RWTH AACHEN UNIVERSITY

Computer Vision II - Lecture 14

Articulated Tracking II

10.07.2014

Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de

leibe@vision.rwth-aachen.de

Outline of This Lecture

- Single-Object Tracking
- Bayesian Filtering
  - Kalman Filters, EKF
  - Particle Filters
- Multi-Object Tracking
  - Data association
  - MHT, (JPDAF, MCMCDA)
  - Network flow optimization
- Articulated Tracking
  - GP body pose estimation
  - (Model-based tracking, AAMs)
  - Pictorial Structures

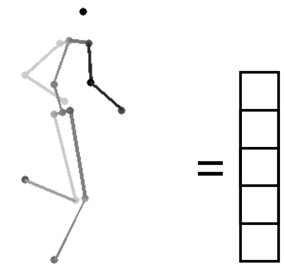
Image sources: Tomasz Svoboda, Deva Ramanan

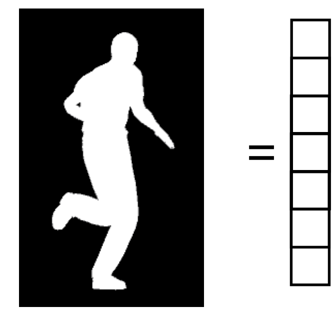
Topics of This Lecture

- Articulated Tracking
  - Motivation
  - Classes of Approaches
- Body Pose Estimation as High-Dimensional Regression
  - Representations
  - Training data generation
  - Latent variable space
  - Learning a mapping between pose and appearance
- Review: Gaussian Processes
  - Formulation
  - GP Prediction
  - Algorithm
- Applications
  - Articulated Tracking under Egomotion

Basic Classes of Approaches

- **Global methods**
  - Entire body configuration is treated as a point in some high-dimensional space.
  - Observations are also global feature vectors.
  - ⇒ View of pose estimation as a high-dimensional regression problem.
  - ⇒ Often in a subspace of "typical" motions...





- **Part-based methods**
  - Body configuration is modeled as an assembly of movable parts with kinematic constraints.
  - Local search for part configurations that provide a good explanation for the observed appearance under the kinematic constraints.
  - ⇒ View of pose estimation as probabilistic inference in a dynamic Graphical Model.

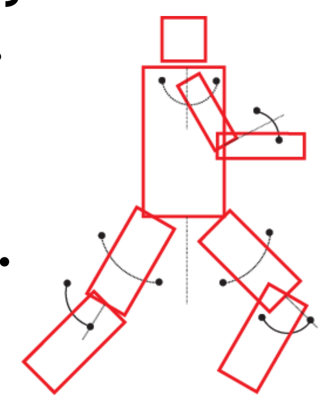
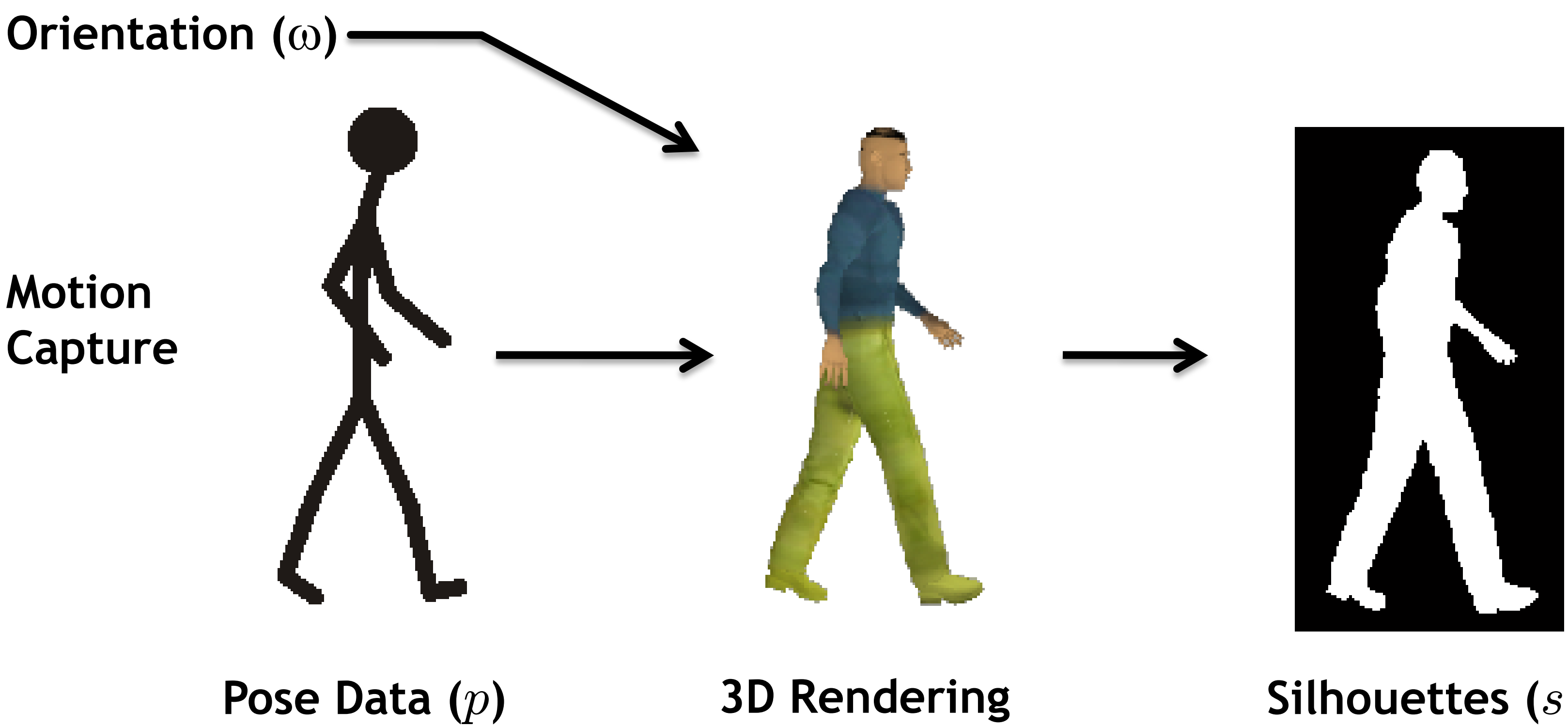


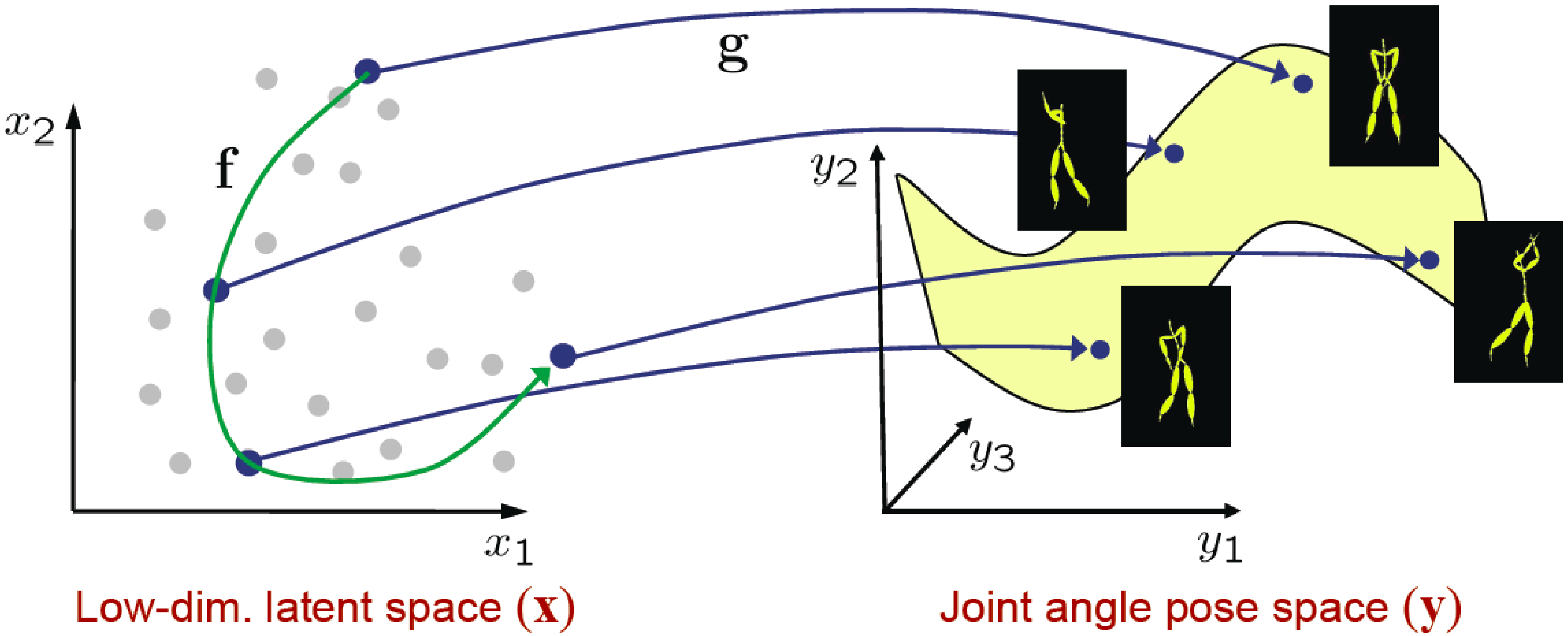
image sources: T. Jaeggli, D. Ramanan, T. Svoboda

Recap: Advantage of Silhouette Data

- Synthetic training data generation possible!
  - Create sequences of „Pose + Silhouette“ pairs
  - Poses recorded with Mocap, used to animate 3D model
  - Silhouette via 3D rendering pipeline



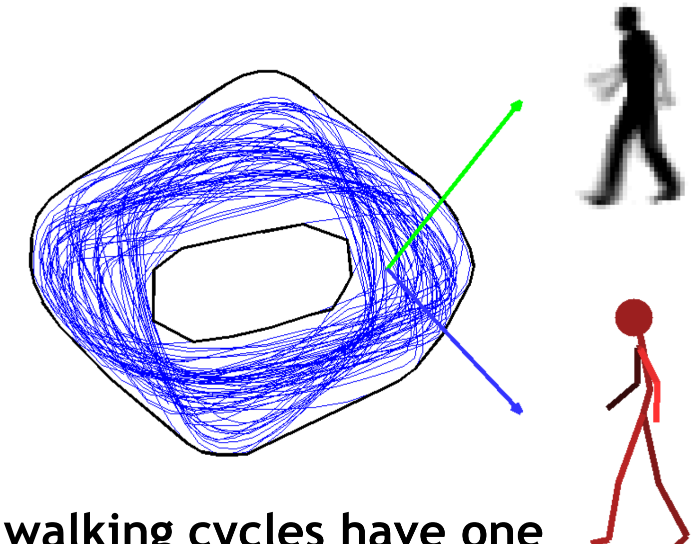
Recap: Latent Variable Models



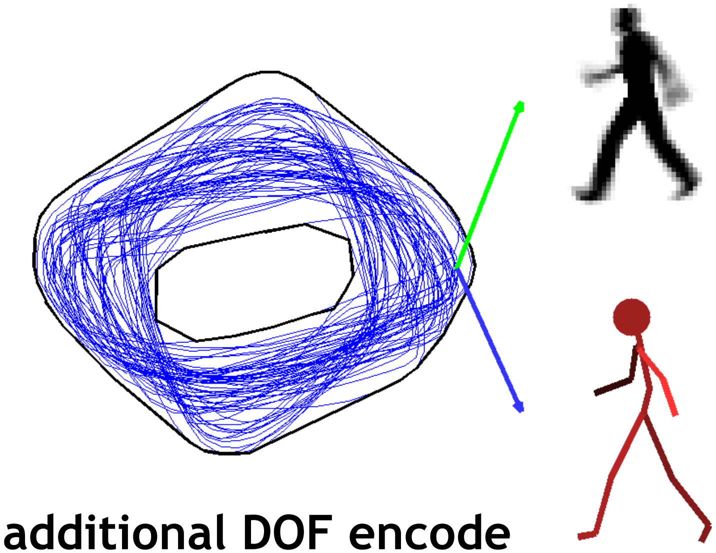
- Joint angle pose space is huge!
  - Only a small portion contains valid body poses.
  - ⇒ Restrict estimation to the subspace of valid poses for the task
  - Latent variable models: PCA, FA, GPLVM, etc.

image source: R. Urtasun

Recap: Articulated Motion in Latent Space



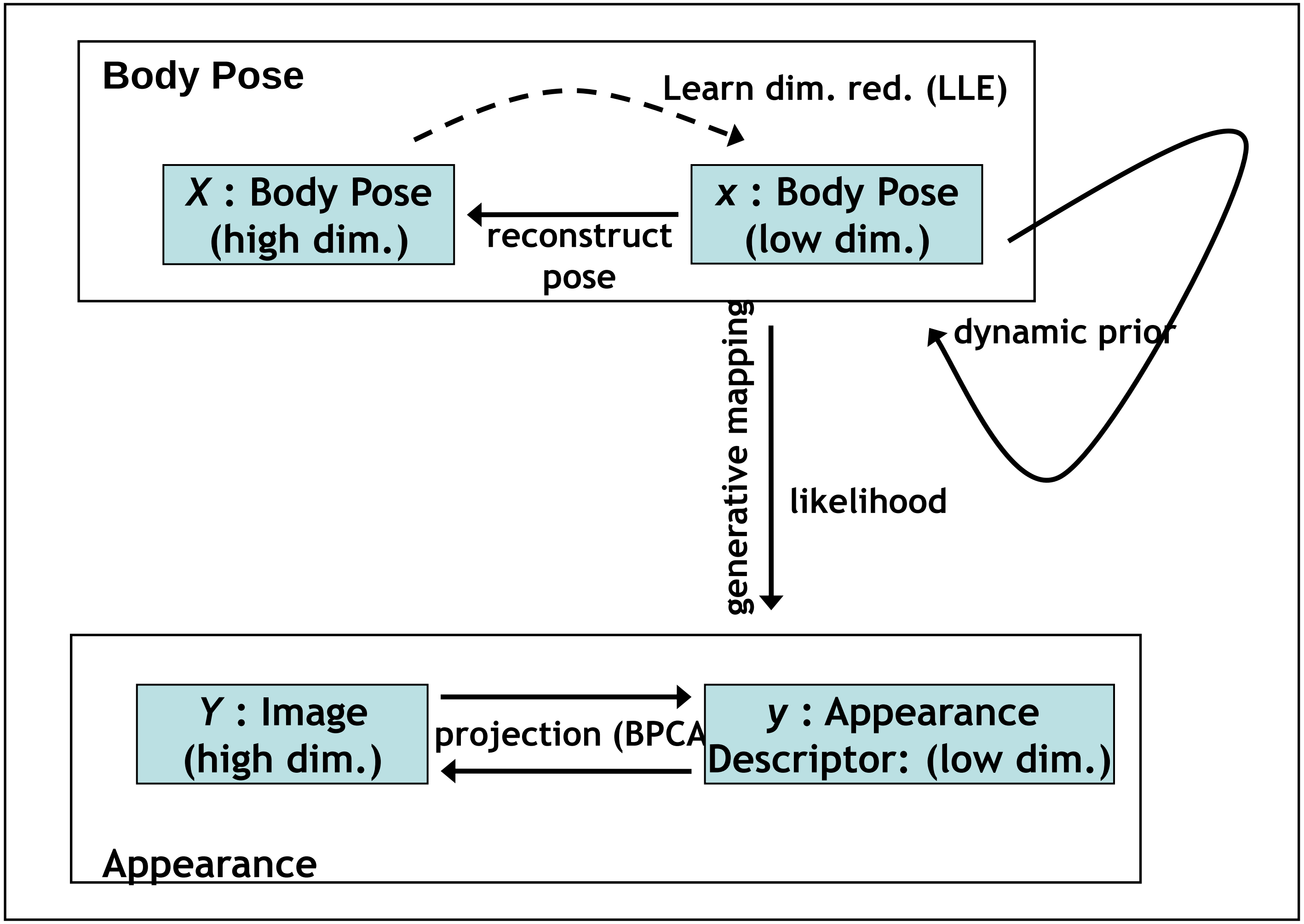
walking cycles have one main (periodic) DOF



additional DOF encode „walking style“

- Regression from latent space to
  - Pose → $p(pose \mid \mathbf{z})$
  - Silhouette → $p(silhouette \mid \mathbf{z})$
- Regressors need to be learned from training data.

Recap: Learning a Generative Mapping

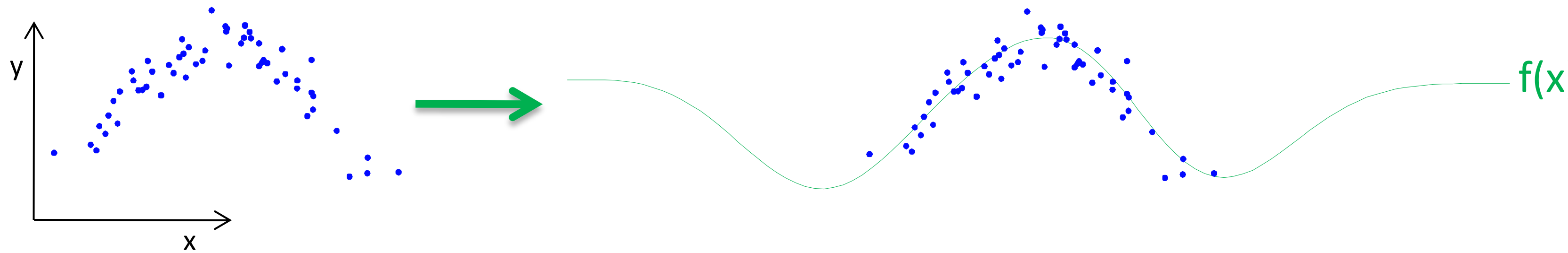


T. Jaeggli, E. Koller-Meier, L. Van Gool, "[Learning Generative Models for Monocular Body Pose Estimation]", ACCV 2007.

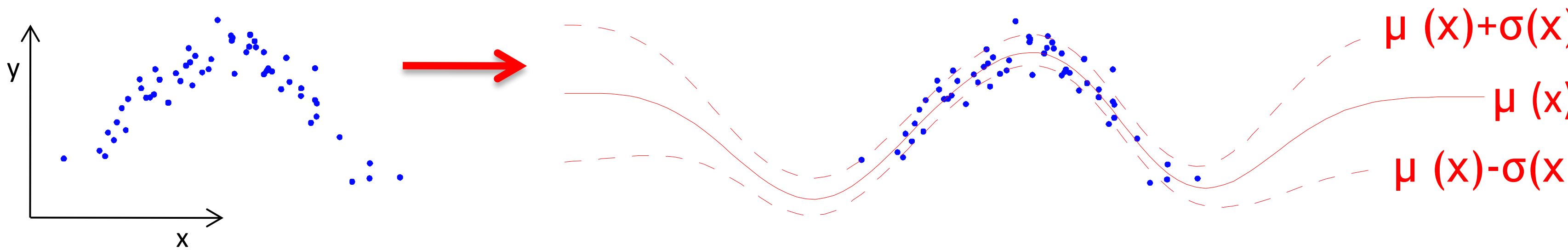
Slide credit: Tobias Jaeggli

Recap: Gaussian Process Regression

- "Regular" regression: $y = f(\mathbf{x})$



- GP regression: $p(y|\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), \sigma(\mathbf{x}))$



Recap: GP Prediction w/ Noisy Observations

- **Calculation of posterior:**
  - Corresponds to conditioning the joint Gaussian prior distribution on the observations:

$$\mathbf{f}_\star | X_\star, X, \mathbf{t} \sim \mathcal{N}(\bar{\mathbf{f}}_\star, \mathrm{cov}[\mathbf{f}_\star]) \qquad \bar{\mathbf{f}}_\star = \mathbb{E}[\mathbf{f}_\star | X, X_\star, \mathbf{t}]$$

  - with:

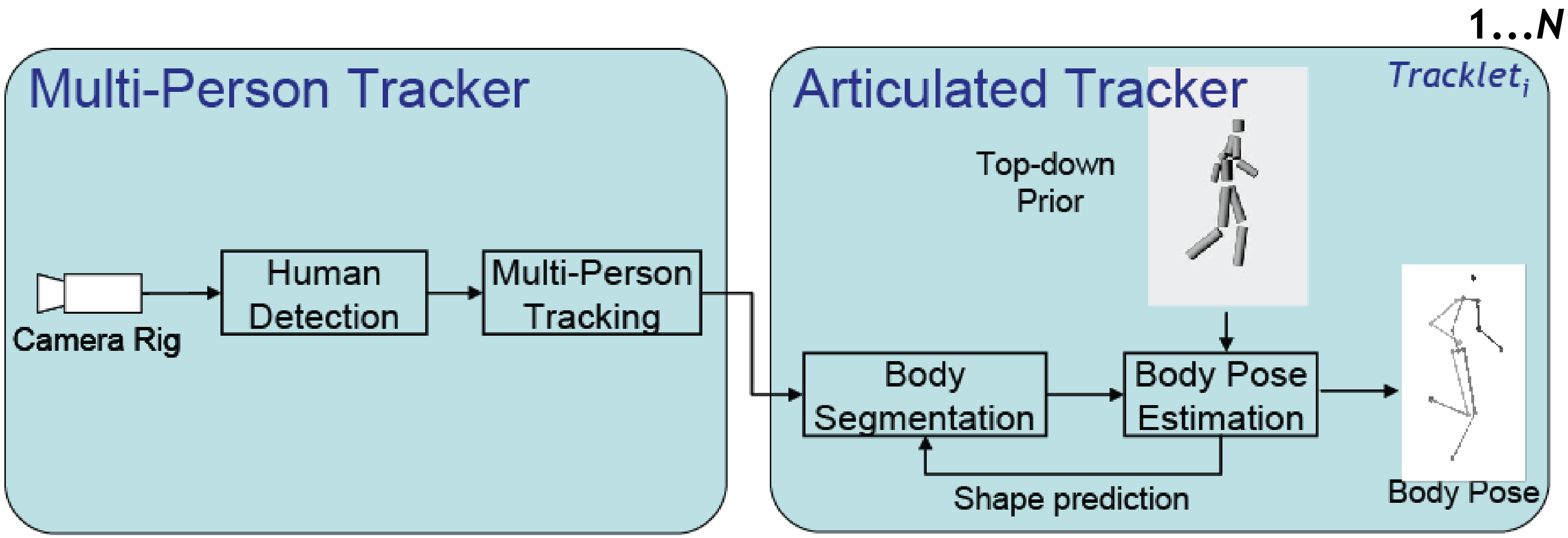
$$\bar{\mathbf{f}}_\star = K(X_\star, X)\left(K(X,X) + \sigma_n^2 I\right)^{-1}\mathbf{t}$$

$$\mathrm{cov}[\mathbf{f}_\star] = K(X_\star, X_\star) - K(X_\star, X)\left(K(X,X) + \sigma_n^2 I\right)^{-1} K(X, X_\star)$$

⇒ **This is the key result that defines Gaussian process regression!**

- The predictive distribution is a Gaussian whose mean and variance depend on the test points X_* and on the kernel $k(\mathbf{x},\mathbf{x}')$, evaluated on the training data X.

Recap: Articulated Multi-Person Tracking



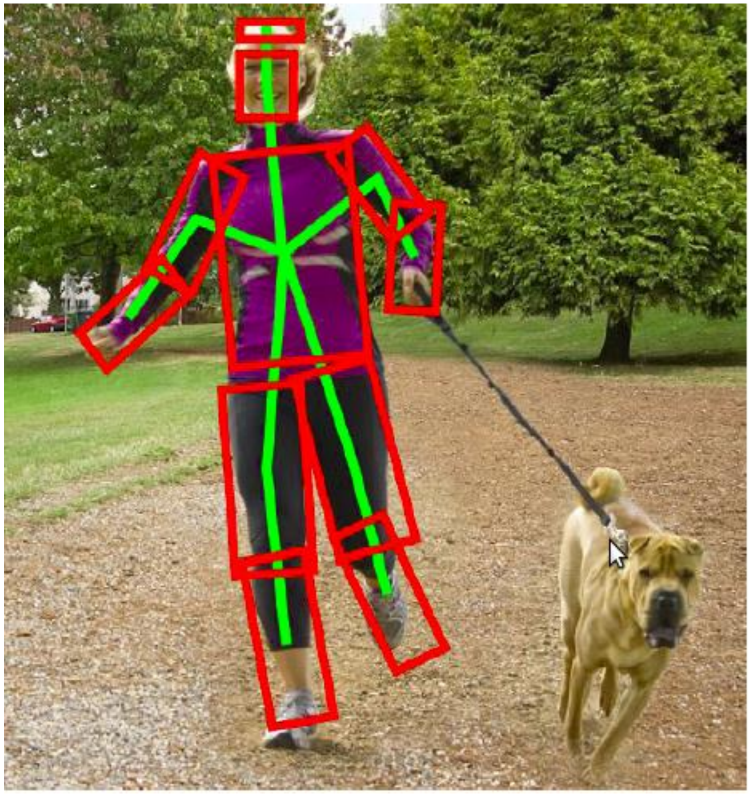
- **Idea: Only perform articulated tracking where it's easy!**
- **Multi-person tracking**
  - Solves hard data association problem
- **Articulated tracking**
  - Only on individual "tracklets" between occlusions
  - GP regression on full-body pose

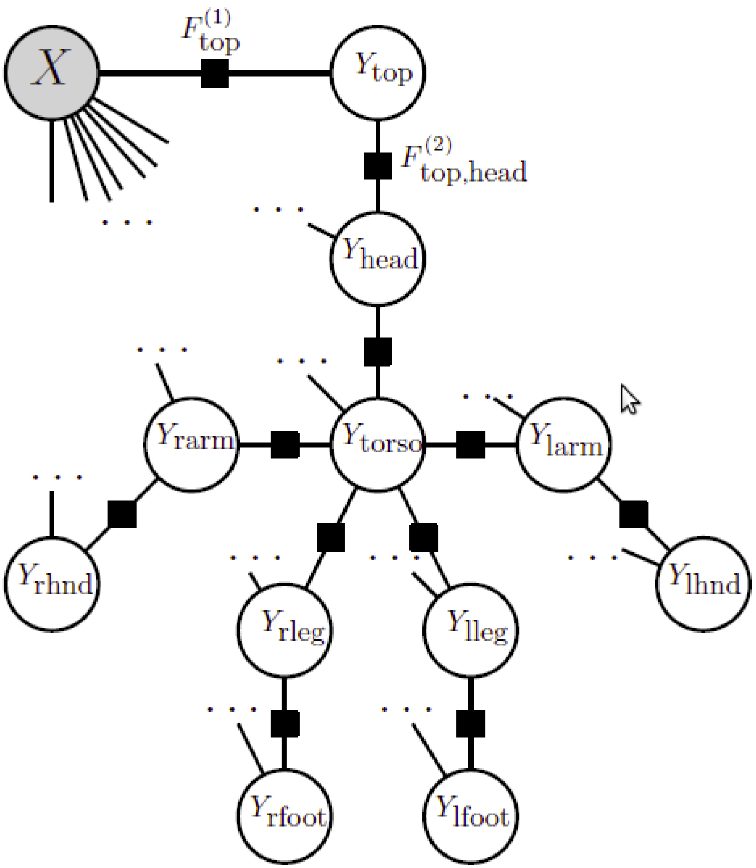
[Gammeter, Ess, Jaeggli, Schindler, Leibe, Van Gool, ECCV'08]

Topics of This Lecture

- Pictorial Structures
  - Model components
  - Prior
  - Likelihood Model
- Recap: Inference
  - Sum-Product algorithm
  - Max-Sum algorithm
- Efficient Inference in Pictorial Structures
  - Generalized Distance Transform
  - Effect on Computation
- Results

Today: Pictorial Structures



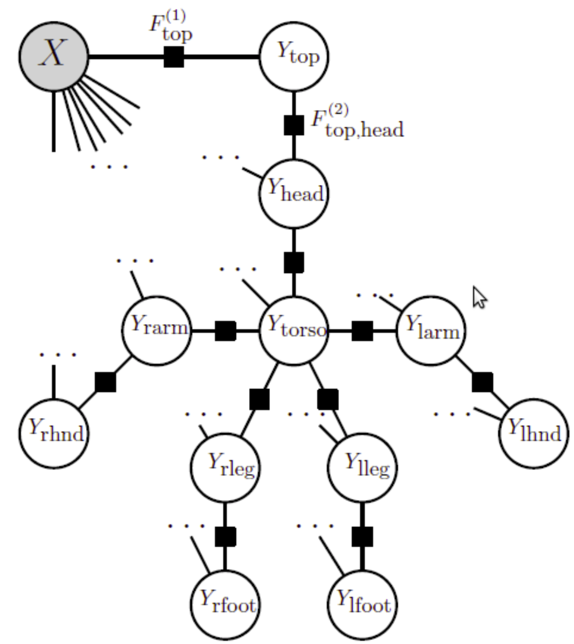


- **Pose estimation as inference in a graphical model**
  - **[Fischler & Elschlaeger, 1973; Felzenszwalb & Huttenlocher, 00]**

Pictorial Structures

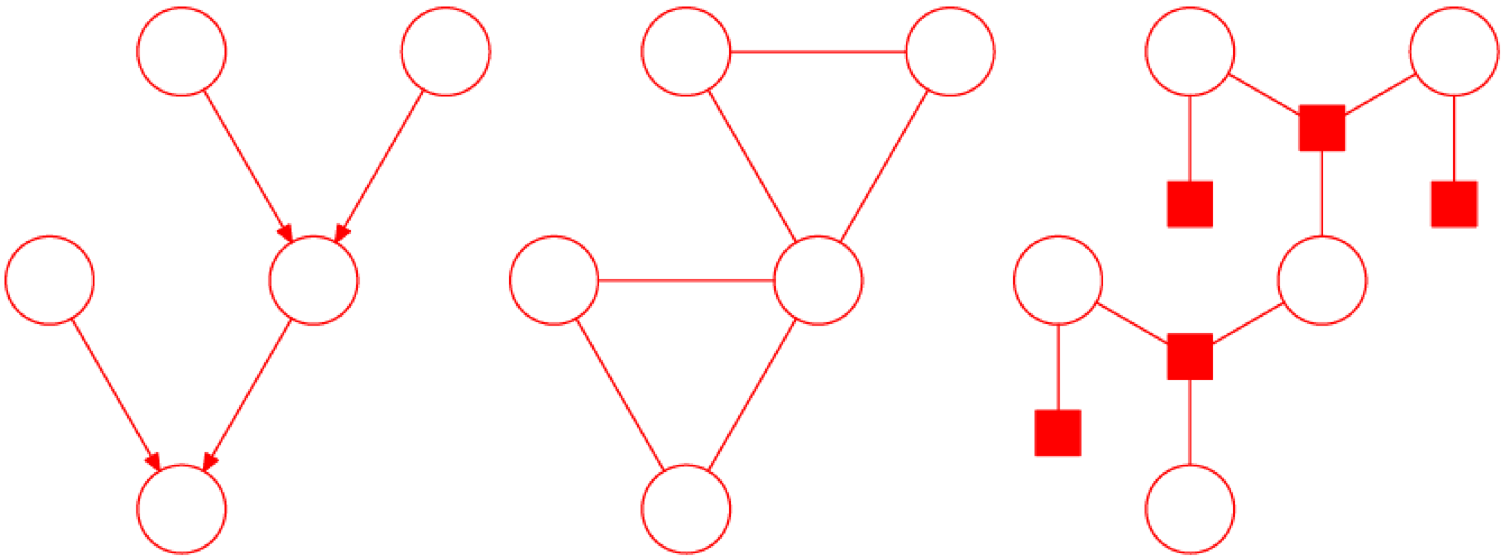
- Each body part one variable node
  - Torso, head, etc. (11 total)
- Each variable represented as tupel
  - E.g., $y_{torso} = (x, y, \theta, s)$ with
  - (x, y) image coordinates
  - θ rotation of the part
  - s scale
- Discretize label space y into L states
  - E.g., size of L for $y = (x, y, \theta, s)$
  - $L = 125 \times 125 \times 8 \times 4 \approx 500'000$

  $\Rightarrow$ Efficient search needed to make this feasible!



P. Felzenszwalb, D. Huttenlocher, Pictorial Structures for Object Recognition, IJCV, Vol. 61(1), 2005.

Recap: Factor Graphs



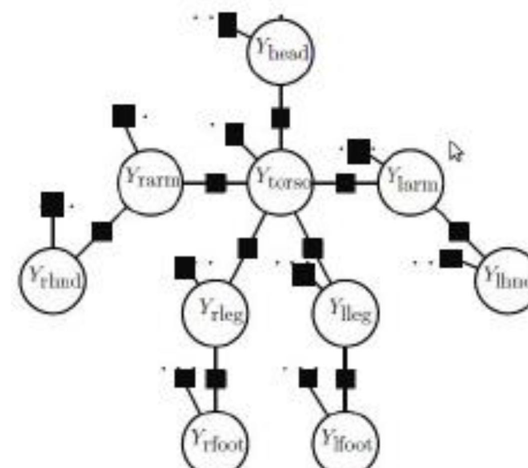
- Joint probability
  - Can be expressed as product of factors: $p(\mathbf{x}) = \frac{1}{Z}\prod_{s} f_s(\mathbf{x}_s)$
  - Factor graphs make this explicit through separate factor nodes.
- Converting a directed polytree
  - Conversion to undirected tree creates loops due to moralization!
  - Conversion to a factor graph again results in a tree!

Image source: C. Bishop, 2006

Two Model Components

- **Prior** $p(L)$
  - Models kinematic dependencies between body parts
  - Tree-structured prior (constraints b/w body parts) lead to efficient inference
  - Generalized distance transform provide additional efficiency

- **Likelihood of body parts** $p(E \mid L)$
  - Models possible appearances of body parts
  - Substantial improvements in recent years in appearance modeling and detection



- **Finding body parts = Pose estimation**

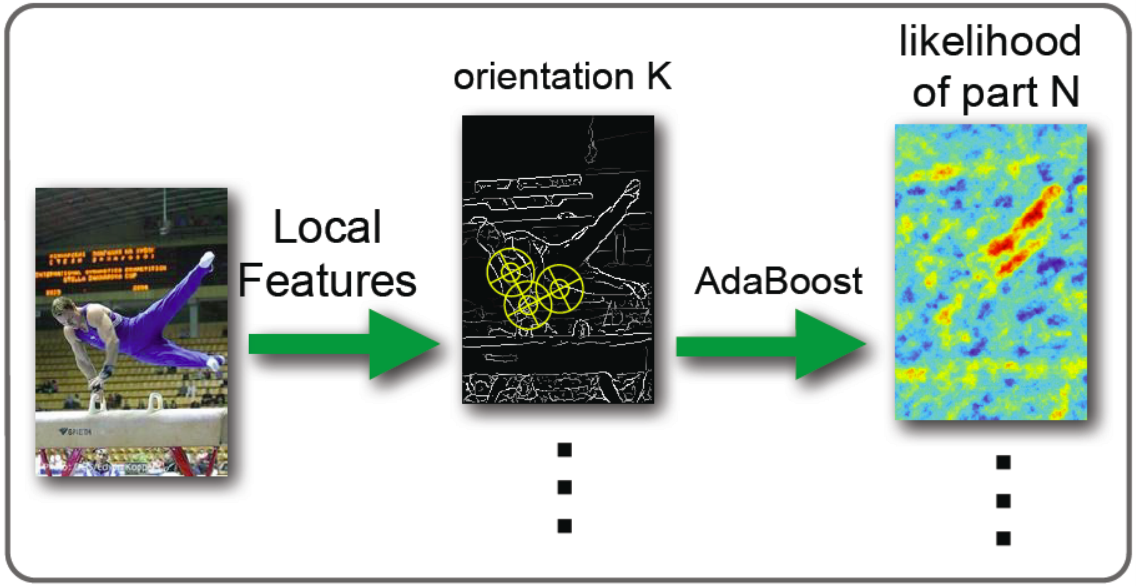
Pictorial Structures: Model Components

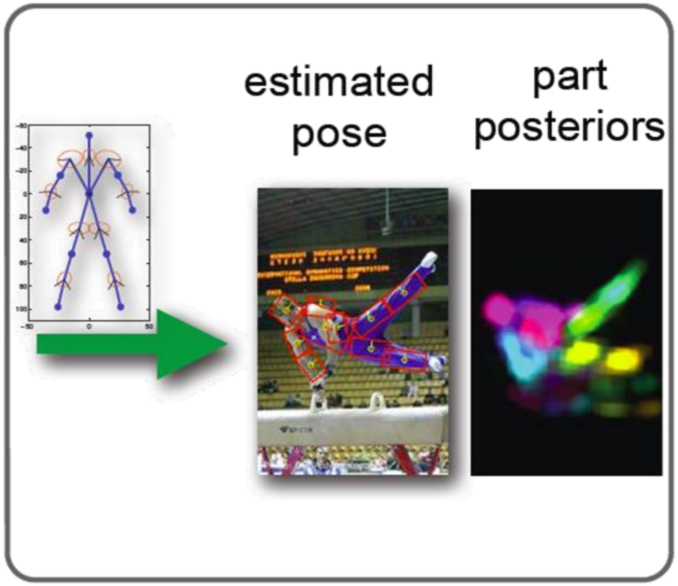
- **Body is represented as flexible combination of parts**

posterior over body poses

$$p(L|E) \propto p(E|L)p(L)$$

likelihood of observations **prior on body poses**





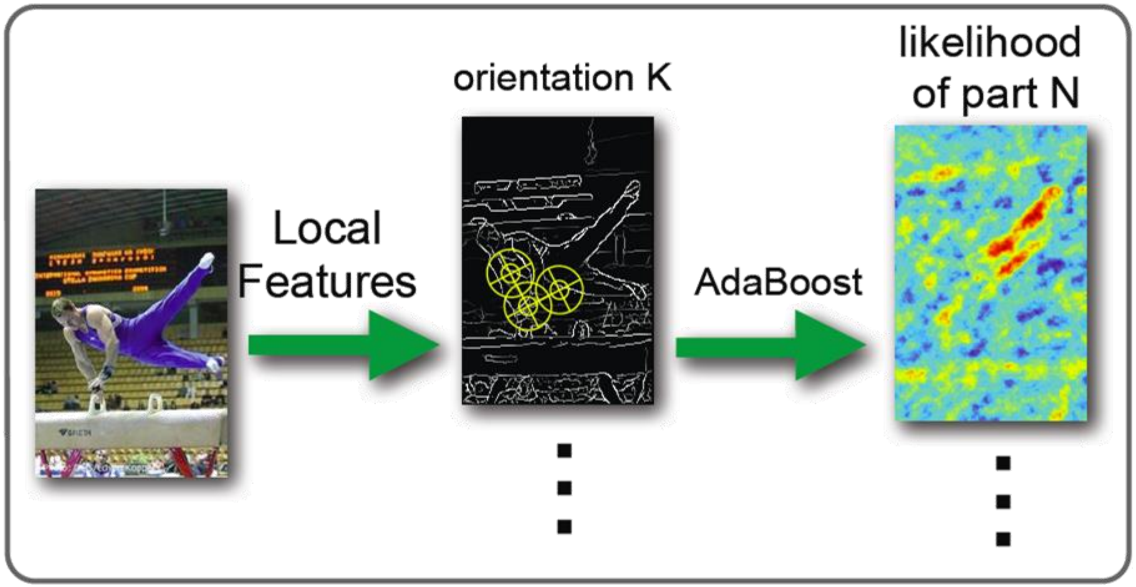
Pictorial Structures: Model Components

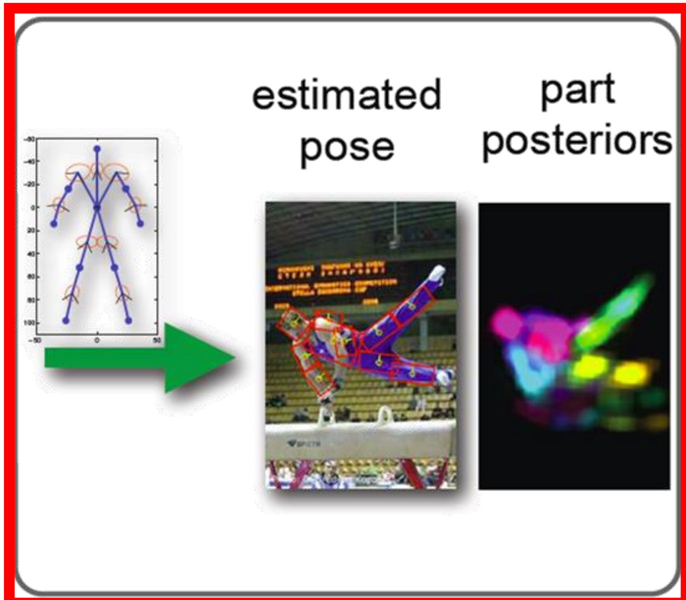
- **Body is represented as flexible combination of parts**

posterior over body poses

$$p(L|E) \propto p(E|L)p(L)$$

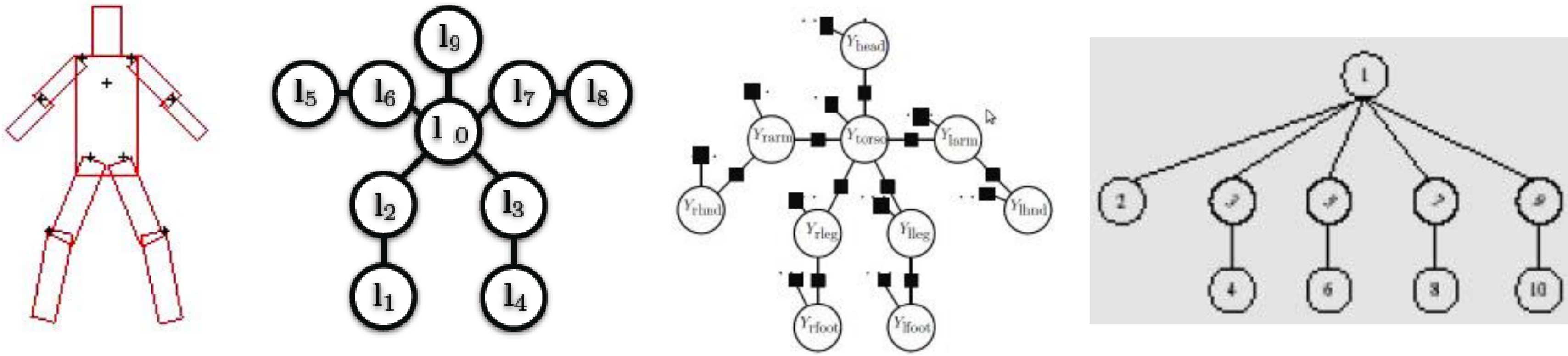
likelihood of observations   prior on body poses



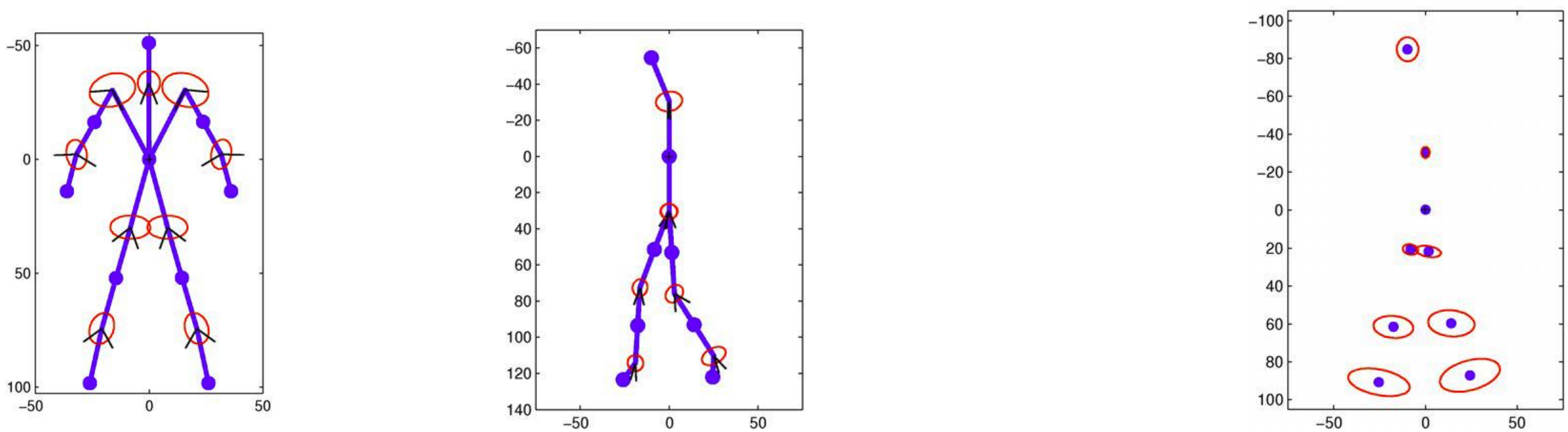


Human Body Pose Models - Prior $p(L)$

- E.g., [Felzenszwalb & Huttenlocher, IJCV'05]

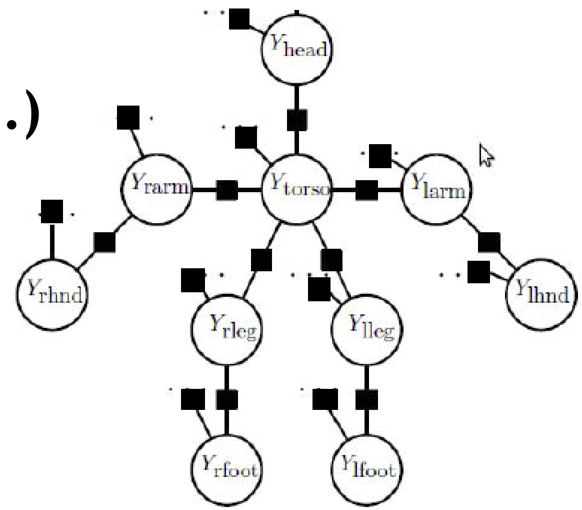


- E.g., [Andriluka et al., IJCV'12]



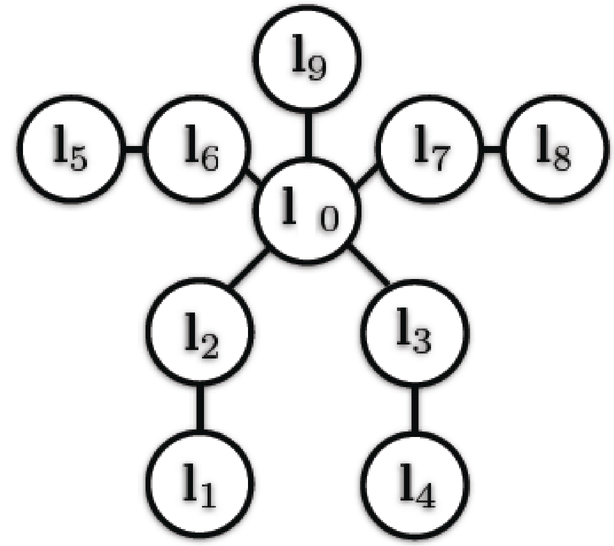
Pictorial Structures

- Potentials (= energies = factors)
  - Unaries for each body part (torso, head, ...)
  - Pairwise between connected body parts



- Body pose estimation
  - Find most likely part location
  - $\Rightarrow$ Sum-product algorithm (marginals)
  - Find the best overall configuration
  - $\Rightarrow$ Max-sum algorithm (MAP estimate)
- Complexity
  - Let k be the number of body parts (e.g., $k = 10$)
  - L is the size of the label space (e.g., several 100k)
  - Max-sum algorithm in general: $\mathcal{O}(k\ L^2)$
  - For specific pairwise potentials: $\mathcal{O}(k\ L)$

Kinematic Tree Prior



- **Notation**
  - (from [Andriluka et al., IJCV'12])
  - Body configuration

$$L = \{l_0, l_1, \ldots, l_N\}$$

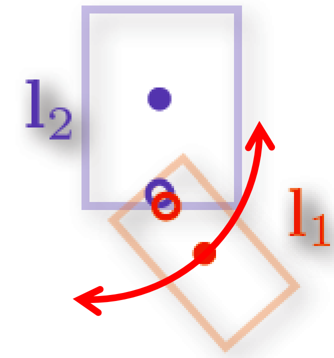
  - Each body part: $l_i = (x_i,\ y_i,\ \theta_i,\ s_i)$

- **Prior**

$$p(L) = p(l_0) \prod_{(i,j)\in G} p(l_i|l_j)$$

  - with $p(l_0)$ assumed uniform
  - with $\mathrm{p}(l_i \mid l_j)$ modeled using a Gaussian

Kinematic Tree Prior



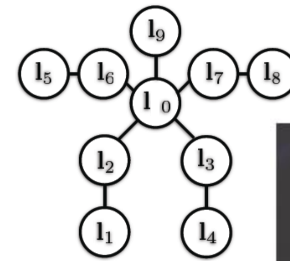
- **Gaussian assumption for $\mathrm{p}(l_i \mid l_j)$**
  - This may seem like a significant limitation.
  - E.g., distribution of forearm configuration given the upper arm is semi-circular, rather than Gaussian!
- **Solution** [Felzenszwalb & Huttenlocher, IJCV'05]
  - Transform part configuration l_i into coordinate system of the joint, where the distribution is captured well by a Gaussian:

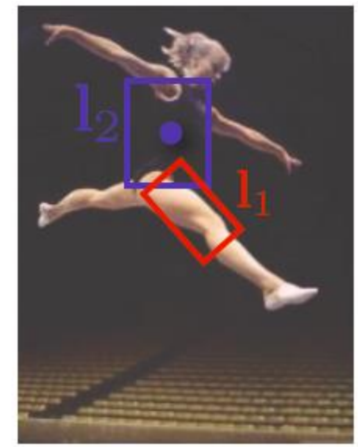
$$T_{ji}(l_i) = \begin{bmatrix} x_i + s_i d_x^{ji} \cos\theta_i - s_i d_y^{ji} \sin\theta_i \\ y_i + s_i d_x^{ji} \sin\theta_i + s_i d_y^{ji} \cos\theta_i \\ \theta_i \\ s_i \end{bmatrix}$$

  - with $d^{ji} = \begin{bmatrix} d_x^{ji} \\ d_y^{ji} \end{bmatrix}$ position of the joint between parts i and j, represented in the coordinate system of part i

Kinematic Tree Prior

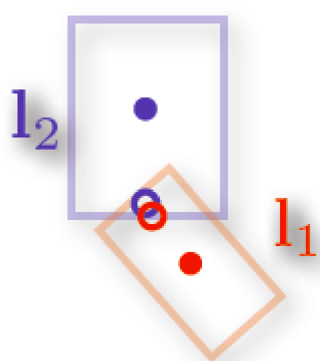
- **Represent pairwise part relations**



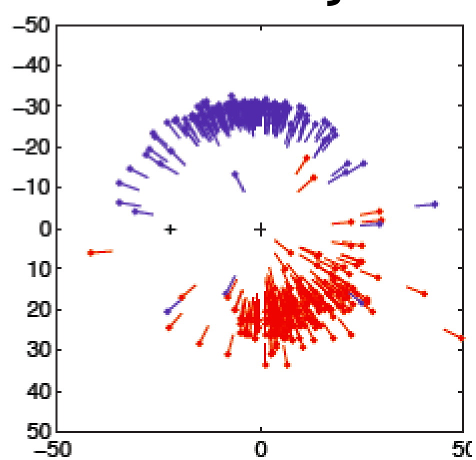


$$p(L) = p(l_0) \prod_{(i,j)\in G} p(l_i|l_j)$$

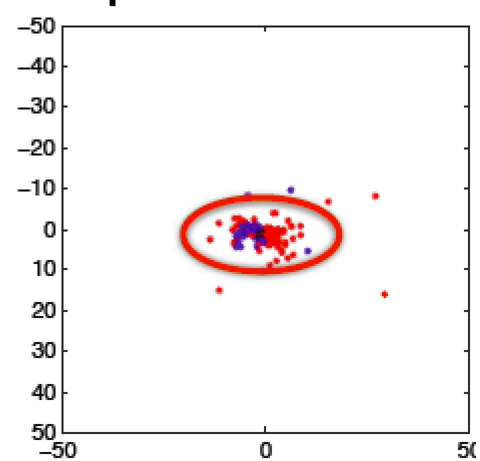
$$p(l_i|l_j) = \mathcal{N}\left(T_{ji}(l_i) - T_{ij}(l_j)|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}\right)$$



Part locations relative to joint



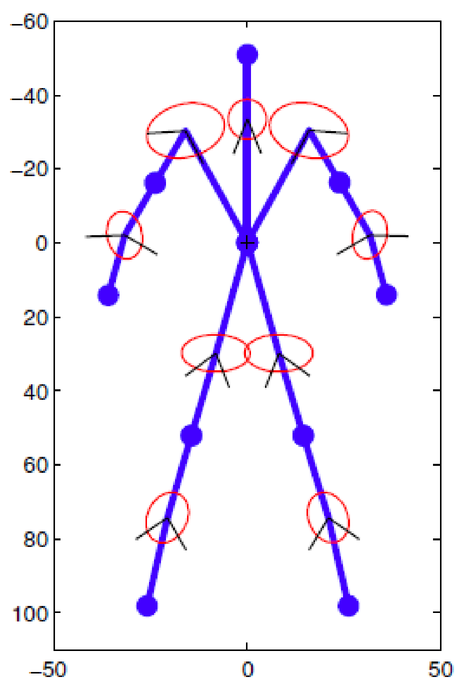
Transformed part locations



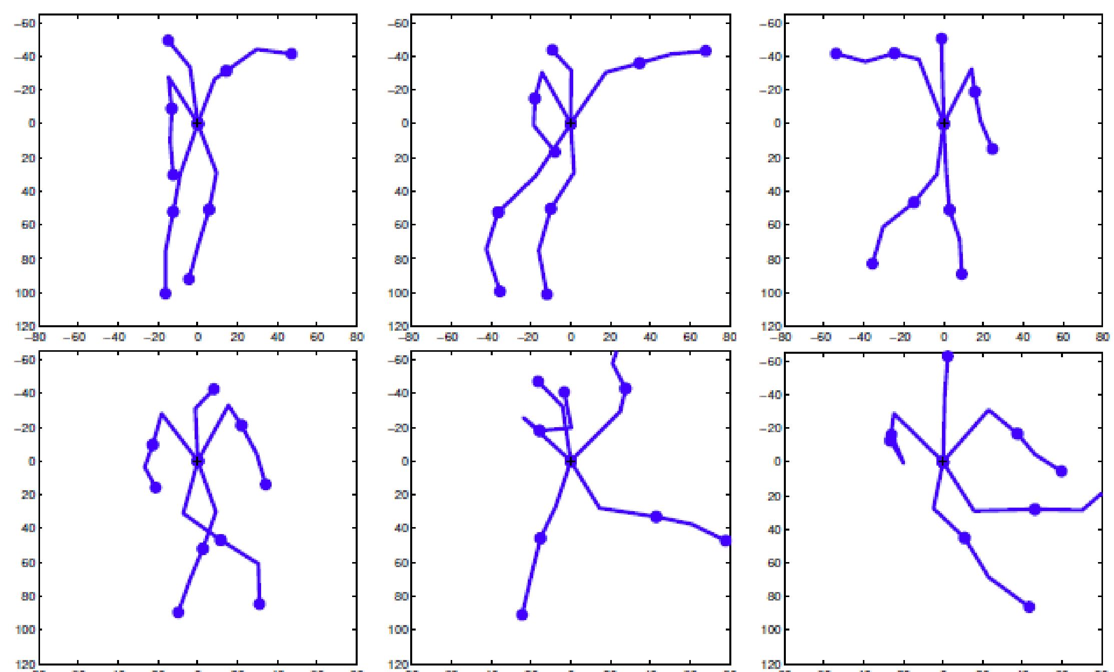
Kinematic Tree Prior

- **Prior parameters** $\{T_{ij},\ \Sigma_{ij}\}$
  - **Learned using maximum likelihood**

Mean pose



Several independent samples



Pictorial Structures: Model Components

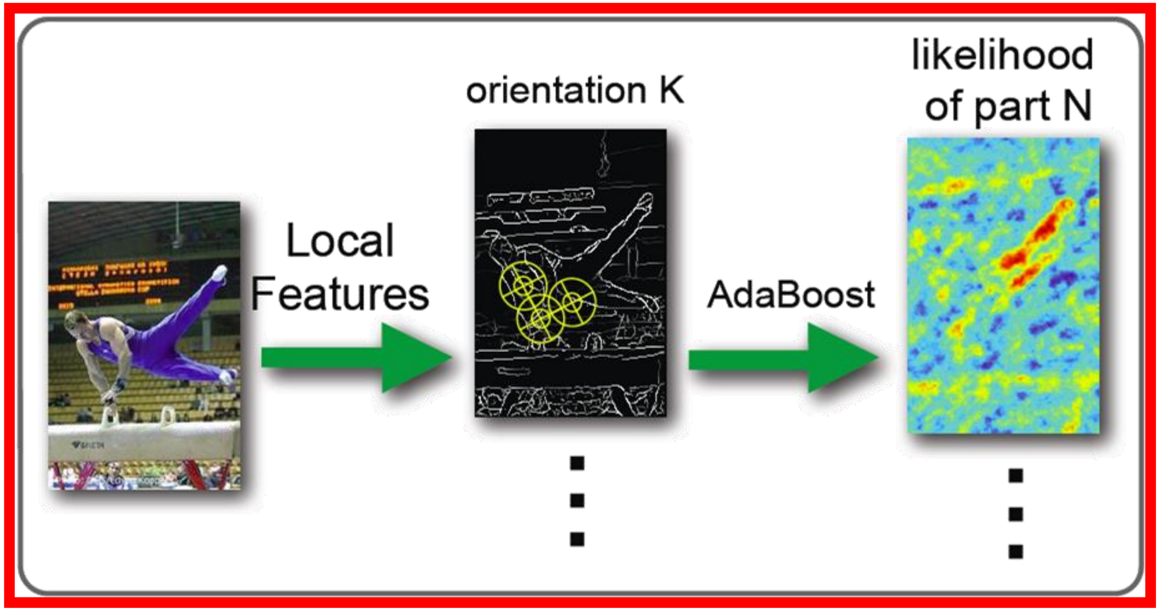
- Body is represented as flexible combination of parts

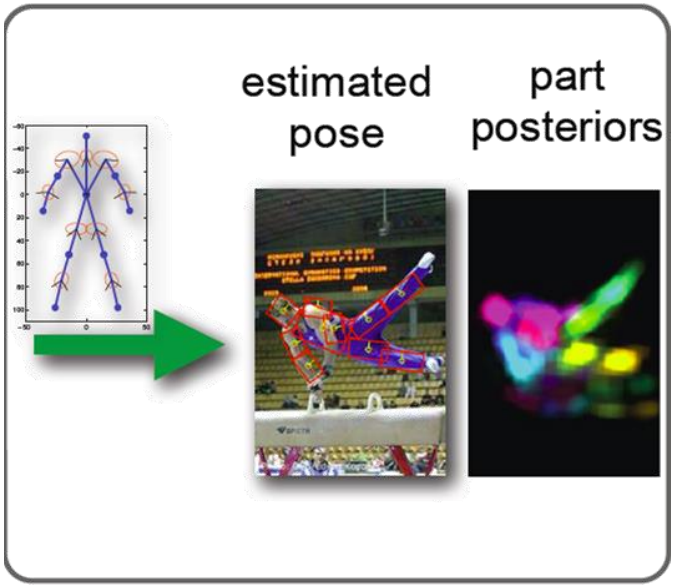
posterior over body poses

$$p(L|E) \propto p(E|L)p(L)$$

likelihood of observations

prior on body poses





Slide credit: Bernt Schiele

Likelihood Model

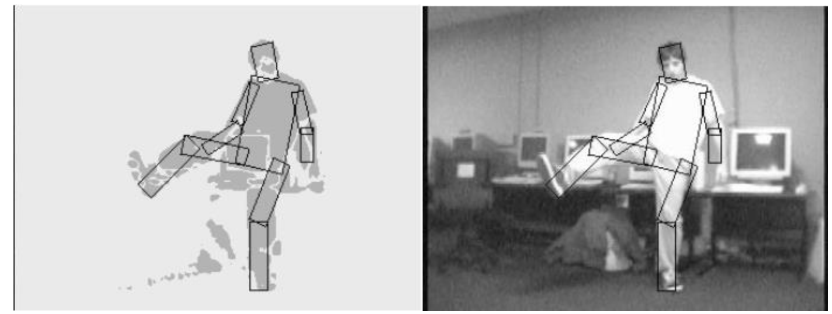
- Assumption
  - Evidence (image features) for each part independent of all other parts

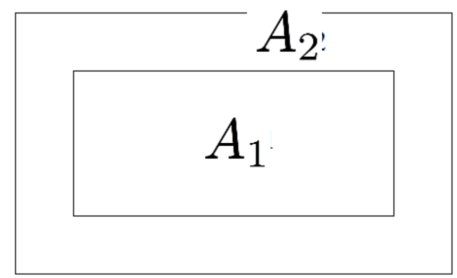
$$p(E|L) = \prod_{i=0}^{N} p(E|l_i)$$

- The assumption is clearly not correct, but
  - Allows efficient computation
  - Works rather well in practice
  - Training data for different body parts should cover "all" appearances

Likelihood Model

- Many variants have been proposed over the years...
  - [Felzenszwalb, IJCV'05]
    - Modeled using rectangular parts based on fg/bg probabilities





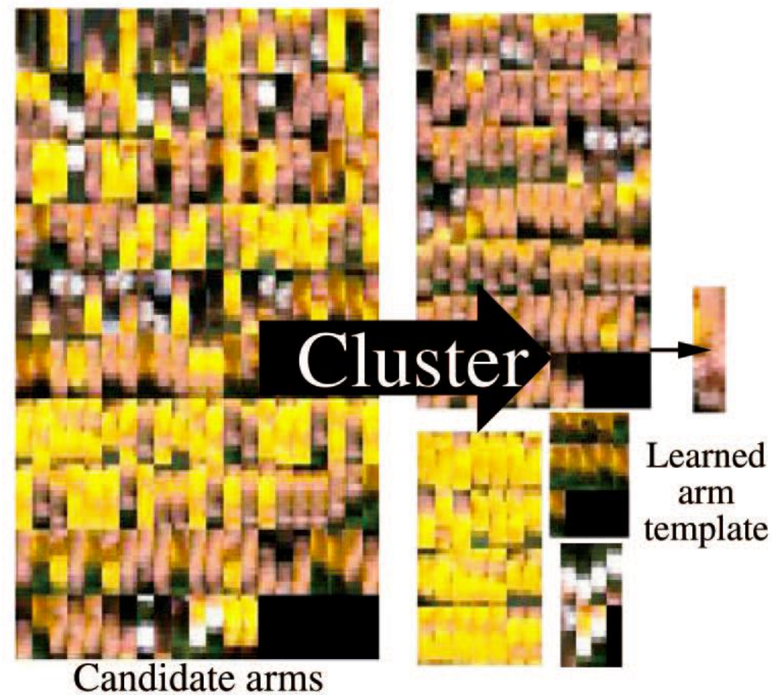
- N_1: #fg pixels inside rectangle
- A_1: size of rectangle
- N_2: #fg pixels inside border
- A_2: size of border area
- t : #pixels in image

- Part likelihood

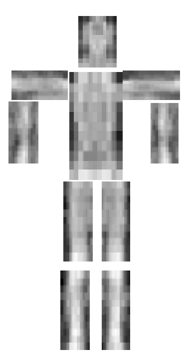
$$p(E|l_i) = q_1^{N_1}(1 - q_1^{A_1 - N_1})q_2^{N_2}(1 - q_2^{A_2 - N_2})0.5^{t - A_1 - A_2}$$

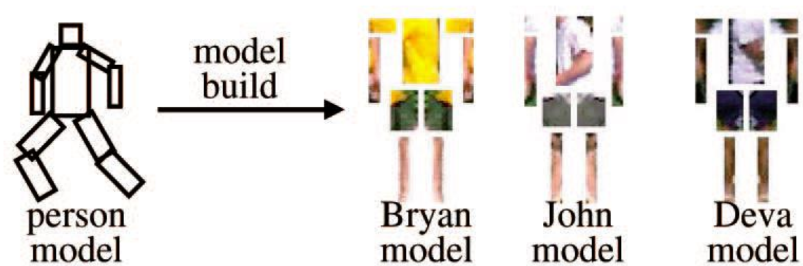
Likelihood Model

- Many variants have been proposed over the years...
  - [Felzenszwalb, IJCV'05]
  - [Ramanan, PAMI'07]
    - Learn person-specific body part appearance models by clustering
    - Initially only color models
    - Later extended by edge models [NIPS'06]



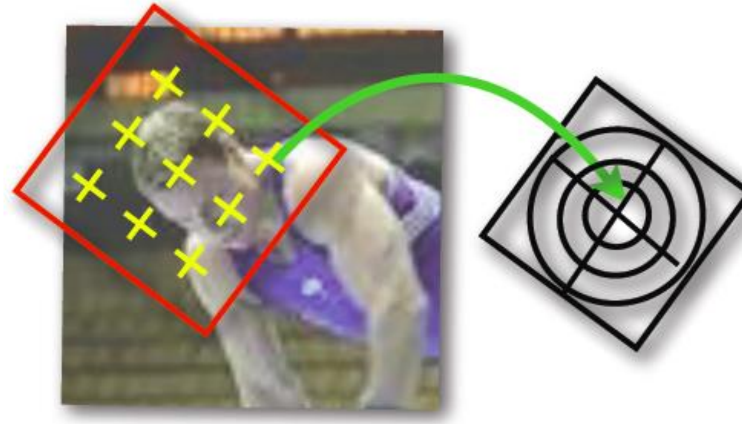
Edge model [NIPS'06]





Likelihood Model

- Many variants have been proposed over the years...
  - [Felzenszwalb, IJCV'05]
  - [Ramanan, PAMI'07]
  - ...
  - [Andriluka, IJCV'12]
    - Boosted classifiers based on local feature descriptors (e.g., Shape context, SIFT)
    - Part likelihood derived from Boosting score



Decision stump weight

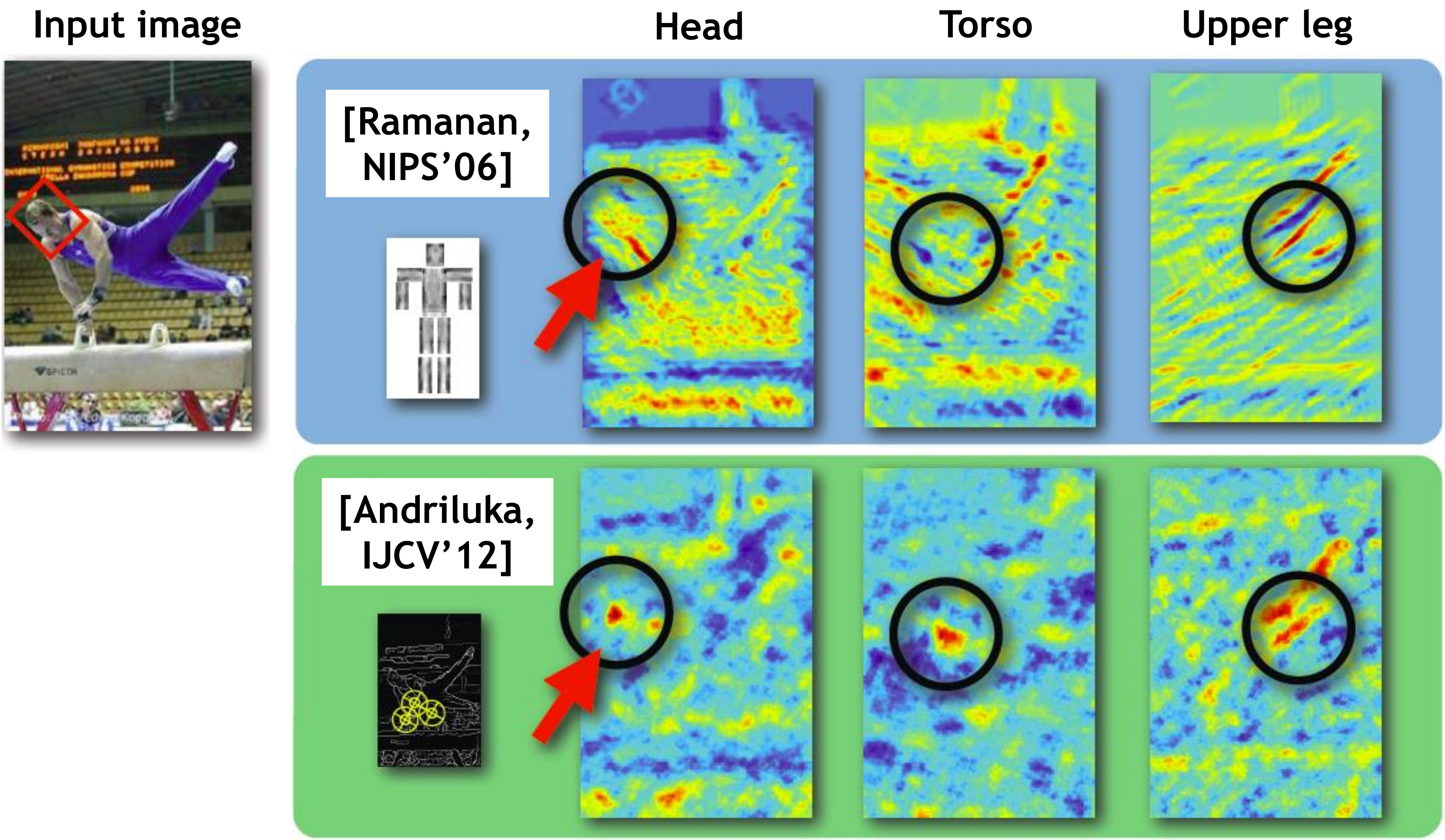
Decision stump output

$$\tilde{p}(E|l_i) = \max\left(\frac{\sum_t \alpha_{i,t} h_t(e_i(l_i))}{\sum_t \alpha_{i,t}}, \varepsilon_0\right)$$

Part location

Small constant to deal with partial occlusions

Likelihood Models – Part Likelihoods



Topics of This Lecture

- Pictorial Structures
  - Model components
  - Prior
  - Likelihood Model
- Recap: Inference
  - Sum-Product algorithm
  - Max-Sum algorithm
- Efficient Inference in Pictorial Structures
  - Generalized Distance Transform
  - Effect on Computation
- Results

Recap: Sum-Product Algorithm

- Objectives
  - Efficient, exact inference algorithm for finding marginals.
- Procedure:
  - Pick an arbitrary node as root.
  - Compute and propagate messages from the leaf nodes to the root, storing received messages at every node.
  - Compute and propagate messages from the root to the leaf nodes, storing received messages at every node.
  - Compute the product of received messages at each node for which the marginal is required, and normalize if necessary.

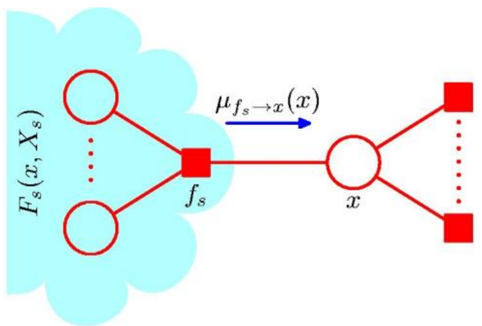
$$p(x) \propto \prod_{s \in \mathrm{ne}(x)} \mu_{f_s \to x}(x)$$

- Computational effort
  - Total number of messages = 2 · number of graph edges.

Recap: Sum-Product Algorithm

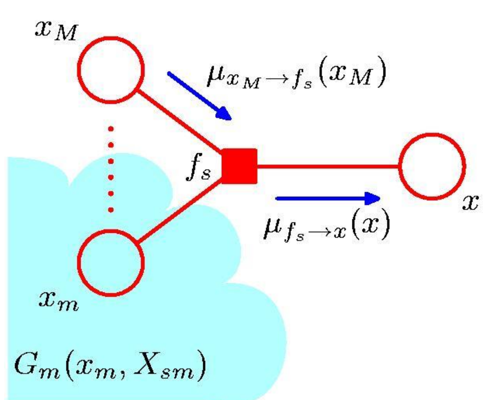
- Two kinds of messages
  - Message from factor node to variable nodes:
    - Sum of factor contributions

$$\mu_{f_s \to x}(x) \equiv \sum_{X_s} F_s(x, X_s)$$
$$= \sum_{X_s} f_s(\mathbf{x}_s) \prod_{m \in \mathrm{ne}(f_s) \setminus x} \mu_{x_m \to f_s}(x_m)$$



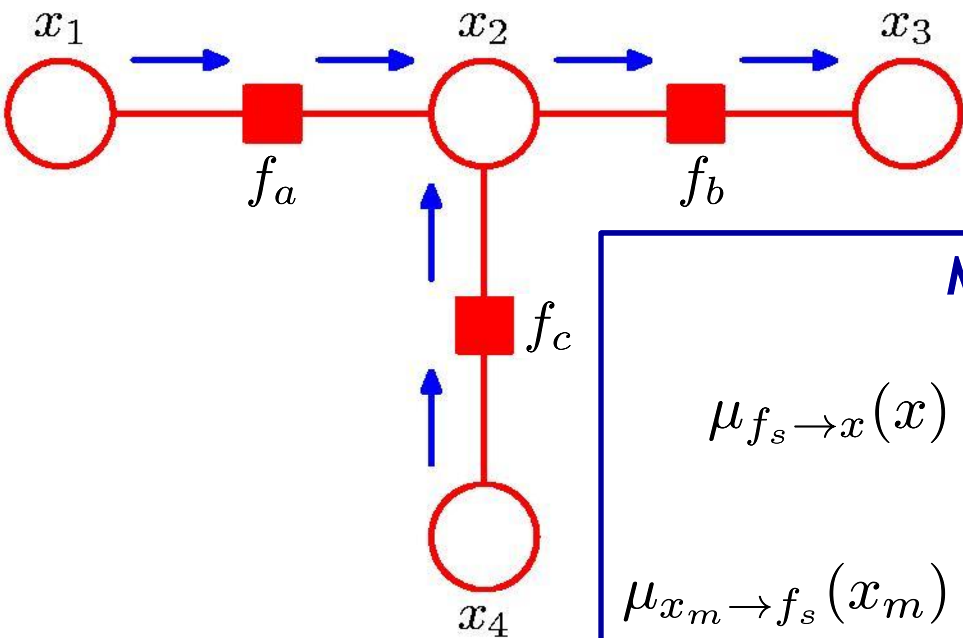
  - Message from variable node to factor node:
    - Product of incoming messages

$$\mu_{x_m \to f_s}(x_m) \equiv \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$$



$\Rightarrow$ Simple propagation scheme.

Recap: Sum-Product from Leaves to Root

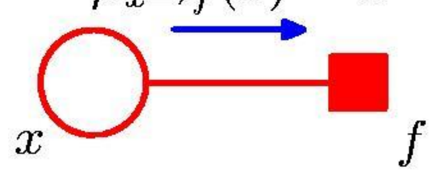


Message definitions:

$$\mu_{f_s \to x}(x) \equiv \sum_{X_s} f_s(\mathbf{x}_s) \prod_{m \in \mathrm{ne}(f_s) \setminus x} \mu_{x_m \to f_s}(x_m)$$

$$\mu_{x_m \to f_s}(x_m) \equiv \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$$

$$\mu_{x \to f}(x) = 1$$



$$\mu_{f \to x}(x) = f(x)$$

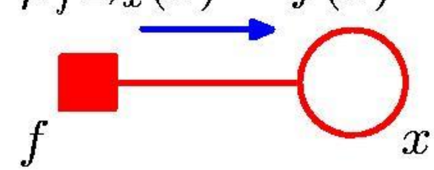
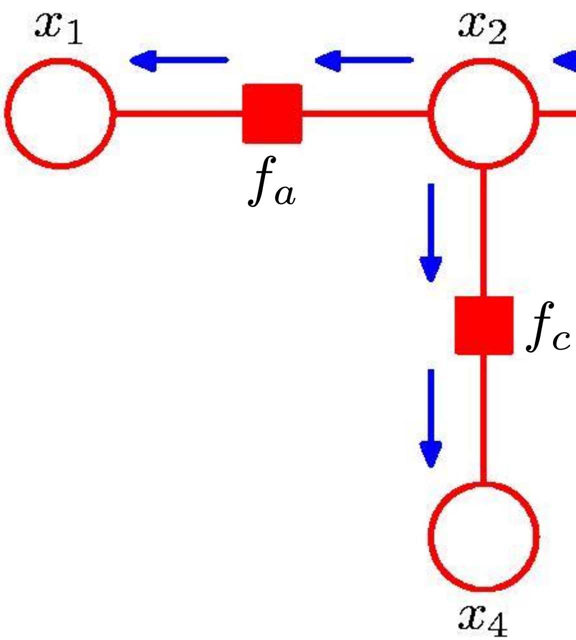


Image source: C. Bishop, 2006

Recap: Sum-Product from Root to Leaves

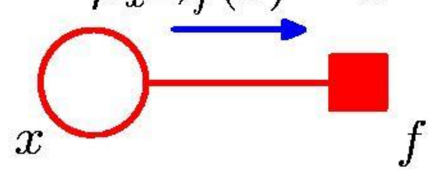


Message definitions:

$$\mu_{f_s \to x}(x) \equiv \sum_{X_s} f_s(\mathbf{x}_s) \prod_{m \in \mathrm{ne}(f_s) \setminus x} \mu_{x_m \to f_s}(x_m)$$

$$\mu_{x_m \to f_s}(x_m) \equiv \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$$

$$\mu_{x \to f}(x) = 1$$



$$\mu_{f \to x}(x) = f(x)$$

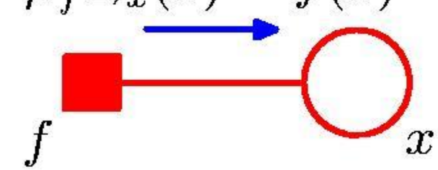


Image source: C. Bishop, 2006

Recap: Max-Sum Algorithm

- Objective: an efficient algorithm for finding
  - Value $\mathbf{x}^{\max}$ that maximises $p(\mathbf{x})$;
  - Value of $p(\mathbf{x}^{\max})$.

  ⇒ Application of dynamic programming in graphical models.

- Key ideas
  - We are interested in the maximum value of the joint distribution

$$p(\mathbf{x}^{\max}) = \max_{\mathbf{x}} p(\mathbf{x})$$

  ⇒ Maximize the product $p(\mathbf{x})$.

  - For numerical reasons, use the logarithm.

$$\ln\left(\max_{\mathbf{x}} p(\mathbf{x})\right) = \max_{\mathbf{x}} \ln p(\mathbf{x}).$$

  ⇒ Maximize the sum (of log-probabilities).

Recap: Max-Sum Algorithm

- **Initialization (leaf nodes)**

$$\mu_{x\to f}(x) = 0 \qquad \mu_{f\to x}(x) = \ln f(x)$$

- **Recursion**
  - Messages

$$\mu_{f\to x}(x) = \max_{x_1,\ldots,x_M}\left[\ln f(x,x_1,\ldots,x_M) + \sum_{m\in\mathrm{ne}(f_s)\setminus x}\mu_{x_m\to f}(x_m)\right]$$

$$\mu_{x\to f}(x) = \sum_{l\in\mathrm{ne}(x)\setminus f}\mu_{f_l\to x}(x)$$

  - For each node, keep a record of which values of the variables gave rise to the maximum state:

$$\phi(x) = \arg\max_{x_1,\ldots,x_M}\left[\ln f(x,x_1,\ldots,x_M) + \sum_{m\in\mathrm{ne}(f_s)\setminus x}\mu_{x_m\to f}(x_m)\right]$$

Recap: Max-Sum Algorithm

- **Termination (root node)**
  - **Score of maximal configuration**

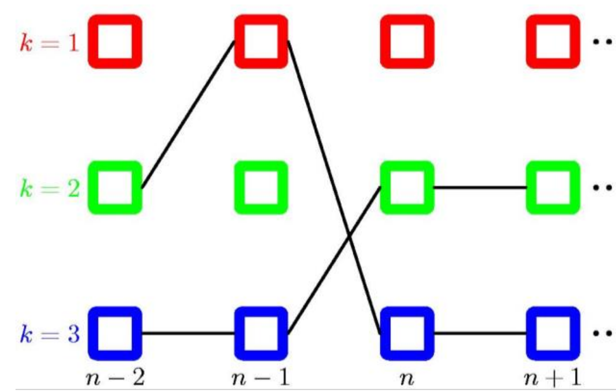
$$p^{\max} = \max_{x}\left[\sum_{s\in\mathrm{ne}(x)} \mu_{f_s\to x}(x)\right]$$

  - **Value of root node variable giving rise to that maximum**

$$x^{\max} = \arg\max_{x}\left[\sum_{s\in\mathrm{ne}(x)} \mu_{f_s\to x}(x)\right]$$

  - **Back-track to get the remaining variable values**

$$x_{n-1}^{\max} = \phi(x_n^{\max})$$



Topics of This Lecture

- Pictorial Structures
  - Model components
  - Prior
  - Likelihood Model
- Recap: Inference
  - Sum-Product algorithm
  - Max-Sum algorithm
- Efficient Inference in Pictorial Structures
  - Generalized Distance Transform
  - Effect on Computation
- Results

Efficient Inference

- **Best location given by MAP**

$$\max_L p(L|E) = \max_L \prod_{i=0}^{N} (p(l_i|l_0)p(e_i|l_i))$$

$$= \min_L \sum_{i=0}^{N} (-\ln p(l_i|l_0) - \ln p(e_i|l_i))$$

  - **Consider case of 2 parts**

$$\min_{l_0,l_1} (-\ln p(e_0|l_0) - \ln p(e_1|l_1) - \ln p(l_1|l_0))$$

  - **Rename things**

$$= \min_{l_0,l_1} (m_0(l_0) + m_1(l_1) + d(l_1, l_0))$$

Efficient Inference

- **Assume d to have quadratic form**

$$d(l_1, l_0) = ||l_1 - T_1(l_0)||^2$$

- **Then**

$$\min_{l_0, l_1} (m_0(l_0) + m_1(l_1) + d(l_1, l_0))$$

$$= \min_{l_0} \left(m_0(l_0) + \min_{l_1} (m_1(l_1) + d(l_1, l_0)) \right)$$

  - **with the second term a generalized distance transform (gDT).**
  - **Algorithms exist to compute gDT efficiently.**
  - **Thus** $= \min_{l_0} (m_0(l_0) + DT_{m_1}(T_1(l_0)))$

$\Rightarrow$ **Finding the best part configuration can be done sequentially, rather than simultaneously!**

Distance Transform

- **Given points $p \in P$ on a grid (e.g., image) G**
  - **Distance Transform associates to each location $x \in G$ the distance to the nearest point $p \in P$**

$$DT_P(x) = \min_{p \in P} \{d(x,p)\}$$

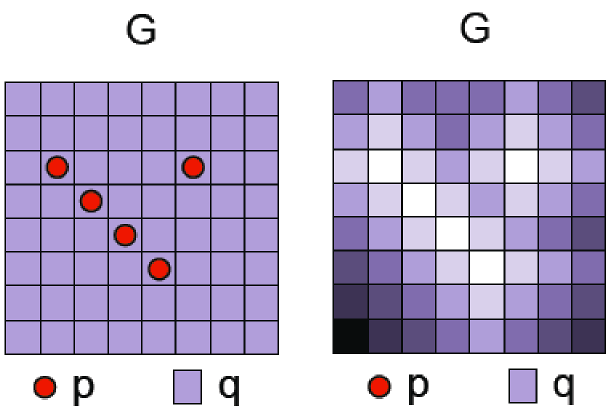
  - **or equivalent**

$$DT_P(x) = \min_{q \in G} \{d(x,q) + 1(q)\} \qquad 1(q) = \begin{cases} 0 & \text{if } q \in P \\ \infty & \text{otherwise} \end{cases}$$

- **Example**

$$d(x,q) = |x - q|$$

$$DT_P(x) = \min_{q \in G} \{|x - q| + 1(q)\}$$



Generalized Distance Transform

- Replace binary function $1(q)$ with general function $f(q)$

$$DT_f(x) = \min_{q \in G} \{d(x,q) + f(q)\}$$

  - We can assign "soft membership of all grid elements to P.
  - $f(q)$ is sampled on the entire grid G.

- In our case
  - f corresponds to m_1.
  - Distance corresponds to $d(l_1, l_0) = ||l_1 - T_1(l_0)||^2$

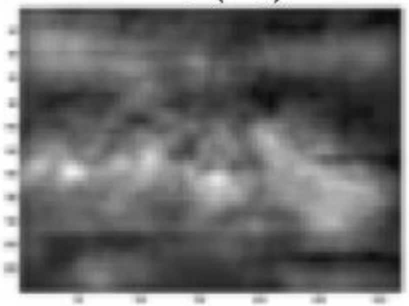
$$DT_{m_1}(T_1(l_0)) = \min_{l_1} \{m_1(l_1) + d(l_1, l_0)\}$$

Example: Part Model of Motorbikes

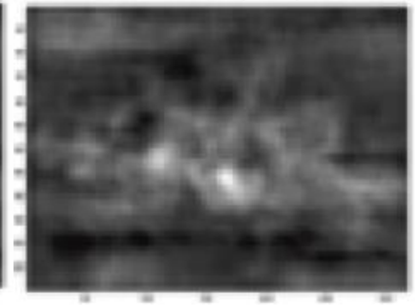
- **Model**
  - 2 parts (use both wheels), simple translation between them given by (x,y) position

1. **Part unaries (log prob)**
   – $m_0(l_0)$ and $m_1(l_1)$

$m_0(l_0)$



$m_1(l_1)$

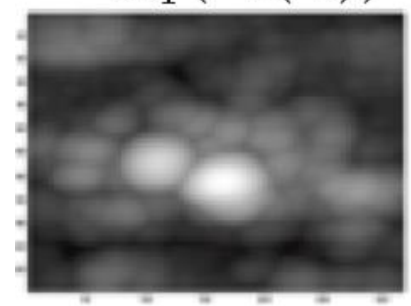


2. **Distance transform of $m_1(l_1)$**

3. **Simply find minimum of sum**

$$\min_{l_0} \left(m_0(l_0) + DT_{m_1}(T_1(l_0))\right)$$

$DT_{m_1}(T_1(l_0))$



Example from Dan Huttenlocher

Topics of This Lecture

- Pictorial Structures
  - Model components
  - Prior
  - Likelihood Model
- Recap: Inference
  - Sum-Product algorithm
  - Max-Sum algorithm
- Efficient Inference in Pictorial Structures
  - Generalized Distance Transform
  - Effect on Computation
- **Results**

Results

- **Tracking and interpreting detailed body motion.**

D. Ramanan, D.A. Forsyth, A. Zisserman. [Tracking People by Learning their Appearance](), PAMI 2007.

References and Further Reading

- Pictorial Structures
  - P. Felzenszwalb, D. Huttenlocher, Pictorial Structures for Object Recognition, IJCV, Vol. 61(1), 2005.

- Human Body Pose Estimation with Pictorial Structures
  - D. Ramanan, D. A. Forsyth, A. Zisserman. Tracking People by Learning their Appearance, IEEE Trans. PAMI., 2007.
  - V. Ferrari, M. Marin, A. Zisserman, Progressive Search Space Reduction for Human Pose Estimation, CVPR, 2008.
  - M. Andriluka, S. Roth and B. Schiele, Discriminative Appearance Models for Pictorial Structures, IJCV, Vol. 99(3), pp.259-280, 2012.