\theta \,\|\, P) - \log P(\mathrm{R}) = \int_{\mathrm{X}} q^\theta(\mathrm{X}) \log \frac{q^\theta(\mathrm{X})}{P(\mathrm{X},\mathrm{R})}$$

The Variational Free Energy Updates

$$F(q^{\theta},P)=\int_{\mathrm{X}} q^{\theta}(\mathrm{X})\log\frac{q^{\theta}(\mathrm{X})}{P(\mathrm{X},\mathrm{R})}=\sum_i\int_{\mathrm{X}} q_i{}^{\theta}(\mathrm{X})\log\frac{q^{\theta}(\mathrm{X})}{P(\mathrm{X},\mathrm{R})}=\sum_i <\log\frac{q^{\theta}(\mathrm{X})}{P(\mathrm{X},\mathrm{R})}>_{q_i{}^{\theta}}$$

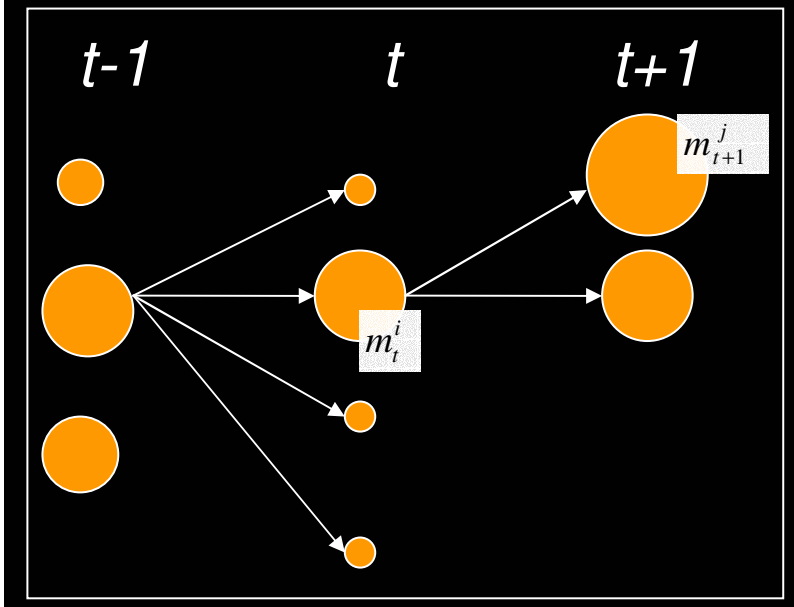
$$\frac{dF(q^{\theta},P)}{d\theta}=\sum_i\int_X q_i{}^{\theta}(\mathrm{X})g_i{}^{\theta}(\mathrm{X})\left(1+\log\frac{q^{\theta}(\mathrm{X})}{P(\mathrm{X},\mathrm{R})}\right)=\sum_i <g_i{}^{\theta}(\mathrm{X})\left(1+\log\frac{q^{\theta}(\mathrm{X})}{P(\mathrm{X},\mathrm{R})}\right)>_{q_i{}^{\theta}}$$

where

$$g_i{}^{\theta}(\mathrm{X})=\frac{d\log q_i{}^{\theta}(\mathrm{X})}{d\theta}$$

- Mixture approximation
- Unconstrained optimization, but enforce mixture constraints by reparameterization
  - Softmax for mixing proportions
  - Cholesky for covariances, positive diagonal

The Embedded Network of Observation Likelihood Peaks



The temporal observation likelihood is approximated by mixtures:

$$p(r_t \mid x_t) = \sum_{i=1}^{N_t} \pi_t^i m_t^i(x_t, \mu_t^i, \Sigma_t^i), t = 1,..,T$$

- Each m_t^i is a node in the network, with value $p(r_t \mid m_t^i)$
- The nodes connect by edges with value $p(m_{t+1}^j \mid m_t^i)$

$$p(r_t \mid m_t^i) = \int_{x_t} m_t^i(x_t) p(r_t \mid x_t)$$

$$p(m_1^i) = \int_{x_1} m_1^i(x_1) p(x_1)$$

$$p(m_{t+1}^j \mid m_t^i) = \int_{x_{t+1}} \int_{x_t} m_{t+1}^j(x_{t+1}) m_t^i(x_t) p(x_{t+1} \mid x_t)$$

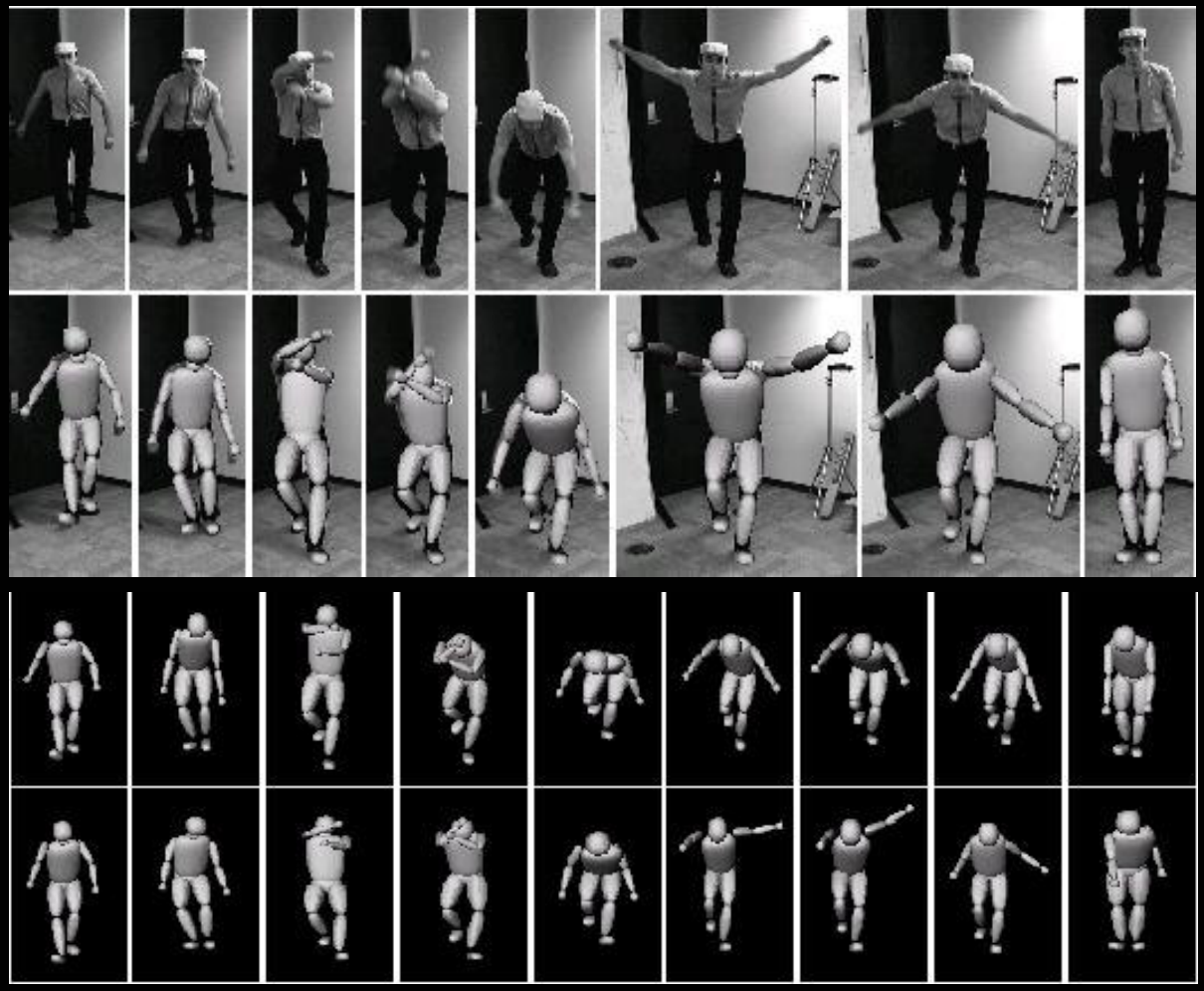
- The Embedded Network is an approximation to the trajectory distribution. It is not a HMM
  - Different number of states per timestep depending on the uncertainty of the observation likelihood
  - Different continuous values for state variables

The Variational Mixture Smoothing Algorithm (VMS)

- **Input:** Filtered mixture distributions at each timestep
- **Output:** Mixture approximation to $P(X|R)$
  - (a) Construct Embedded Network of likelihood peaks
    - Lift the dynamics and the observation likelihood from point-wise (as defined in a non-linear dynamical system) to mode-wise
    - Find probable trajectories between initial and final mode pairs
  - (b) Refine non-linearly (smoothing)
    - Initialize using (a)
    - Mixture of MAP estimates (means + inv. Hessians at maxima)
  - (c) Variationally optimize a mixture approximation to $P(X|R)$
    - Initialize using (b)

Variational Mixture Smoothing

Sminchisescu & Jepson, CVPR2004



2 (out of several) plausible trajectories

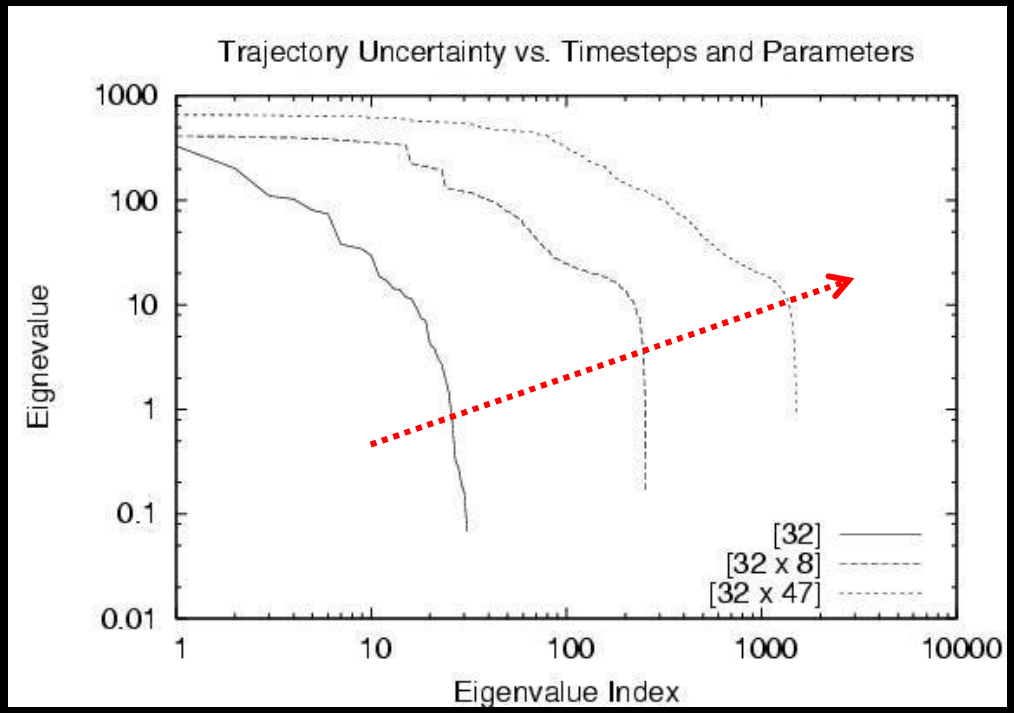
Variational Mixture Smoothing <v>

Sminchisescu & Jepson, CVPR2004

Model / image Filtered Smoothed

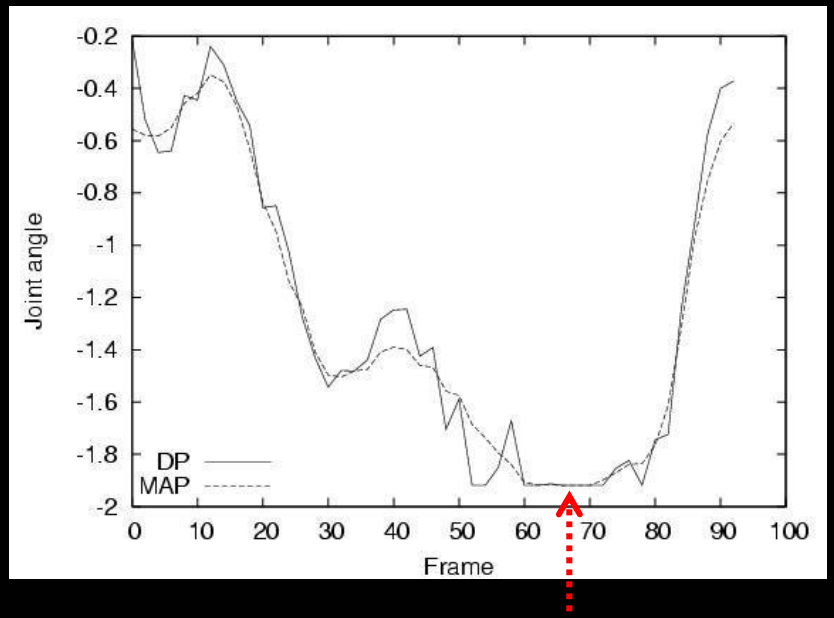
2 (out of several) plausible trajectories

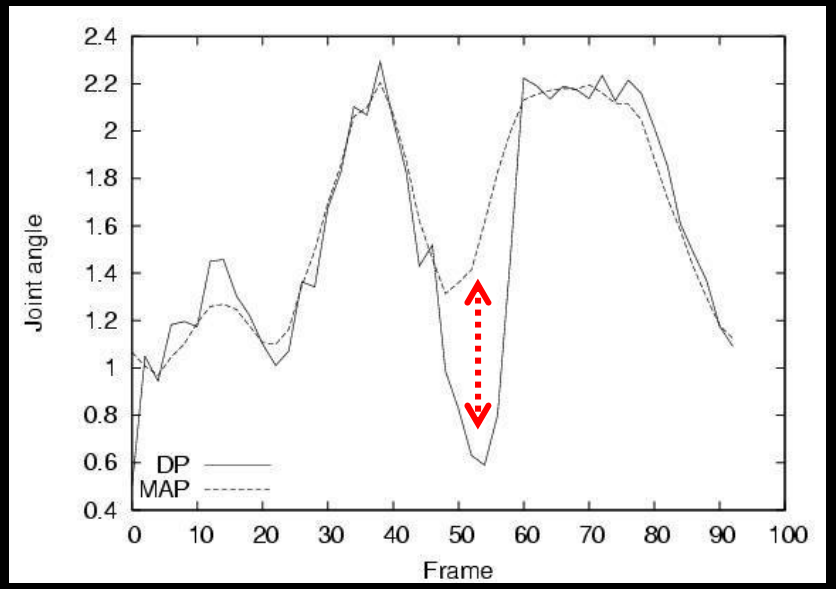
Smoothing over long sequences reduces the local optima uncertainty



- Covariance eigenspectrum for trajectories having 32, 256 and 1504 variables
  - Largest / smallest eigvalue ratio: 5616 (1 frame) , 2637 (8), 737 (47)
- Notwithstanding larger dimensionality, estimation based on longer sequences is better constrained

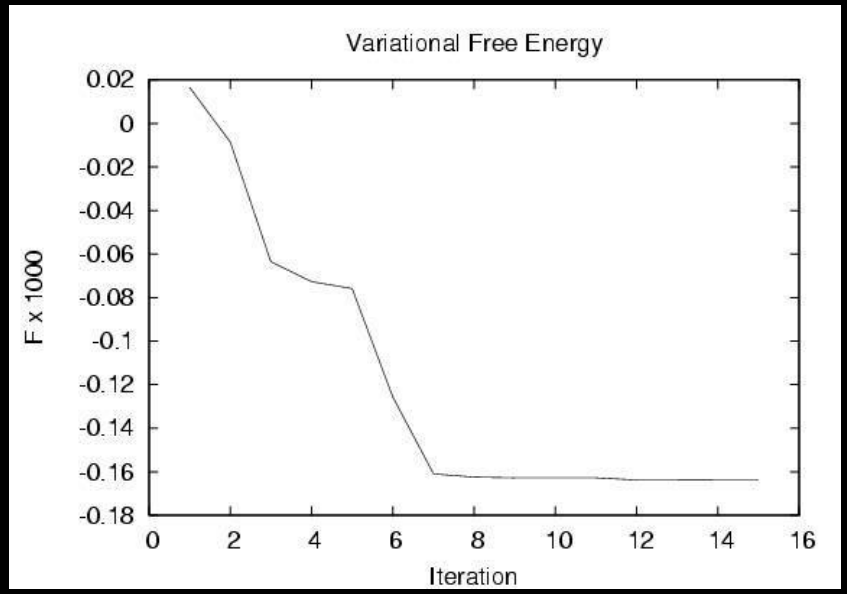
MAP Smoothing





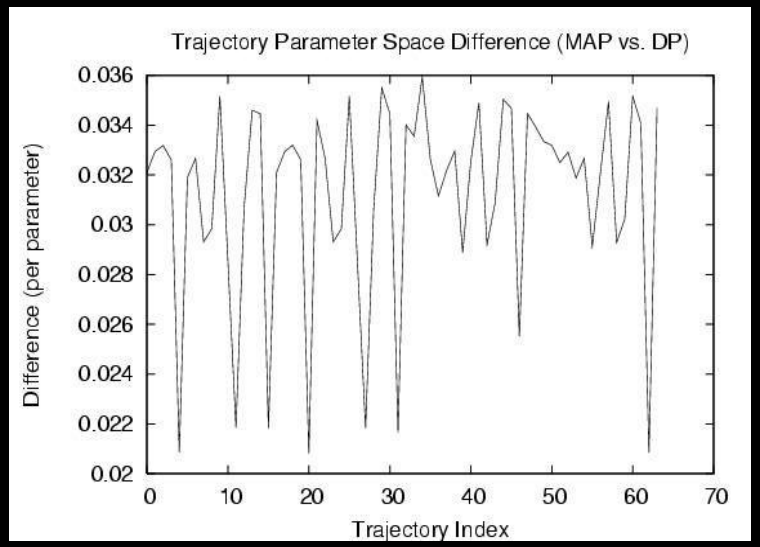
- Correctly preserves physical constraints (left)
- Discovers new configurations (*e.g.* frames 50-60, right)
  - Trajectory differences concentrated in a few temporal states

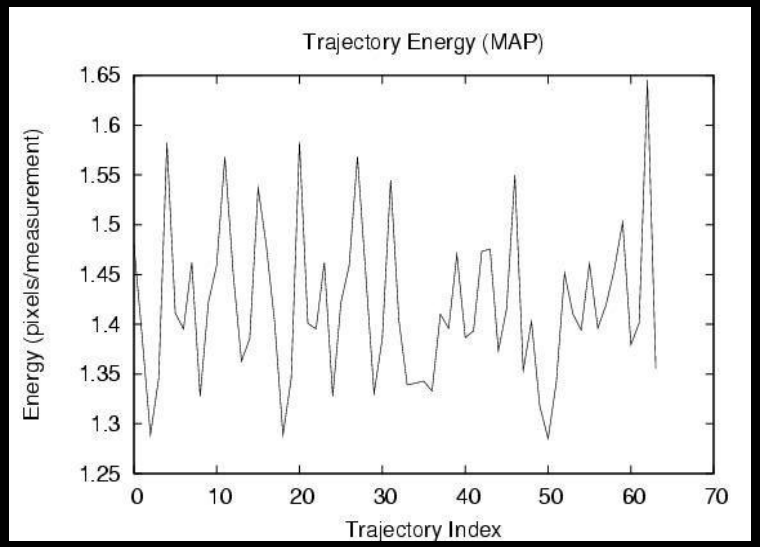
Free Energy Minimization



- Mixture approximation initialized from MAP (maxima and inverse Hessians)
- Estimate the mixing proportions, means and covariance inflation factors
- Plateau after 6-8 refinement steps
- Good approximation for ill conditioned vision problems: sharp priors + uncertain likelihoods => MAPs that underestimate surrounding volumes
- For peak emphasis, optimize parameter subsets (e.g. with means fixed)

Trajectory Statistics
(state space measured in *radians*, energy measured in *pixels*)





- Average change *per state variable* 2-3 degrees
  - but changes are concentrated in only a few variables
  - the qualitative difference between trajectories is often quite large

- Low energy (pixel error, right) shows that smoothing preseves the model-image matching quality

So what can we say now?

- We can live with this, extensive search and lots of ambiguity is part of life
- We are definitely on the wrong path
- We are not yet done
  - Aha …

Teddies appalled by inference.

Probability and Statistics

- *Probability:* inferring probabilistic quantities for the data given fixed models as before…

- *Statistics:* inferring a model given fixed data observations
  - Do we have the best models for probability calculations?

Models = representation + parameters

We will learn both

- supervised and unsupervised procedures

The Need for Representation Learning

- The unconstrained generative human model is useful for tracking unexpected general motions
- But high-dimensional inference is expensive
- Multiple trajectories consistent with the image evidence
  - Human perception less ambiguous due to better `modeling', priors, *etc*
- Missing data (occlusion of limbs) difficult to deal with

➢ Many human motions may be repetitive or low-dimensional
  - *e.g.* running, walking, human conversations

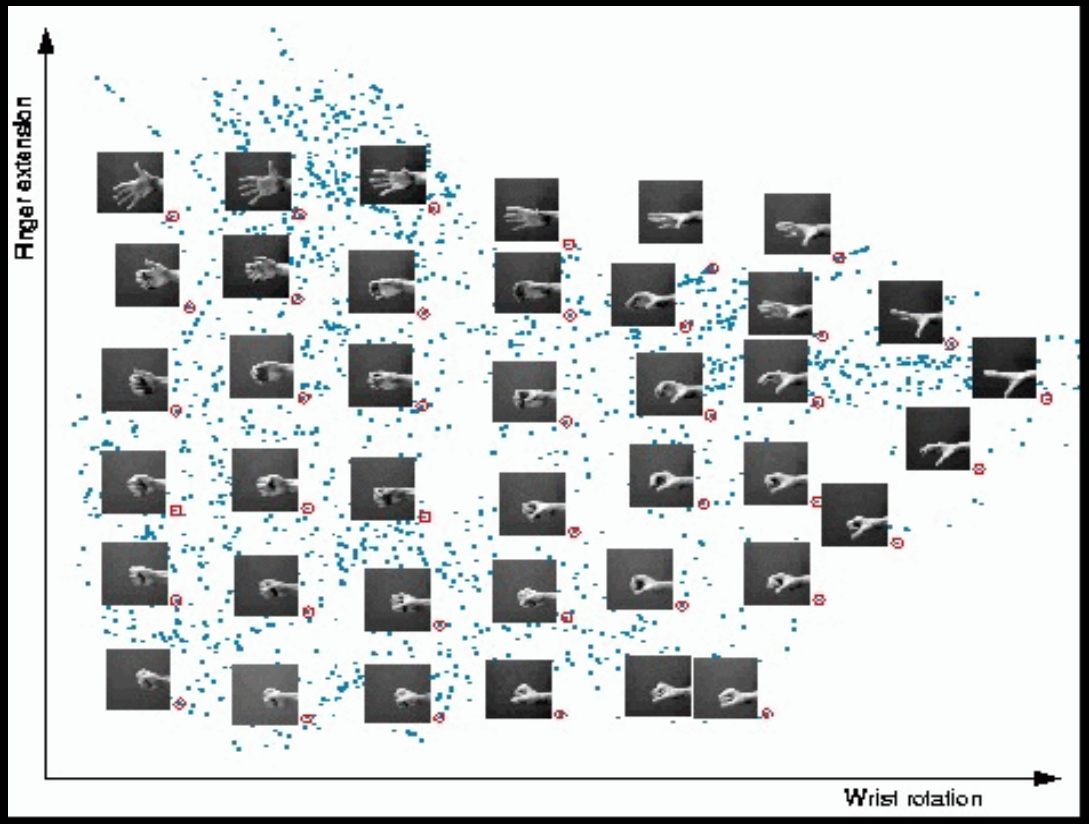
➢ Need a better state representation (prior)
  - Learn correlations between variables, lower the dimension

Presentation Plan

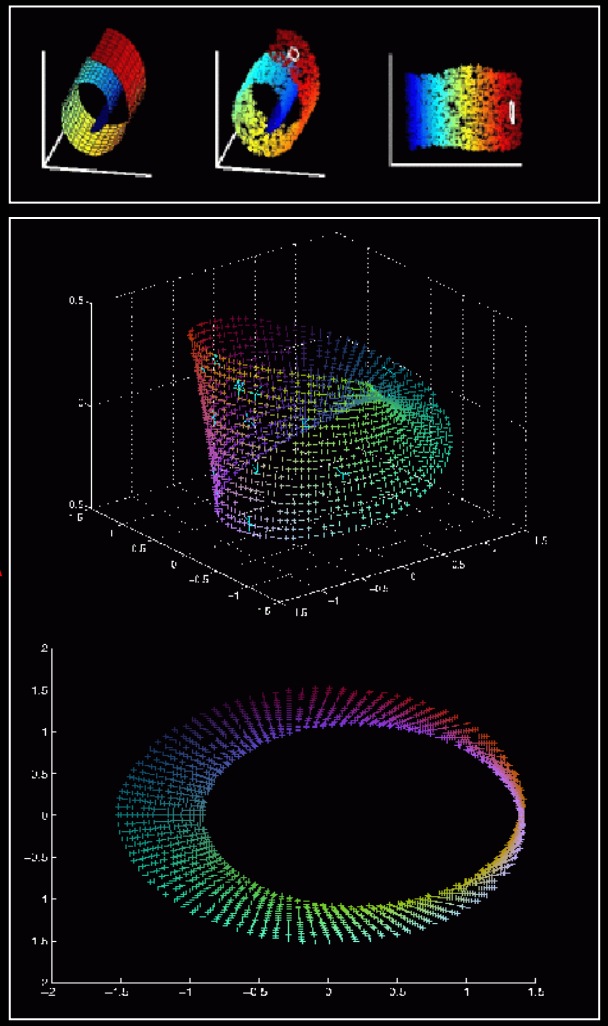
- Introduction, history, applications
- State of the art for 2d and 3d, human detection, initialization
- 3D human modeling, generative and discriminative computations
- Generative Models
  - Parameterization, shape, constraints, priors
  - Observation likelihood and dynamics
  - Inference algorithms
  - Learning non-linear low-dimensional representations and parameters
- Conditional (discriminative) models
  - Probabilistic modeling of complex inverse mappings
  - Observation modeling
  - Discriminative density propagation
  - Inference in latent, kernel-induced non-linear state spaces
- Conclusions and perspectives

From Images to Hidden Causes

- The true few hidden causes for data variability are embedded in the high-dimensional ambient space of images



(Tenenbaum et al '01)



Low-dimensional non-linear embedding preserves original high-dimensional manifold geometry

Manifolds?

The topology of an intrinsic low-dimensional manifold for some physical process ultimately depends on what is observed (*e.g.* joint angles *vs.* image features, *etc*)

- *e.g.* a manifold based on image silhouettes of humans walking parallel to camera plane self-intersects at half-cycles whereas one based on human 3d joint angle does not (more generally, the latter representation is viewpoint invariant whereas the former is not)

Issues:

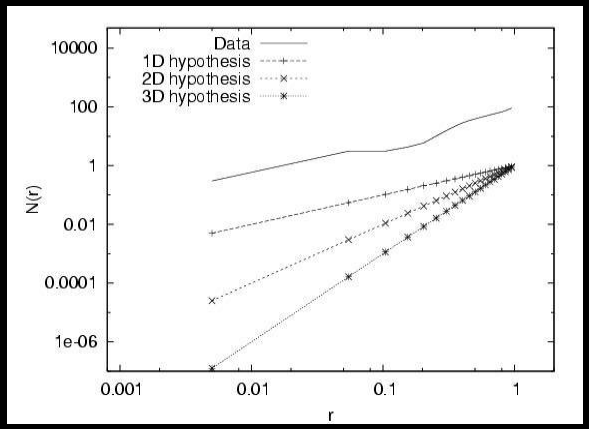
- How to estimate the intrinsic dimensionality
- How to handle non-linearity and physical constraints
- How to efficiently use the intrinsic (latent) space for inference
  - Need a global latent coordinate system and a continuous low-dimensional model (*i.e.* priors and observation likelihood)
    - for *e.g.* non-linear optimization, hybrid MCMC sampling)

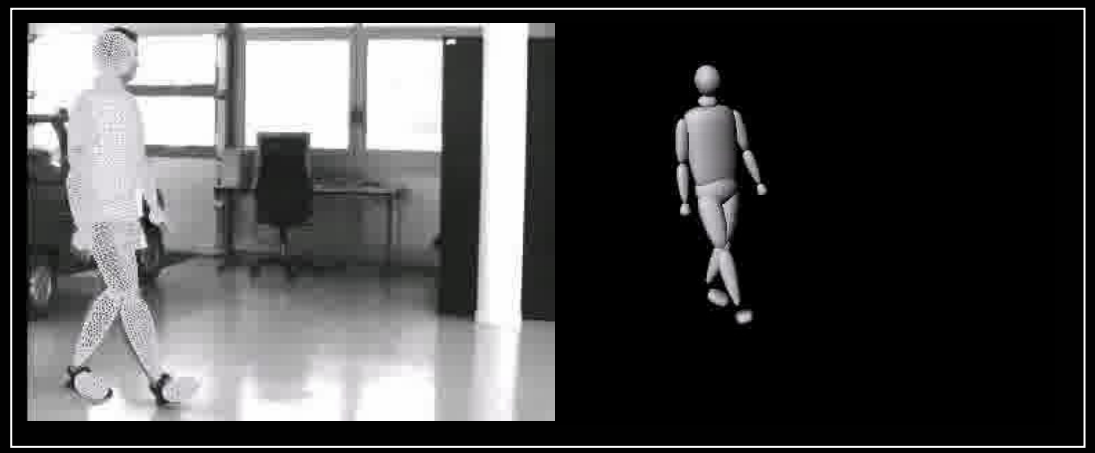
Intrinsic Dimension Estimation <v> and Latent Representation for Walking

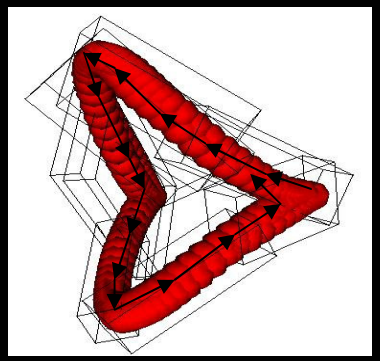
- 2500 samples from motion capture
- The Hausdorff dimension (*d*) is effectively 1, lift to 3 for more flexibility
- Use non-linear embedding to learn the latent 3d space embedded in an ambient 30d human joint angle space

Intrinsic dimension estimation

$$d = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)}$$



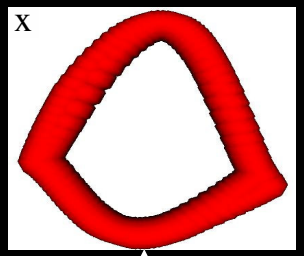


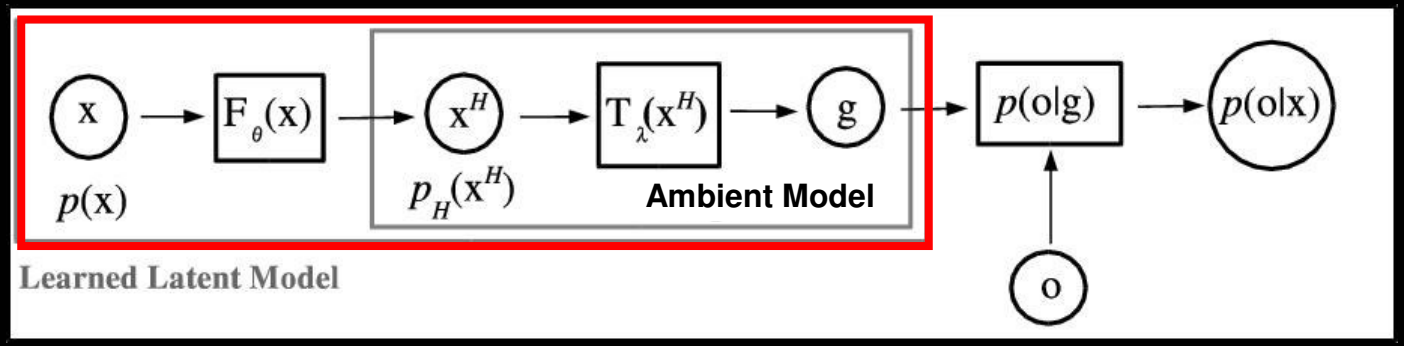


3d latent space

Learning a Latent Variable Generative Model

Sminchisescu & Jepson, ICML 2004





$$\text{Manifold prior } p(x) \propto p_M(x) \cdot p_H(F_\theta(x)) \, |J_{F_\theta}{}^T J_{F_\theta}|^{1/2}$$

- Global representation
  - Obtained with non-linear dimensionality reduction
  - To support intrinsic curvature, use *e.g.* Laplacian Eigenmaps
- Continuous generative mapping: latent -> ambient space
  - Use sparse kernel regression (simple map due to embedding) $F_\theta(x)$
- Consistent latent prior combines
  - Ambient priors (back-transferred to latent space) $p_H(F_\theta(x)) \, |J_{F_\theta}{}^T J_{F_\theta}|^{1/2}$
  - Training-data density in latent space (mixture) $p_M(x)$

Visual Inference in a 12d Space

6d rigid motion + 6d learned latent coordinate

Interpretation #1
Says `bye' when conversation ends (before the turn)

Interpretation #2
Points at camera when conversation ends (before the turn)

Visual Inference in a 12d Space <v>

6d rigid motion + 6d learned latent coordinate

Interpretation #1

Says `bye' when conversation ends

(before the turn)

Interpretation #2

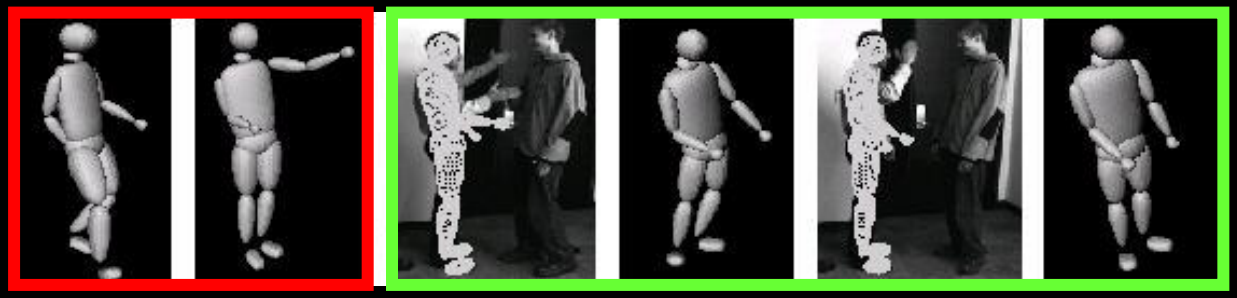
Points at camera when conversation ends

(before the turn)

- 9000 samples from running, walking, conversations (29d)
- Learning correlations between state variables is useful during occlusion
  - Occlusion would otherwise lead to singularities or arbitrary `default' behavior
- The state distribution is still multimodal. Need inference based on MH
  - Branching at activity switch or inside (*e.g.* right arm during conversation)
  - Image matching ambiguities

Learning Representations

Caveats and Pitfalls



No transferred ambient physical priors

Only trained with running data

- More efficient to compute but need good modeling
- Need statistically representative training set
- Need effective ways to enforce constraints
  - ➢ Learned representations could lack physical interpretability

Learning Parameters

Random Fields

Sminchisescu, Welling and Hinton '03, '05

$$p_\theta(\mathrm{X} \mid \mathrm{R}) = \frac{1}{Z_\theta(\mathrm{R})} \exp\left[-E_\theta(\mathrm{X}, \mathrm{R})\right]$$

$$Z_\theta(\mathrm{R}) = \int_{\mathrm{X}} \exp\left[-E_\theta(\mathrm{X}, \mathrm{R})\right]$$

- X is the joint state vector at all T timesteps $\mathrm{X} = (x_1, x_2, ..., x_T)$, a trajectory
- R is the joint observation vector $\mathrm{R} = (\mathrm{r}_1, \mathrm{r}_2, ..., \mathrm{r}_T)$
- θ is a vector of parameters (e.g. body shape, observation likelihood, dynamics)
- $E_\theta(\mathrm{X}, \mathrm{R})$ is a sum of energy functions of neaby temporal states and observations, e.g. unormalized `dynamics' and `observation' functions d_θ and l_θ:

$$E_\theta(\mathrm{X}, \mathrm{R}) = \sum_{t=1}^{T} \left[l_\theta(\mathrm{x}_t, \mathrm{r}_t) + d_\theta(\mathrm{x}_t, \mathrm{x}_{t-1}) \right]$$

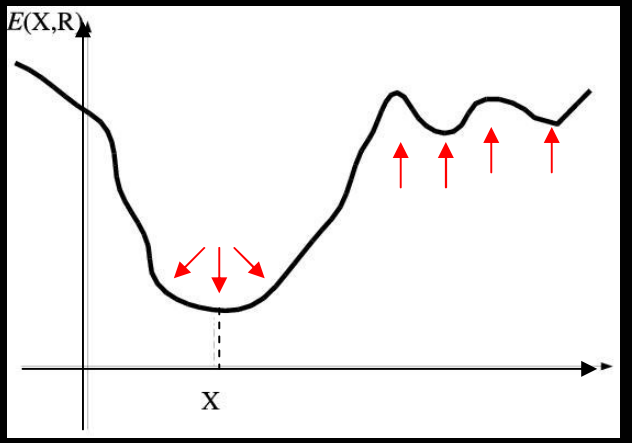
- Instead of learning pieces off-line, do end-to-end sequence training in a *globally normalized* model
- Alternatively, can learn the dynamics and the observation likelihood separately, and work by *locally normalizing*
  - Effective (due to smaller problems) but possibly suboptimal

Learning the model parameters
Optimize the conditional log-likelihood

$$L_\theta = -\sum_{d=1}^{D} \log\ p_\theta(\mathrm{X} \mid \mathrm{R}) = \sum_{d=1}^{D} \left(E_\theta(\mathrm{X}, \mathrm{R}^d) + \log Z_\theta(\mathrm{R}^d) \right)$$

$$\frac{dL_\theta}{d\theta} = \sum_{d=1}^{D} \left(\frac{dE_\theta(\mathrm{X}^d, \mathrm{R}^d)}{d\theta} - \int_{\mathrm{X}} p_\theta(\mathrm{X} \mid \mathrm{R}^d) \frac{dE_\theta(\mathrm{X}, \mathrm{R}^d)}{d\theta} \right) = \left\langle \frac{dE_\theta}{d\theta} \right\rangle_{\text{data}} - \left\langle \frac{dE_\theta}{d\theta} \right\rangle_{\text{model}}$$

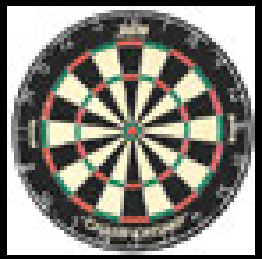
- The gradient involves `contrastive terms'
  - Difference between averages over the data and the model distribution
  - Essentially a `learning by inference' loop
  - Educated version of parameter hand-tuning
- The normalization enforces discrimination: making the true response likely effectively makes competing, incorrect interpretations unlikely

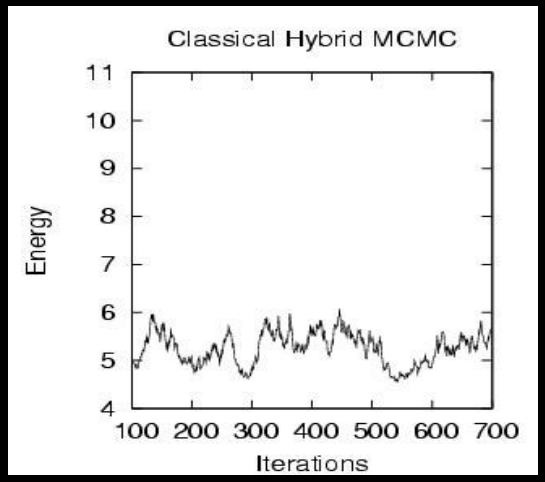


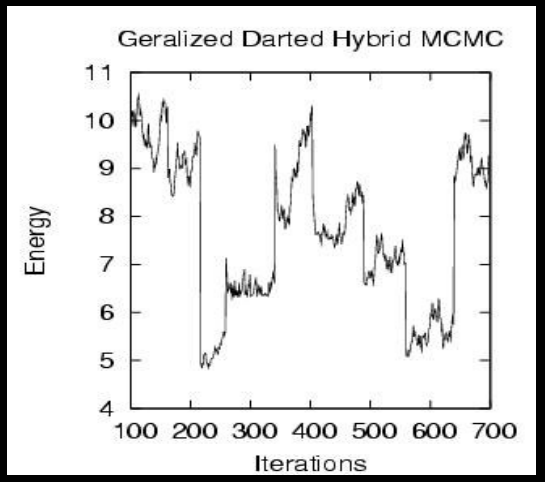
True solution

Computing Partition Functions
Generalized Darting

Sminchisescu, Welling and Hinton '03, Sminchisescu & Welling'05

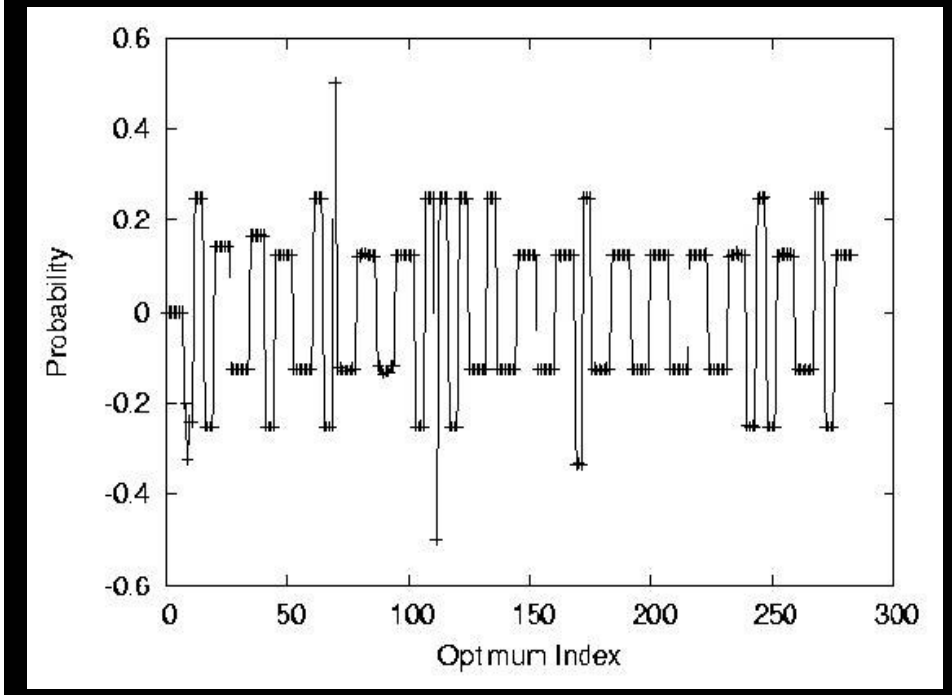




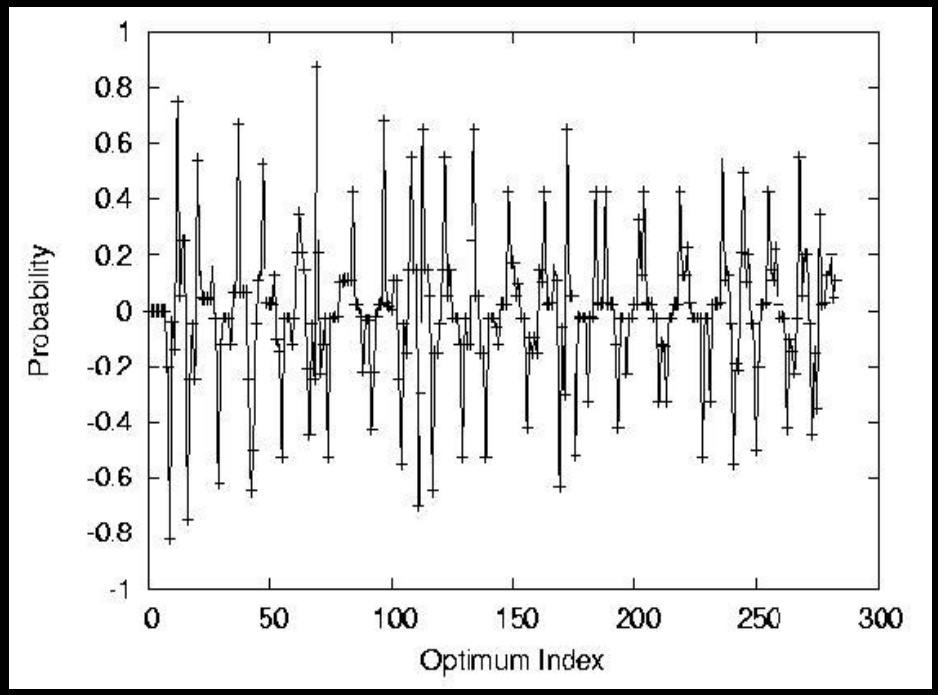


- A fast, auxiliary variable method for sampling an equilibrium distribution given knowledge of its peaks
  - e.g statically using ET / HS / KJS or dynamically using VMS
- Combines short-term moves (classical MC) with long-range jumps between peaks in a way that obeys detailed balance
  - A complicated constraint to fulfill
  - Ad-hoc procedures are typically incorrect because they fail to precisely account for the local volumes when computing acceptance probabilities

The effect of learning on the trajectory distribution



Before



After

- Learn body proportions + parameters of the observation model (weighting of different feature types, variances, *etc*)
- Notice reduction in uncertainty
- The ambiguity diminishes significantly but does not disappear

Generative Modeling Summary

- Useful for tracking complex motions
- Components can be learned (prior distributions, compact representations, parameters of the observation model, *etc*)
  - Supervised and unsupervised procedures
- Learning improves the model, allowing more confident percepts
  - Inference is complex (expensive) but essential
  - Need gradient information to optimize high-dimensional models
  - Need compact mixture approximations
- The entropy of the state posterior after learning reflects the fundamental limits of modeling and gives good intuition about run-time speed and accuracy

Can we use models that are easier to infer and learn?
Do we need to generate the image or rather to condition on it?

"Thanks for choosing Quality Courier."

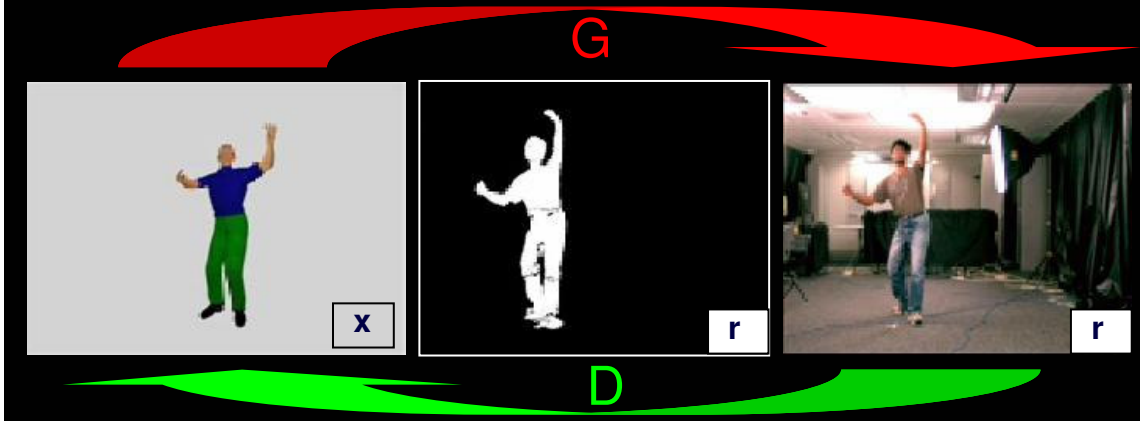
Presentation Plan

- Introduction, history, applications
- State of the art for 2d and 3d, human detection, initialization
- 3D human modeling, generative and discriminative computations
- Generative Models
  - Parameterization, shape, constraints, priors
  - Observation likelihood and dynamics
  - Inference algorithms
  - Learning non-linear low-dimensional representations and parameters
- Conditional (discriminative) models
  - Probabilistic modeling of complex inverse mappings
  - Observation modeling
  - Discriminative density propagation
  - Inference in latent, kernel-induced non-linear state spaces
- Conclusions and perspectives

Rationale for Discriminative Modeling

- For humans, it seems much easier to recognize a body posture than to draw it
- Some of the pose recognition computations in the brain appear to be dominantly feed-forward and obey stringent time constraints
- The above considerations do not rule out visual feedback (analysis by synthesis, generative modeling) but question its optimal placement in a robust artificial system for pose perception
- This part of the lecture describes a complementary feed-forward, bottom-up, probabilistic discriminative approach

Generative vs. Discriminative Modelling



x is the model state

r are image observations

Goal: $p_\theta(\mathbf{x}\,|\,\mathbf{r})$

θ are parameters to learn given training set of (**r**,**x**) pairs

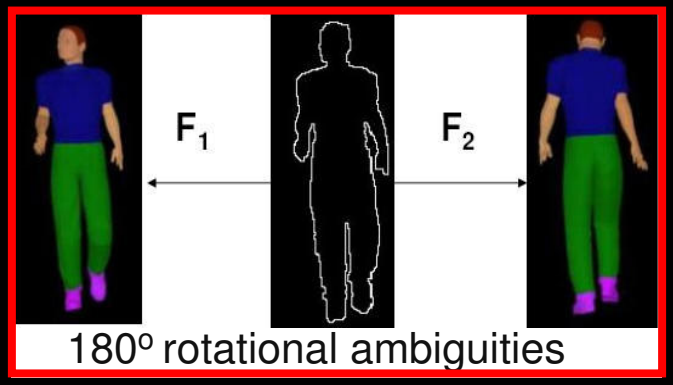
$$p_\theta(\mathbf{x}\,|\,\mathbf{r}) \;\alpha\; p_\theta(\mathbf{r}\,|\,\mathbf{x})\cdot p(\mathbf{x})$$

- Learning to `invert' perspective projection and kinematics is difficult and produces multiple solutions
  - *Multivalued mappings* ≡ *multimodal conditional state distributions*
- Probabilistic temporal framework lacking until now
  - What distributions to model?
  - Which propagation rules to use?

- Learn
  - State representations and priors
  - Observation likelihood; but difficult to model human appearance
  - Temporal dynamics
- Sound probabilistic framework
  - Mixture or particle filters
  - State inference is expensive, need powerful inference methods

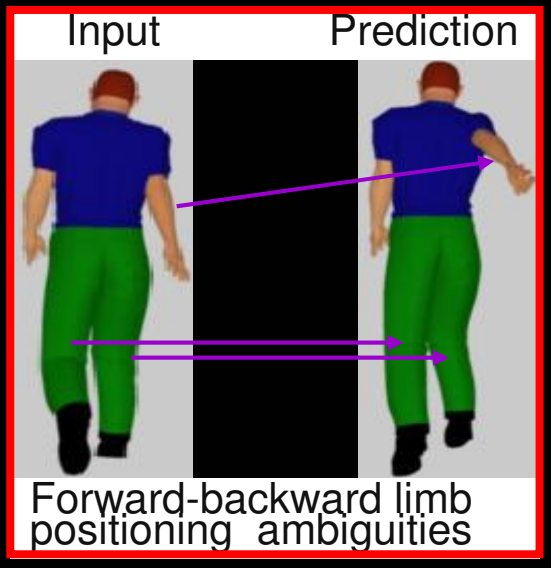
`Discriminative' Difficulties in Inverting the Monocular Projection and the Human Kinematics

Some issues persist, disregarding the approach



180º rotational ambiguities

Leg assignment ambiguities



Forward-backward limb positioning ambiguities

Conditional Visual Inference

Discriminative Density Propagation

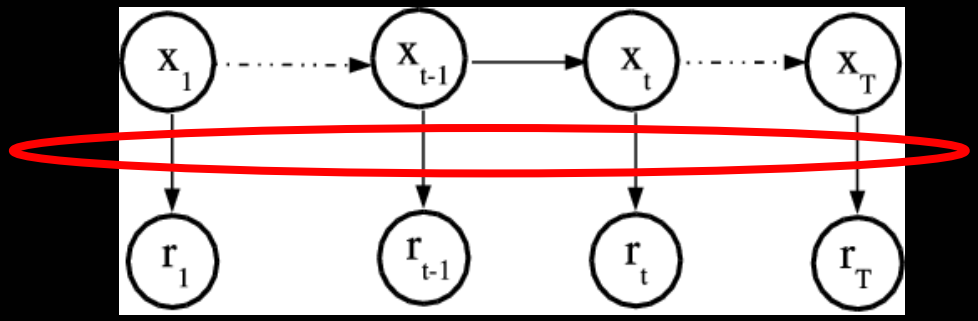
Sminchisescu, Kanaujia, Li, Metaxas, CVPR 2005

Key Aspects

- How to formalize the problem
  - Structure of the graphical model, the state and observation
  - Temporal propagation rules
- How to accurately model complex multi-valued, observation-to-state (inverse) mappings
  - Correctly reflect contextual dependencies
- How to do computation compactly and efficiently
  - Sparsity, lower-dimensionality, mixtures, *etc*

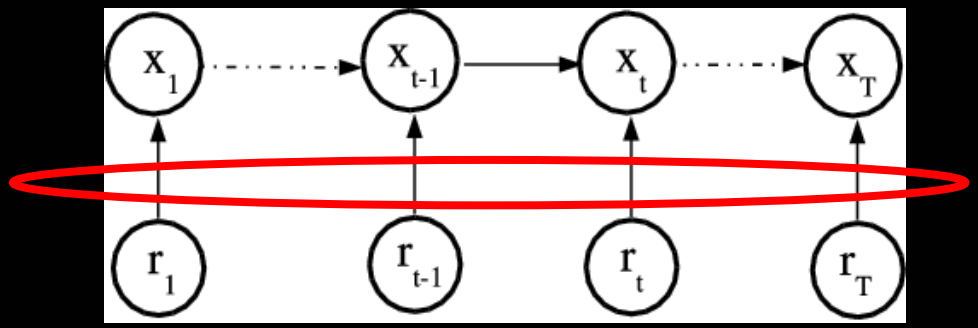
Temporal Inference (tracking)

- Generative (top-down) chain models



← *Models* the observation

- Discriminative (bottom-up) chain models



← *Conditions on* the observation

Image Observations

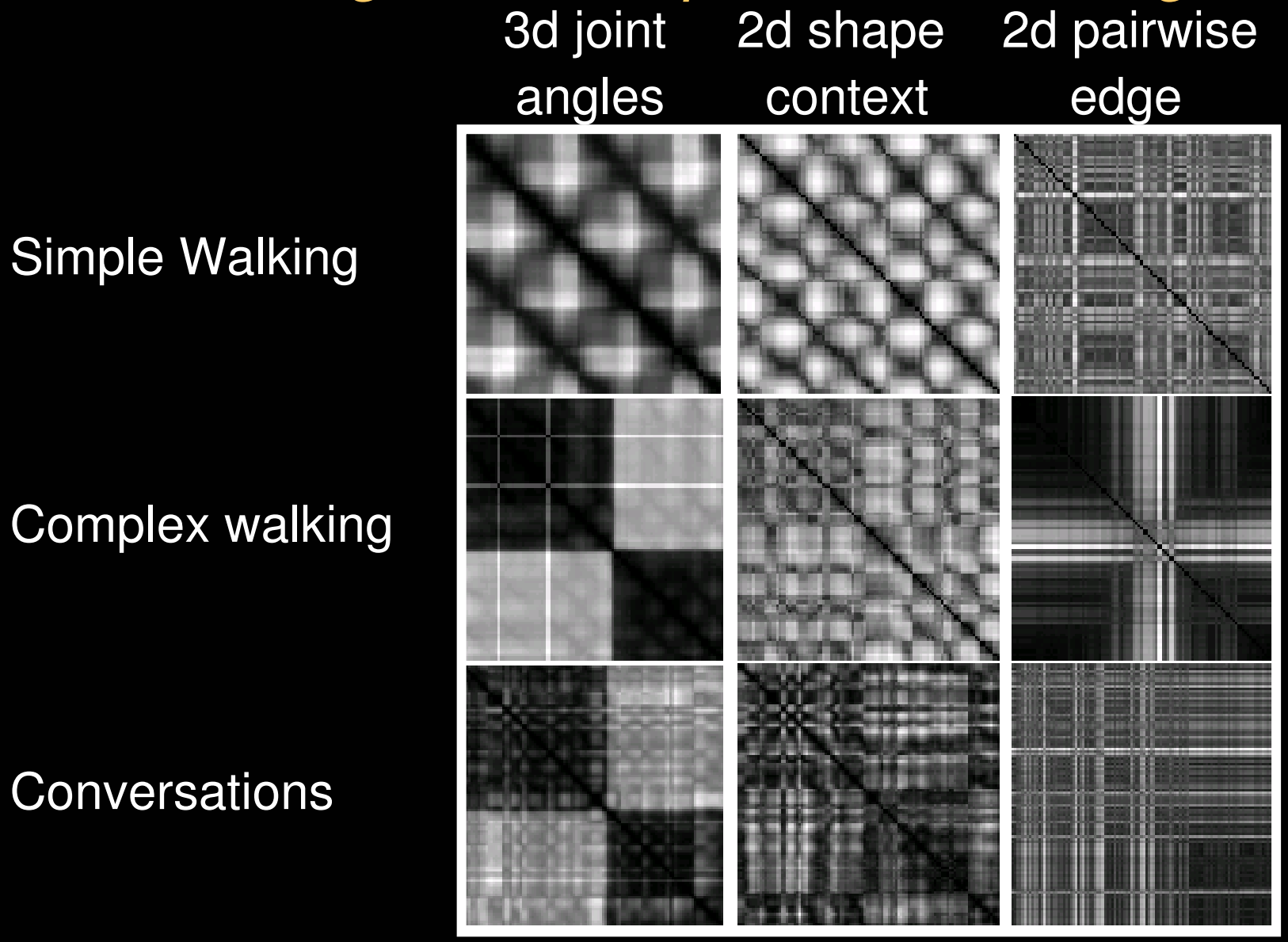
Histograms of silhouettes with internal edges

- Silhouettes capture the essential pose information
  - ▲ Low-level, arguably extractable from image sequences
    - e.g. using the 2d detectors reviewed in the state-of-the art
  - ▲ Insensitive to surface attributes: texture, color, clothing
  - ▼ Internal details may be spurious due to clothing folds
  - ▼ Often distorted by poor foreground segmentation, attached shadows

- Histogram representation (*cf.* approaches in object recognition)
  - Sample edge points on the silhouette
  - Compute local shape context (SC) and pairwise edge orientation (EO)
  - Obtain generic codebook by clustering in the SC and EO space
  - Represent each new silhouette *w.r.t* to codebook
    - vector quantize to obtain histogram

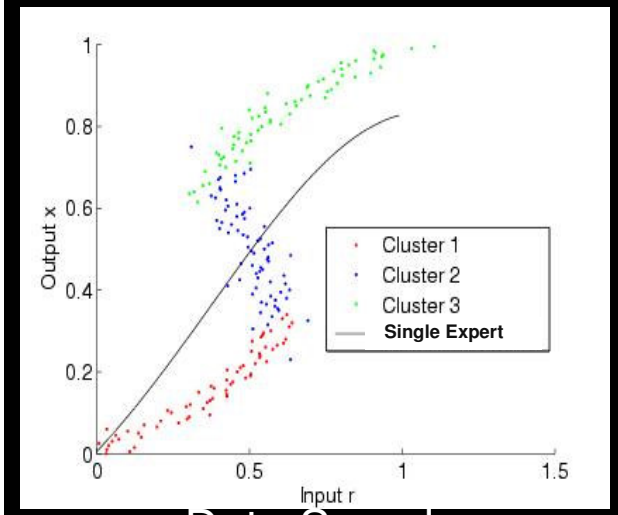
Image Features (affinity matrices)

Pairwise edge and shape context histograms

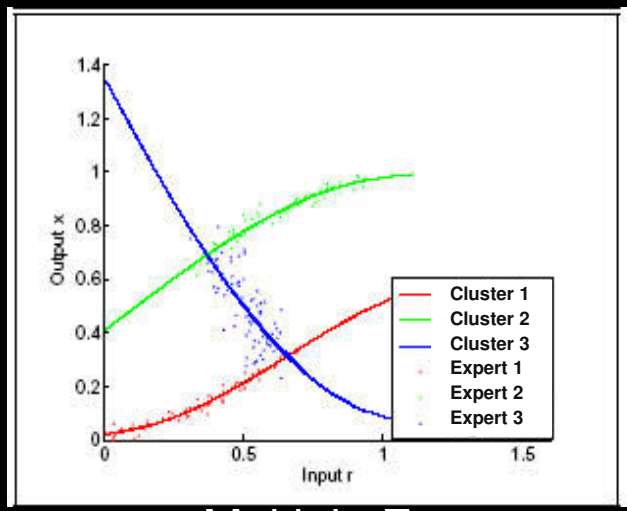


Modeling Multivalued Inverse Mappings

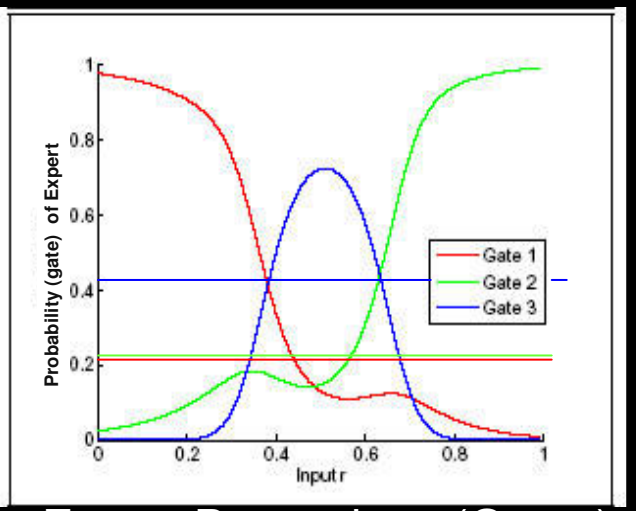
Bayesian ***Conditional*** Mixtures of Experts



Data Sample



Multiple Experts



Expert Proportions (Gates)
vs. uniform coefficients (Joint)

- A single expert cannot represent multi-valued relations
- Multiple experts can focus on representing parts of the data
  - Constrained clustering problem under the assumed input – output (state - observation) dependency, *not* the Euclidean similarity between datapoints
- But the expert contribution (importance) is contextual
  - Disregarding context introduces systematic error (invalid extrapolation)
  - Essential to work with the conditional and not with the joint distribution
- Therefore, the experts need input dependent mixing proportions

Bayesian ***Conditional*** Mixtures of Experts

Parameters

$\boldsymbol{\theta} = (\mathbf{W}, \boldsymbol{\Sigma}, \boldsymbol{\delta})$

$$p(\mathbf{x} \mid \mathbf{r}, \mathbf{W}, \boldsymbol{\Sigma}, \boldsymbol{\delta}) = \sum_{i=1}^{M} g(\mathbf{r} \mid \boldsymbol{\delta}_i) \cdot p(\mathbf{x} \mid \mathbf{r}, \mathbf{W}_i, \boldsymbol{\Sigma}_i)$$

$$g(\mathbf{r} \mid \boldsymbol{\delta}_i) = \frac{\exp(\boldsymbol{\delta}_i^{\mathbf{T}} \mathbf{r})}{\sum_{k=1}^{M} \exp(\boldsymbol{\delta}_k^{\mathbf{T}} \mathbf{r})}$$

$$p(\mathbf{x} \mid \mathbf{r}, \mathbf{W}_i, \boldsymbol{\Sigma}_i) \sim N(\mathbf{x} \mid \mathbf{W}_i \Phi(\mathbf{r}), \boldsymbol{\Sigma}_i)$$

Mixture of experts

Gating function - confidence in expert ***i***, given input **r**

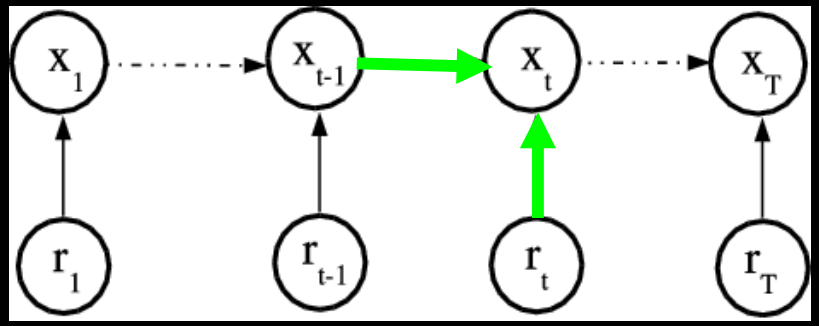
Expert ***i*** is e.g. a kernel regressor with weight W_i

- ARD (automatic relevance determination) model
  - Sparse solutions, empirically observed to avoid overfitting
- Estimate parameters θ, using double loop EM
  - Learn the experts *and* how to predict their gates
  - Maximum-likelihood type II approximations

Discriminative Temporal Graphical Model

Discriminative Temporal Inference

- `Bottom-up' chain



Conditions on the observation

$$p(\mathbf{x}_t \mid \mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{r}_t)\, p(\mathbf{x}_{t-1} \mid \mathbf{R}_{t-1}), \quad where\ \mathbf{R}_t = (\mathbf{r}_1, ..., \mathbf{r}_t)$$

Local conditional

Temporal (filtered) prior

- The *temporal prior* is represented as a Gaussian mixture
- The *local conditional* is a Bayesian mixture of Gaussian experts
- For robustness, include a `memoryless' importance sampler
  - *i.e.* $p(x_t|r_t)$ also used for initialization
- Integrate pair-wise products of Gaussians analytically
  - Linearize (possibly non-linear) kernel experts
  - The mixture grows exponentially, prune at each timestep

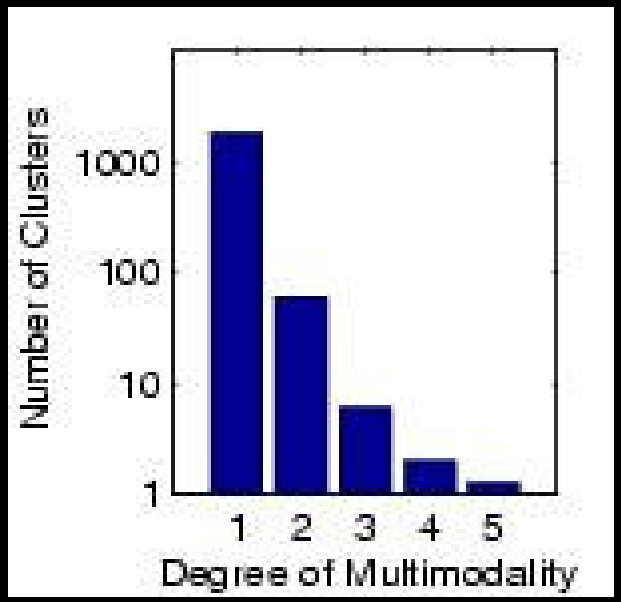
Experimental Setting

- Motion capture data from the CMU database
- Animate a computer graphics (CG) human model
- Generate training pairs of *(state, observation)=(joint angles, silhouettes)* using CG rendering (Maya)
- **State** = 56 joint angles
- **Observation** = 60d silhouette shape context + internal edge histogram descriptor
- Silhouettes computed using statistical background subtraction (courtesy of A. Elgammal)

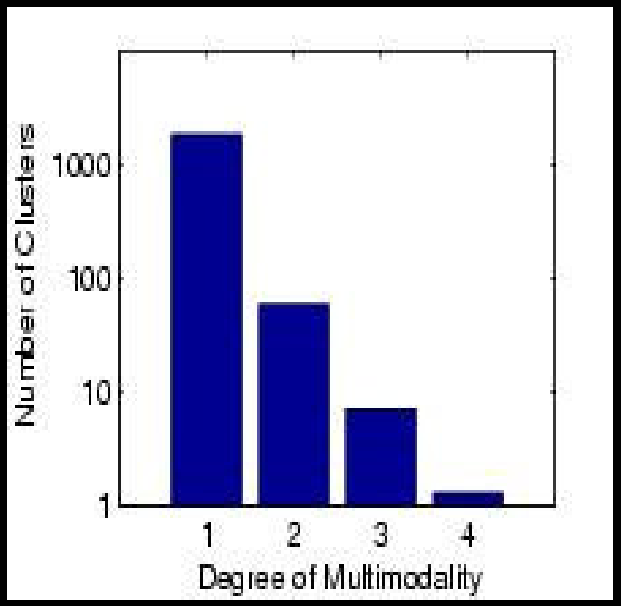
- Learn local conditionals with 5 experts, 5-15% sparsity
  - Learn predictors for each state dimension, independently
- Temporal distribution also represented using 5 components

Analysis of the Training Set

$p(\mathbf{x}_t \mid \mathbf{r}_t)$



observation vs. *state*



$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{r}_t)$

(current observation, previous state) vs. *current state*

- Histograms obtained by clustering independently in the state and observation space, count state clusters within each observation cluster and histogram the values
- Multimodality is not wild but significant for both predictors
  - Likely to increase with the training set

Results on artificially generated silhouettes with 3d ground truth

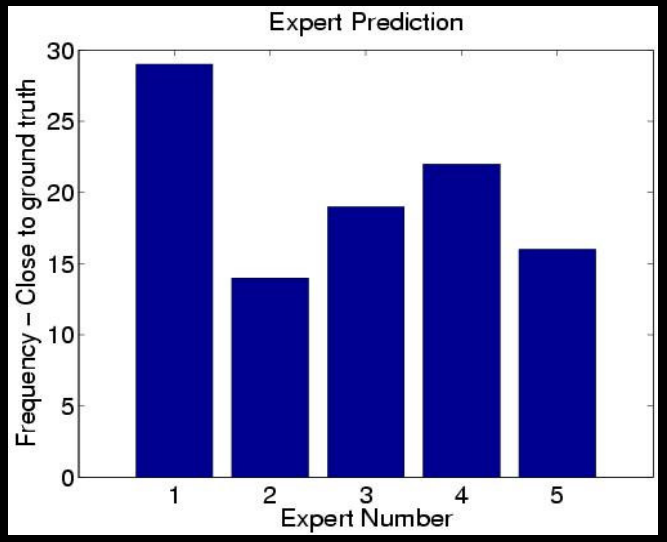
(average error / average maximum error, per joint angle)

Sequence	$p(\mathbf{x}_t \mid \mathbf{r}_t)$			$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{r}_t)$		
	NN	RVM	BME	NN	RVM	BME
NORMAL WALK	4 / 20	2.7 / 12	2 / 10	7 / 25	3.7 / 11.2	2.8 / 8.1
COMPLEX WALK	11.3 / 88	9.5 / 60	4.5 / 20	7.5 / 78	5.67 / 20	2.77 / 9
RUNNING	7 / 91	6.5 / 84	5 / 94	5.5 / 91	5.1 / 108	4.5 / 76
CONVERSATION	7.3 / 26	5.5 / 21	4.15 / 9.5	8.14 / 29	4.07 / 16	3 / 9
PANTOMIME	7 / 36	7.5 / 53	6.5 / 25	7.5 / 49	7.5 / 43	7 / 41

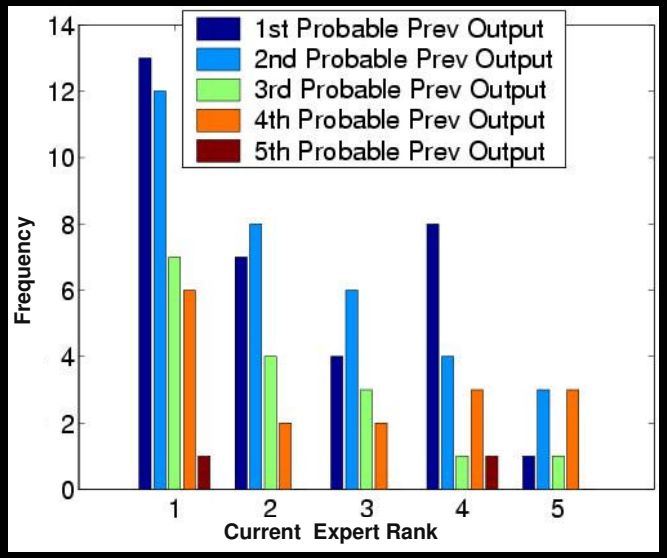
- Notice smaller maximum error for BME

Prediction Quality and Peak Dynamics

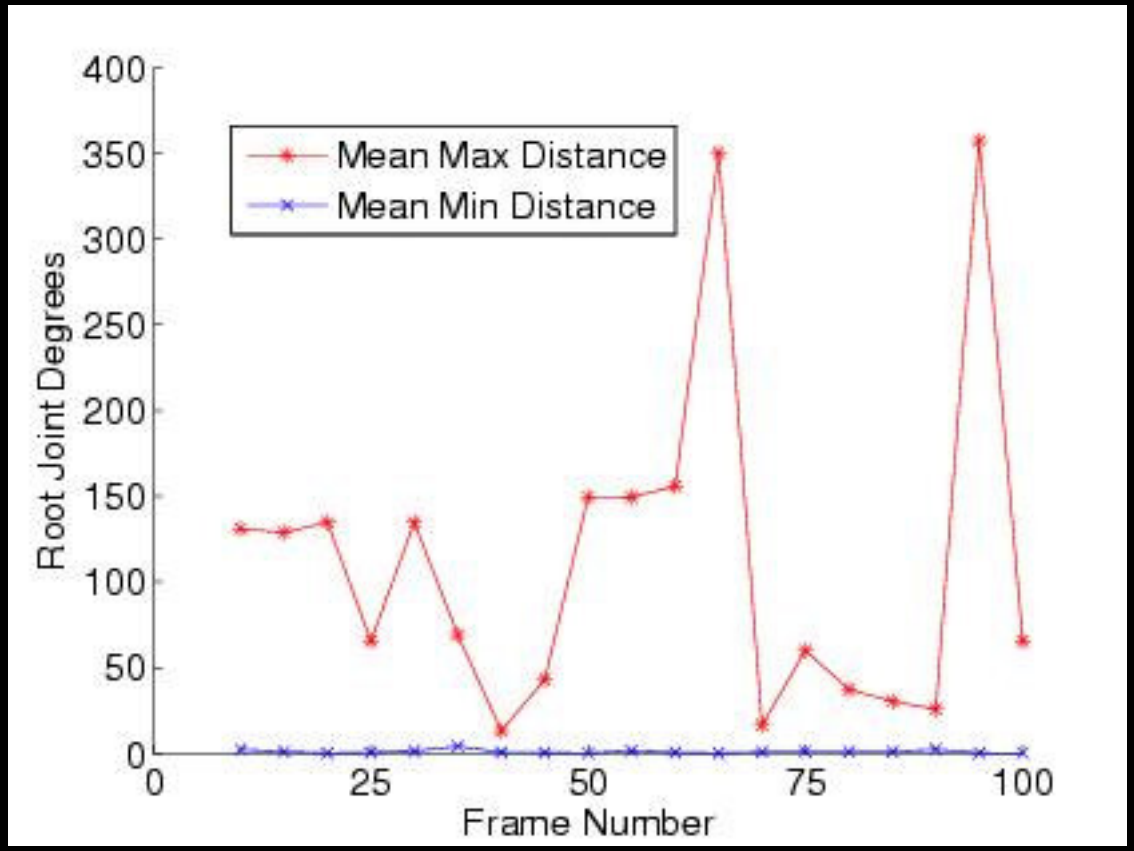
- The most probable model prediction is not always the most accurate one
  - But usually the correct solution is among the ones predicted
  - In a multiple hypotheses framework this `approximately correct' behavior enables recovery despite transient failures



- The `peak dynamics' shows that the top most probable experts tend to be more stable at preserving their rank across successive timesteps, compared to the less probable ones

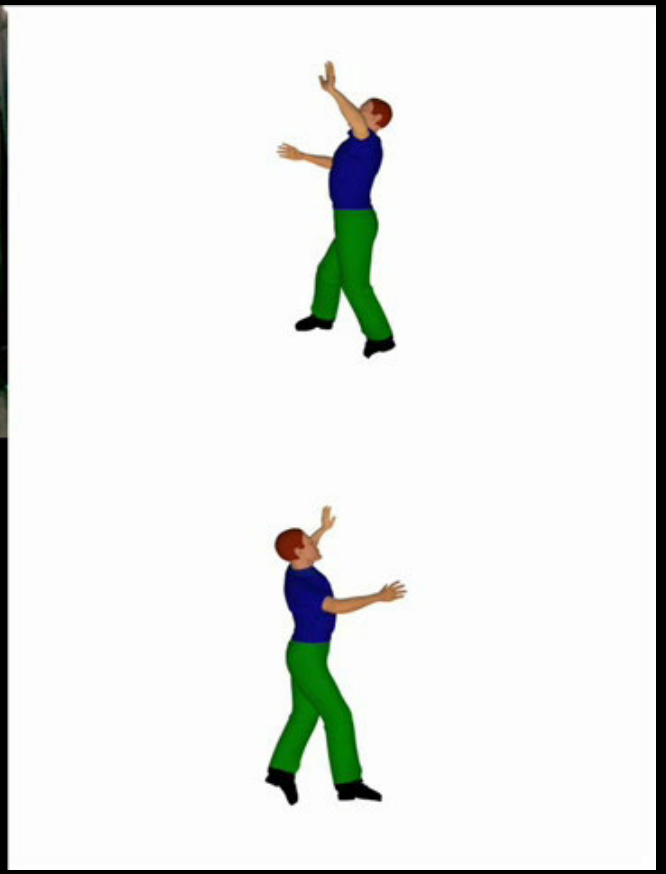


Temporal Mode Statistics



- Modes are often distant

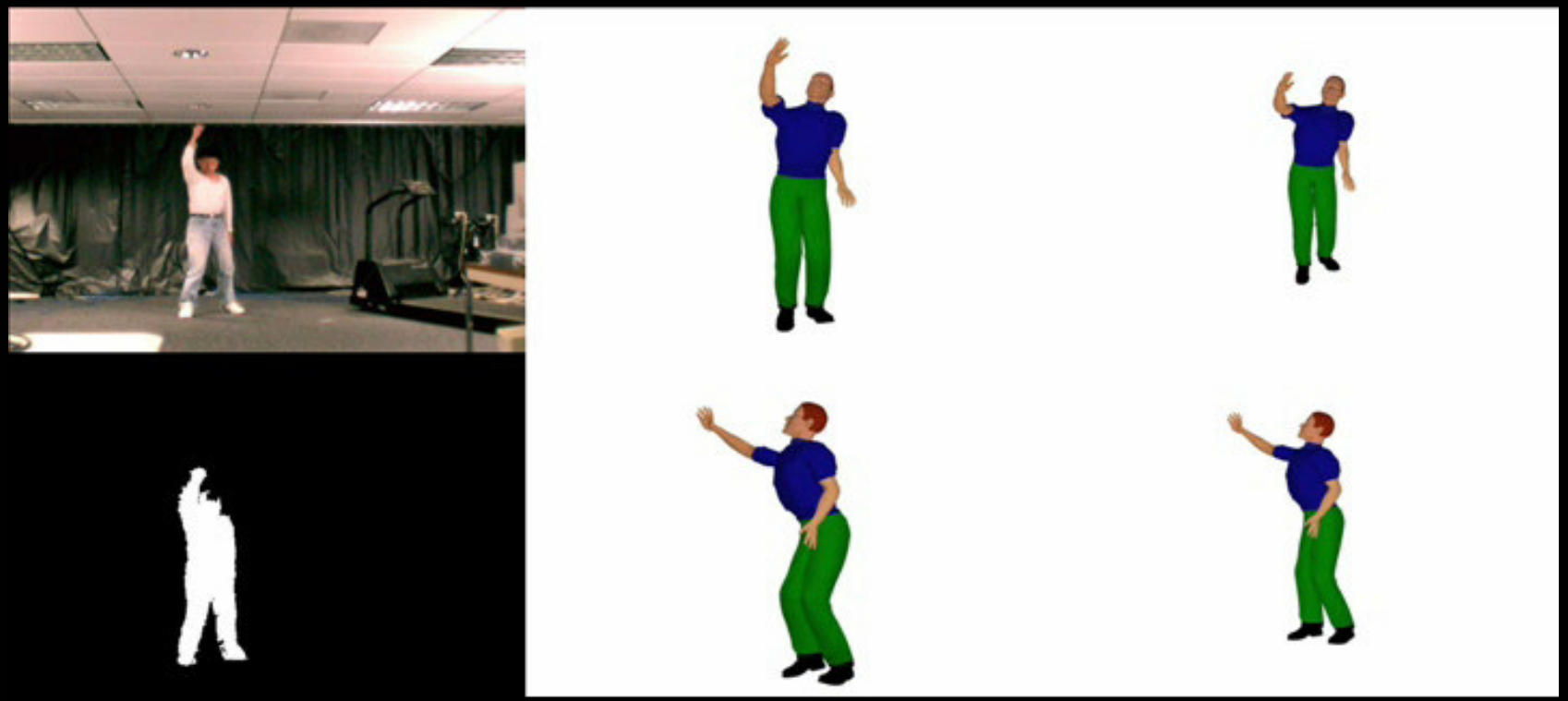
Turn during Dancing <v>



Notice imperfect silhouettes

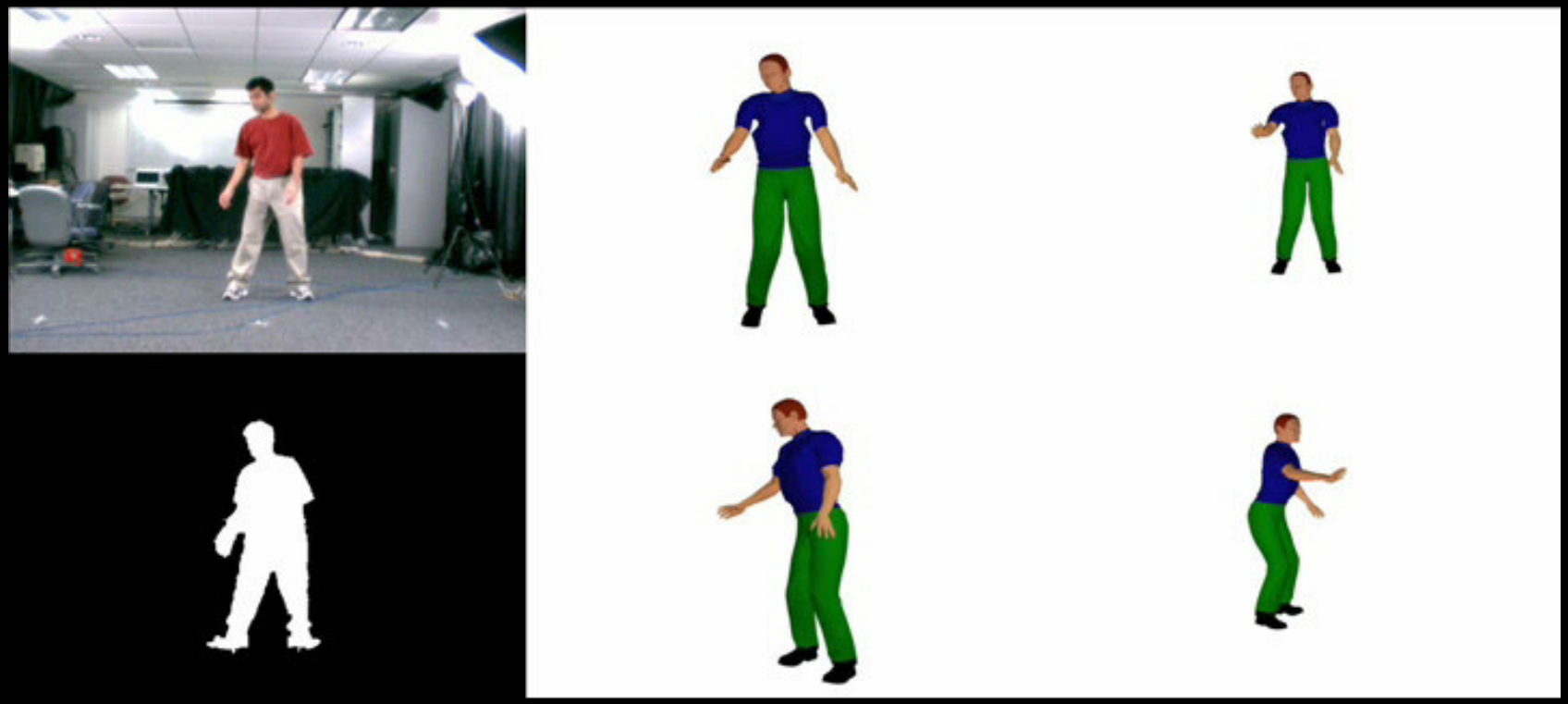
Picking from the Floor <v>

Input *Filtered* *Smoothed*



Washing a Window <v>

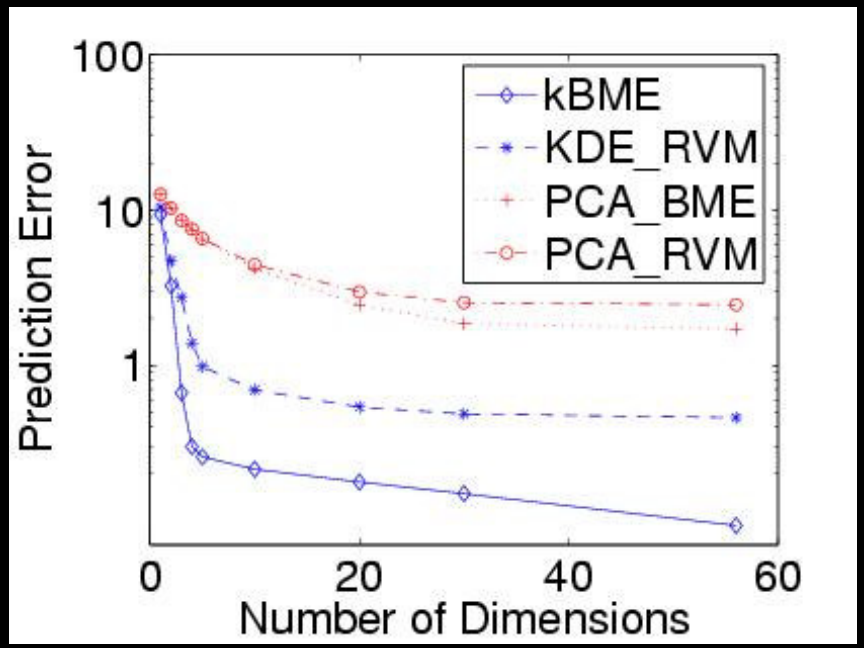
Input Filtered Smoothed



Low-dimensional State Inference

Sminchisescu, Kanaujia, Li, Metaxas, NIPS 2005

- The pose prediction problem is highly structured
  - Human joint angles are correlated, not independent
  - The intrinsic dimensionality of the human state space is typically lower than any original (ambient) one



RVM – Relevance Vector Machine
KDE – Kernel Dependency Estimator

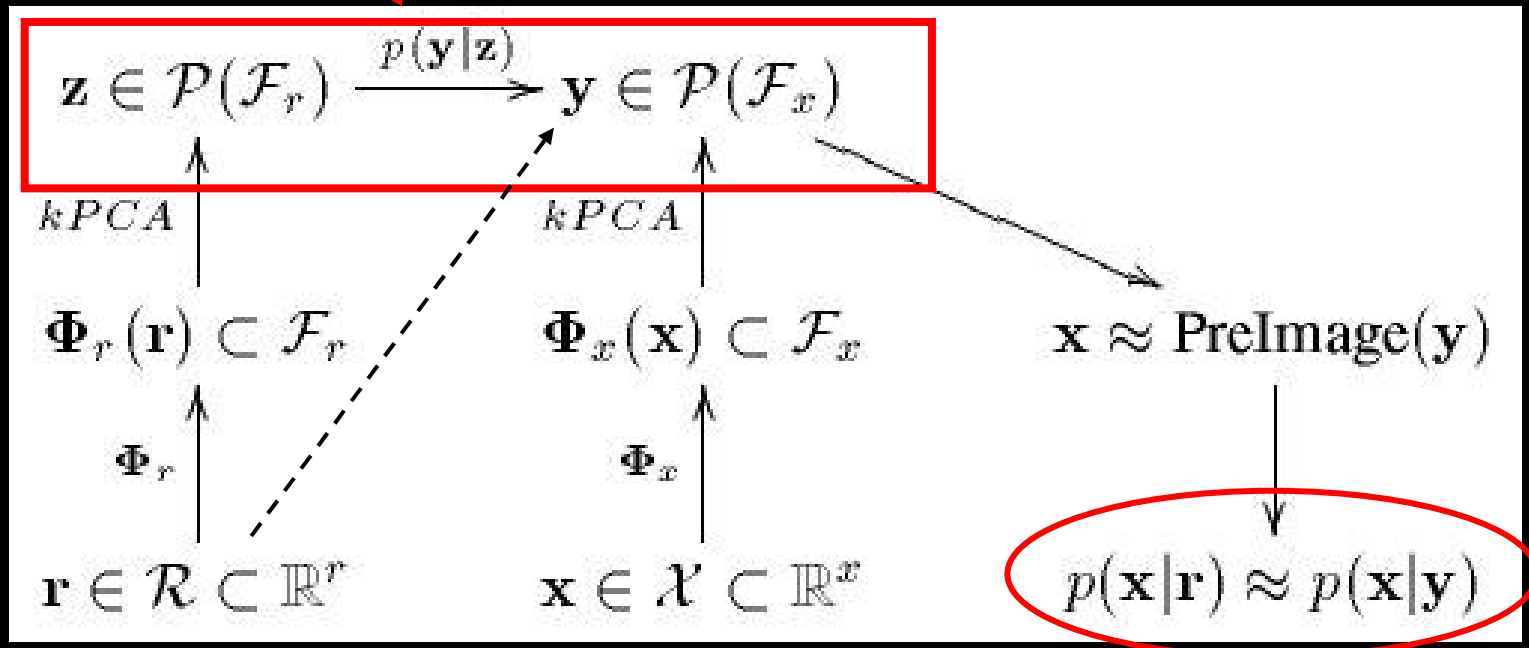
A Conditional Bayesian Mixture of Non-linear Kernel Induced Latent Experts

Inference in latent space → $p(\mathbf{y}_t|\mathbf{Z}_t) = \int_{\mathbf{y}_{t-1}} p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{z}_t) p(\mathbf{y}_{t-1}|\mathbf{Z}_{t-1})$

model this

Conditional mixture of experts

latent state



observation

original state

predictive distribution (e.g. for visualization)

Quantitative Comparisons

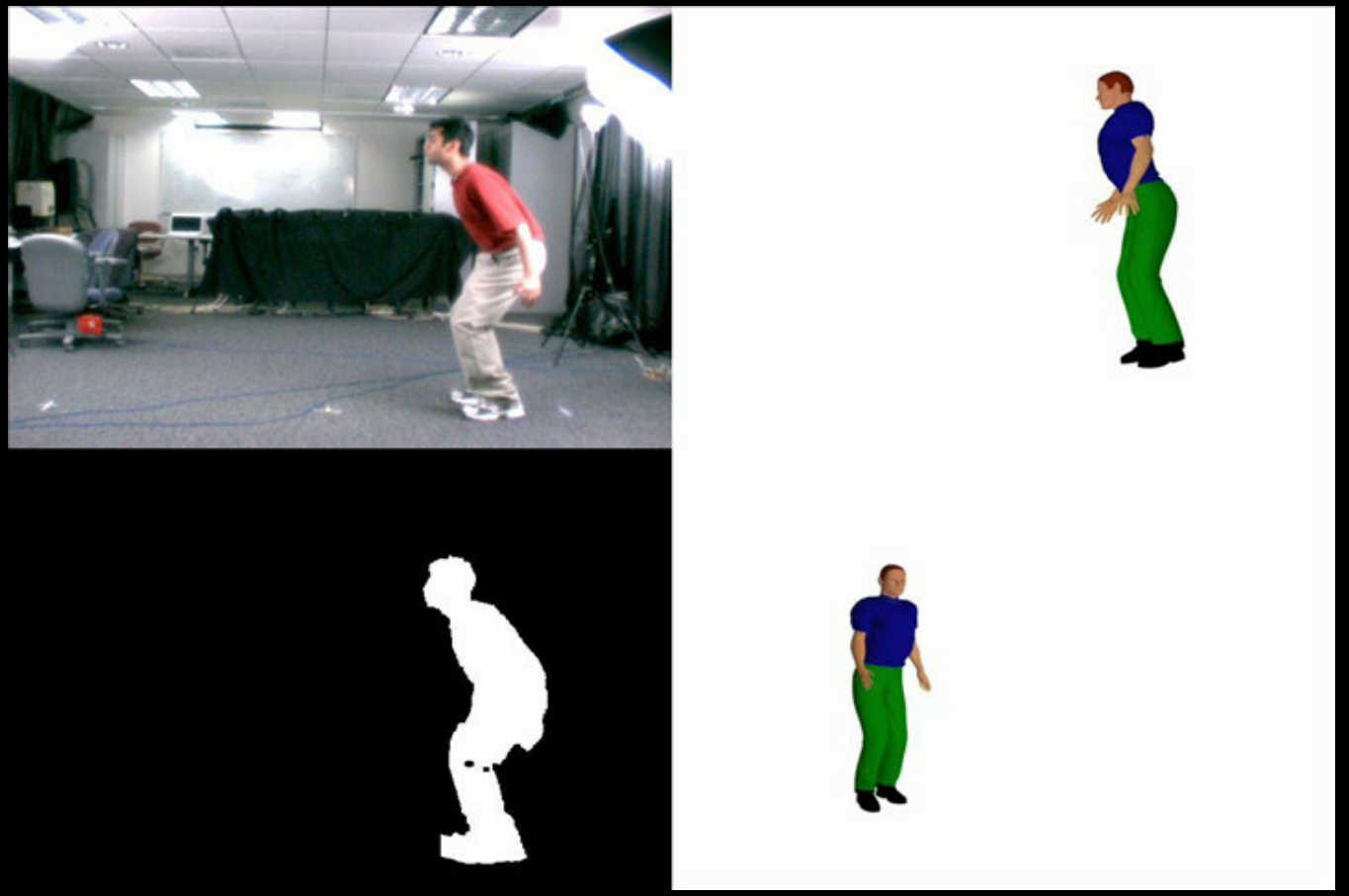
(average error / joint angle)

	KDE-RR	RVM	KDE-RVM	BME	kBME
Walk and turn back	10.46	4.95	7.57	4.27	4.69
Conversation	7.95	4.96	6.31	4.15	4.79
Run and turn left	5.22	5.02	6.25	5.01	4.92

	KDE-RR	KDE-RVM	kBME
Walk and Turn	7.59	7.15	3.72
Run and Turn	17.7	16.08	8.01

- Existing structured predictors (*e.g.* KDE) cannot handle multimodality
- Low-dimensional models (kBME) perform close to the high-dimensional ones at a lower computational cost
- Training and inference is about 10 time faster with kBME

Long Jump <v>



- Inference in a 6d latent space
- Notice self-occlusion (right side of the body)

Summary of Conditional Modeling

- Sound probabilistic framework for discriminative inference
- Robust and fast, effective for initialization (generative)
  - can be sped up using low-dimensional latent variable models, sparsity
- Very general – applies, in principle, to any problem where Kalman or particle filtering has been used
  - *e.g.* reconstruction, tracking, state inference under uncertainty
- Key ingredients (*Discriminative Density Propagation*)
  - Graphical model based on a discriminative temporal chain
  - Flexible modeling of complex multimodal *conditional* distributions
  - Compact representation and inference based on mixtures

What we have seen today

- Introduction, history, applications
- State of the art for 2d and 3d, human detection, initialization
- 3D human modeling, generative and discriminative computations
- Generative Models
  - Parameterization, shape, constraints, priors
  - Observation likelihood and dynamics
  - Inference algorithms
  - Learning non-linear low-dimensional representations and parameters
- Conditional (discriminative) models
  - Probabilistic modeling of complex, contextual inverse mappings
  - Observation modeling
  - Discriminative density propagation
  - Inference in latent, kernel-induced non-linear state spaces
- Conclusions and perspectives

Monocular 3D Reconstruction Prospects
Paths and Cycles

- Conditional / discriminative models will play a key role for robust 3d perception, including initialization and recovery from failure
  - The silhouette observation (as presented here) is not a major limitation
  - Can be easily relaxed to other feature types + FOA mechanisms
  - Generative modeling is expensive even with perfect foreground segmentation. Low-dimensional, latent variable models, and parameter learning can be a fix
  - Discriminative modeling replaces inference with prediction
- Generative model useful for verification, the 3d model important for training and occlusion reasoning
- Learn both representations and parameters
  - Hierarchical, low-dimensional, non-linear and sparse

Conclusions

- Modeling, inference and learning algorithms applied to **monocular 3D** human motion reconstruction
- We have discussed both generative and discriminative algorithms
  - Sound probabilistic formulations, efficient optimization and learning methods
- The human model is specific to a particular problem, but the optimization and learning methods we discussed are widely applicable

Open problems

- From the lab to the real world
  - Multiple people, occlusions, partial views, lighting changes
- The choice of representation
  - Extract automatically, handle variability
  - Accommodate multiple levels of detail, hierarchies
- Learning methods with good generalization properties
  - Small training sets, yes, if possible
  - But this will not preclude the choice of good structure to learn, or of smart on-line data-generation methods
  - Smoothness priors (*e.g.* `infinite' models, Bayesian integration) is not enough
    - Only an indication of high uncertainty (due to lack of data)
- Inference algorithms
  - Efficiently exploit generative and discriminative methods
  - The role of context in ambiguity resolution

Workshop at the upcoming CVPR

June 22nd, 2006: **Learning, Representation and Context for Human Sensing in Video**

(organized by Fernando De la Torre and me)

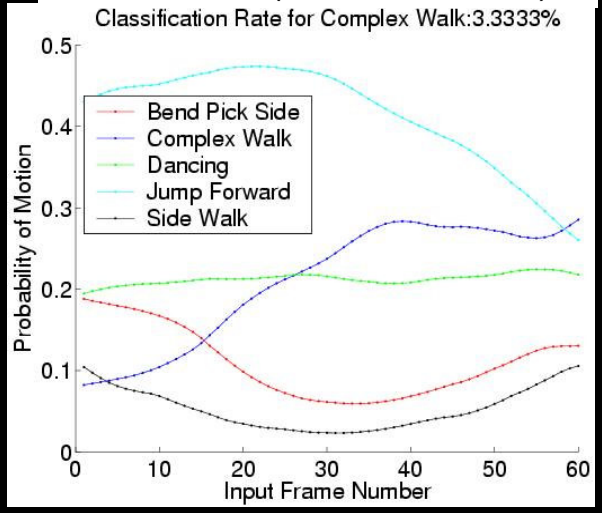
- Planned as a set of invited talks and panels
- Both retrospective and prospective
- See web page for topics and participants:

http://www.cs.toronto.edu/~crismin/workshop06.html

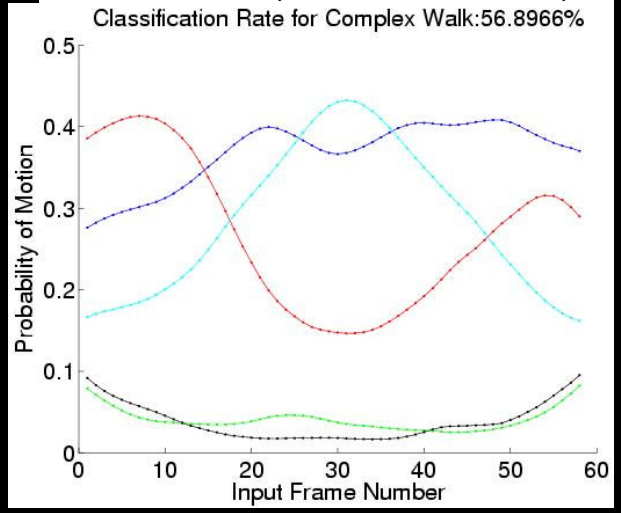
See this: The role of (observation) context for recognition performance

Conditional Models for Human Motion Recognition

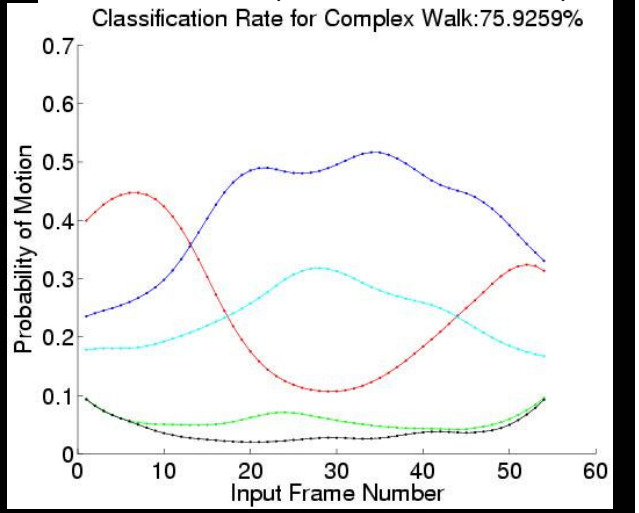
CRF, *W=0 (1 observation)*



CRF, *W=1 (3 observations)*



CRF, *W=3 (7 observations)*



- An HMM tested on the same sequence misclassifies the complex walk (1.5% accuracy), which is close to the performance of a conditional model with no context

If interested, pass by our poster

Conditional Models for Contextual Human Motion Recognition

C. Sminchisescu, A. Kanujia, Z. Li, D. Metaxas, ICCV 2005

Poster # 39, 13:30-16:00, October 20th, last day of conference

- *Poster available at:*

  http://www.cs.toronto.edu/~crismin/PAPERS/iccv05_poster.pdf

- *Paper available at:*

  http://www.cs.toronto.edu/~crismin/PAPERS/iccv05.pdf

- Please, use the web paper version. Due to a technical error, the camera ready version we submitted for the proceedings has not been the most recent

Thanks to my collaborators on various elements of the research I described

- Geoffrey Hinton (Toronto)
- Allan Jepson (Toronto)
- Atul Kanaujia (Rutgers)
- Zhiguo Li (Rutgers)
- Dimitris Metaxas (Rutgers)
- Bill Triggs (INRIA)
- Max Welling (UCIrvine)

Thanks for generously providing materials (slides, videos) from their work to:

- Michael Black (Brown)
- Bill Triggs (INRIA)

Thanks to the Organizers of ICCV 2005, for making this tutorial possible

Thank you for attending!

Materials, papers, videos available online at:

http://www.cs.toronto.edu/~crismin

http://www.cs.toronto.edu/~crismin/PAPERS/tutorial_iccv05.pdf