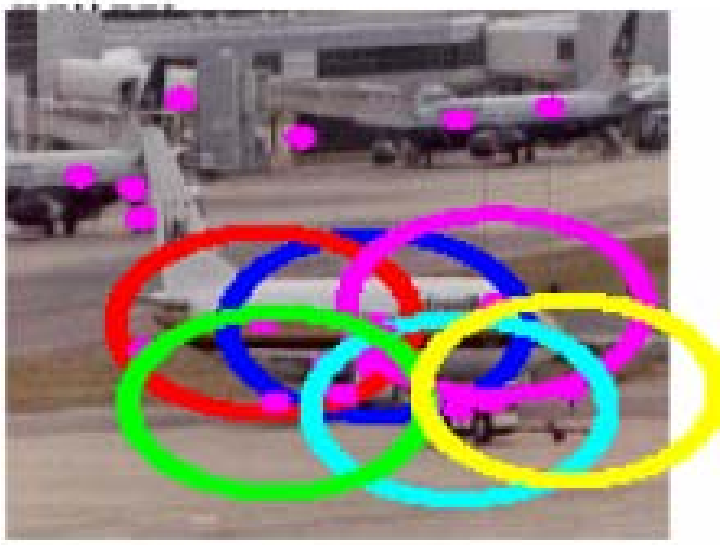
Pictorial Structures for Object Recognition

Pedro F. Felzenszwalb

Presented by Hanlin Tang

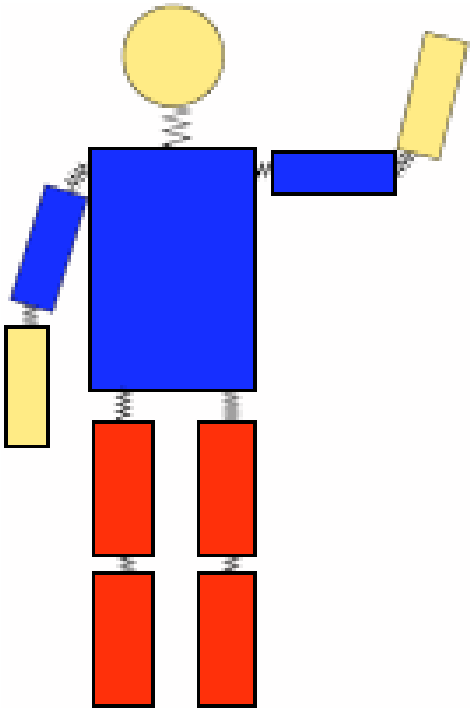
COS/PSY 598b

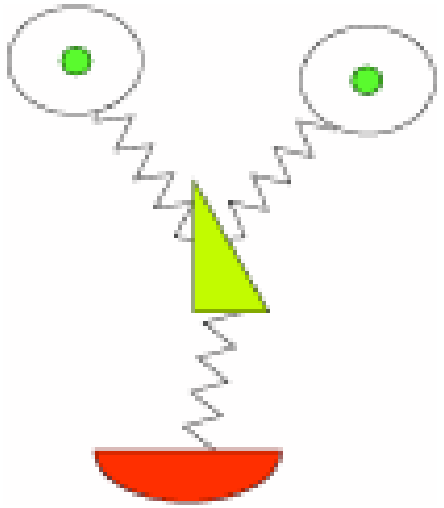
Feature detection



Felzenswab Talk, 2007

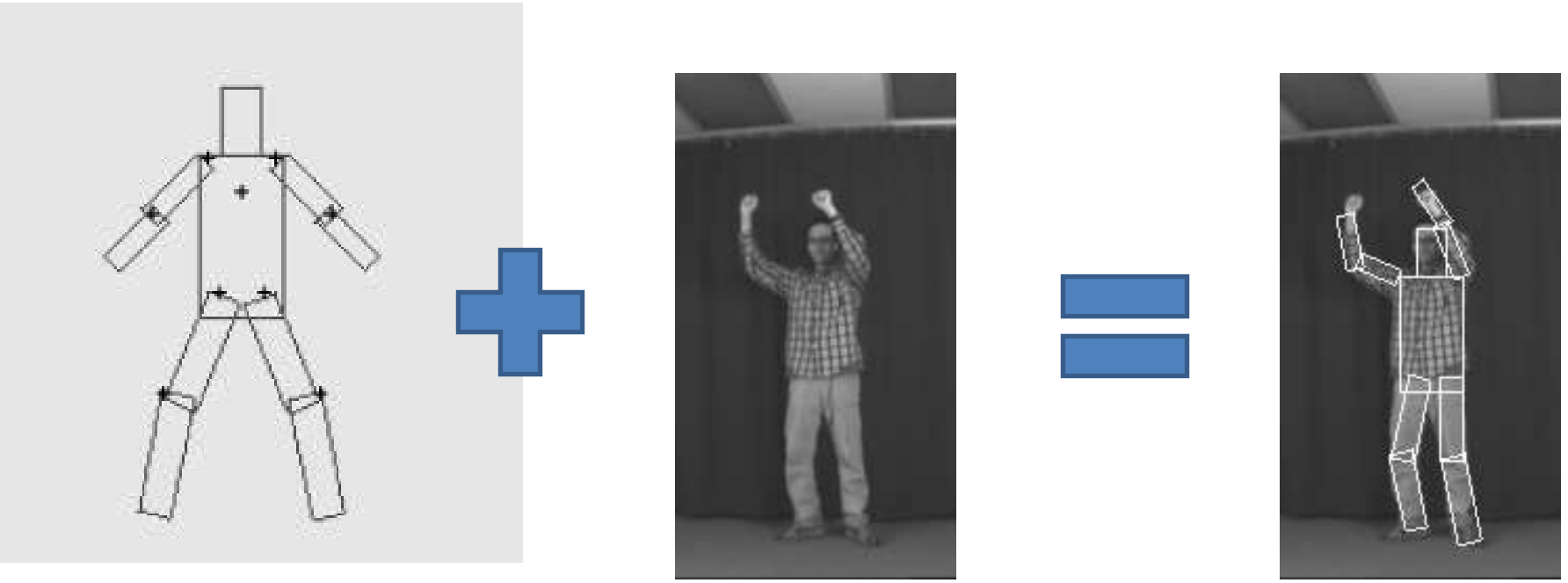
Pictorial Recognition



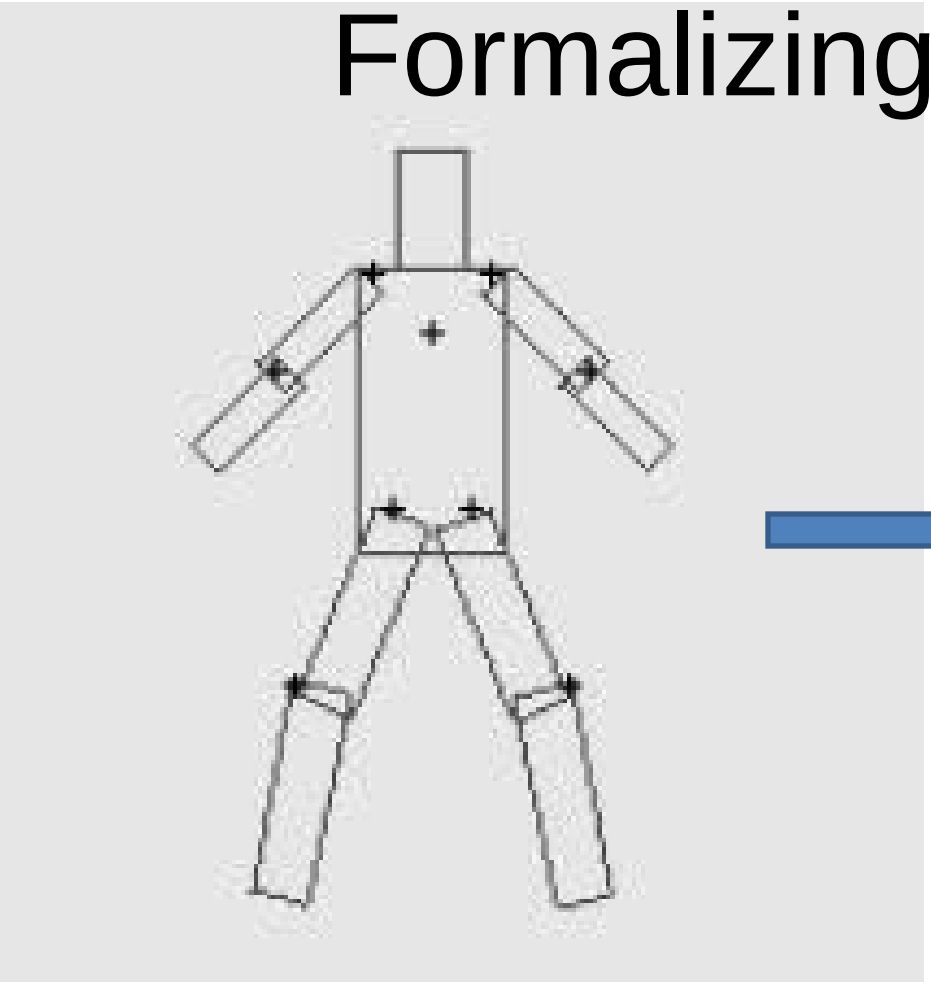


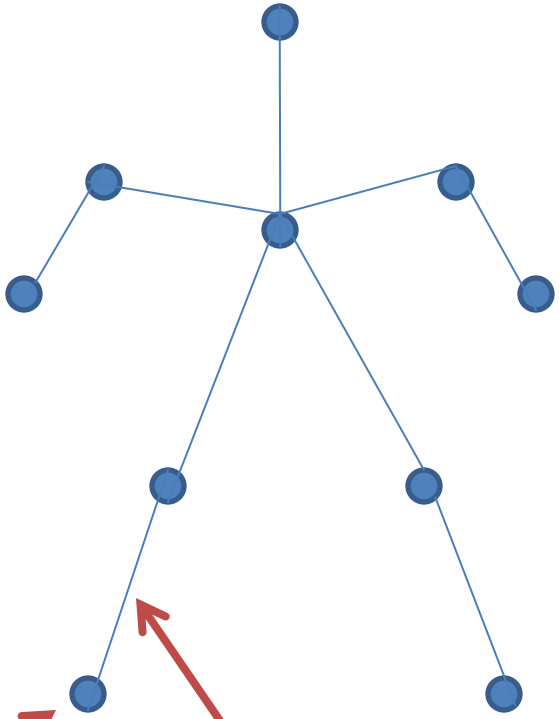
Felzenswab Talk, 2007

The Task



Formalizing this intuition

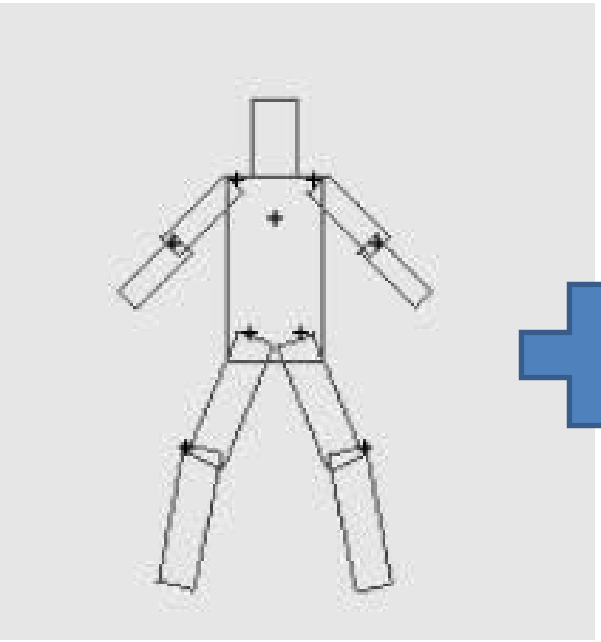


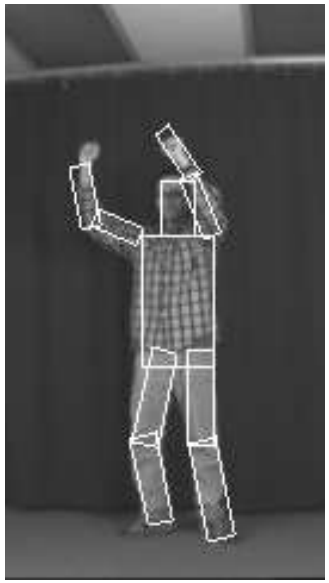


Vertex v_i has $\{u_j\}$

Edge has c_{ij}

The Task





Θ ={*u*,*c*} I L = {l_1, l_2 , l_3 ..

$$L^* = \underset{L}{\operatorname{argmax}}\, p(L|I,\theta)$$

Bayes Rule to the Rescue…

$$p(L|I,\theta) \propto p(I|L,\theta)p(L|\theta)$$

Assuming part independence,

$$P(L|I,\theta) \propto \left(\prod_{i=1}^{n} p(I|l_i,u_i) \prod_{(v_i,v_j)\in E} p(l_i,l_j|c_{ij}) \right)$$

A more intuitive formulation:

$$L^* = \underset{L}{\operatorname{argmax}}\, p(L|I,\theta)$$

$$= \underset{L}{\operatorname{argmin}} -\log[p(L|I,\theta)]$$

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$

Mismatch with image

Deformation Cost
(cost of stretching springs!)

$$m_i(l_i) = -\log p(I|l_i, u_i)$$

$$d_{ij}(l_i, l_j) = -\log p(l_i, l_j|c_{ij})$$

Fischler and Eschlager
(1972)

The Grand Assumption

Want to define $d_{ij}(l_i, l_j)$

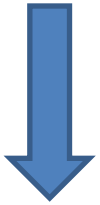
Let: $x = l_1 - l_2$

In 1-dimension: $\left(\frac{x-\mu}{\sigma}\right)^2$

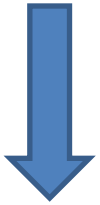
In N-dimensions: $(x-\mu)^T \Sigma^{-1} (x-\mu)$

The Grand Assumption

$$d_{ij}(l_i, l_j) = \overbrace{(T_{ij}(l_i) - T_{ji}(l_j))}^{x}{}^T \mathbf{D}_{ij}^{-1} (T_{ij}(l_i) - T_{ji}(l_j))$$

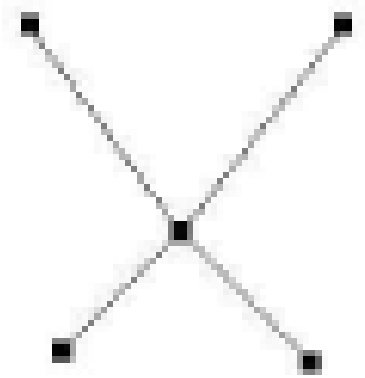


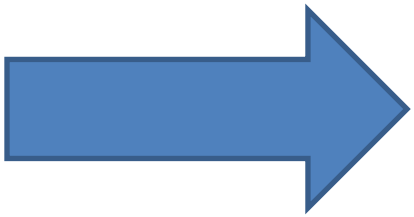
$$d_{ij}(l_i, l_j) = -\log p(l_i, l_j | c_{ij})$$

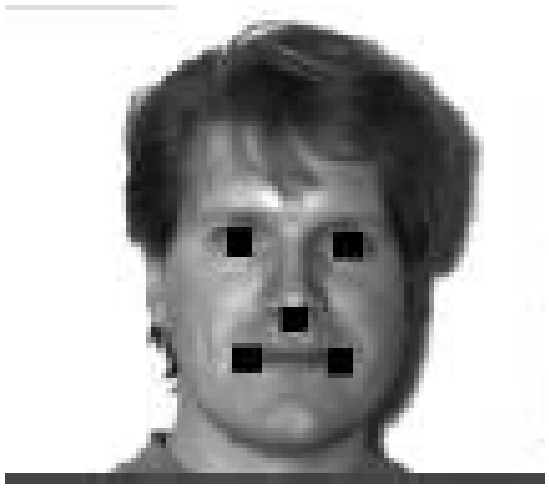


$$p(l_i, l_j | c_{ij}) = \mathcal{N}(T_{ij}(l_i) - T_{ji}(l_j), 0, D_{ij})$$

Iconic Models - Faces







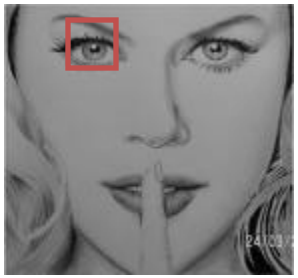
$\Theta = \{u, c\}$

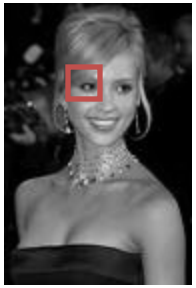
$L = \{l_1, l_2, l_3 ..$

$$L^* = \underset{L}{\operatorname{argmax}}\, p(L|I, \theta)$$

What an eye should look like

- The naïve way:



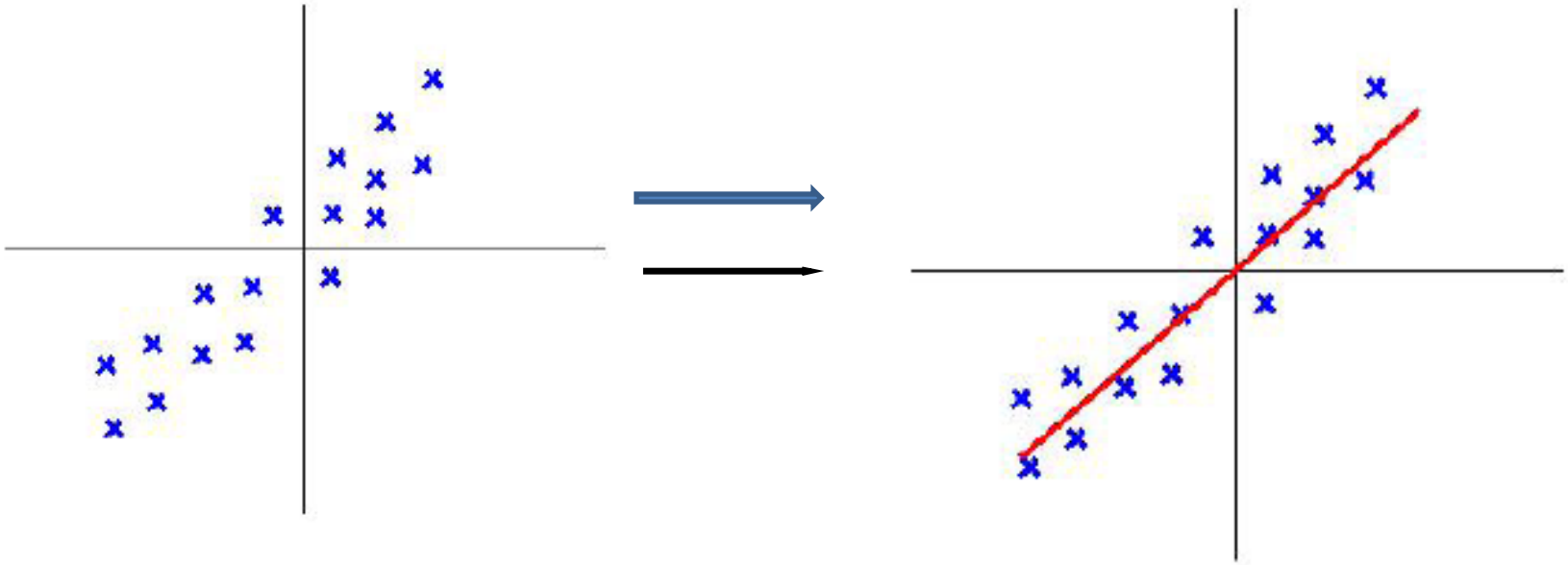




$$u = \{\mu, \sigma\}$$

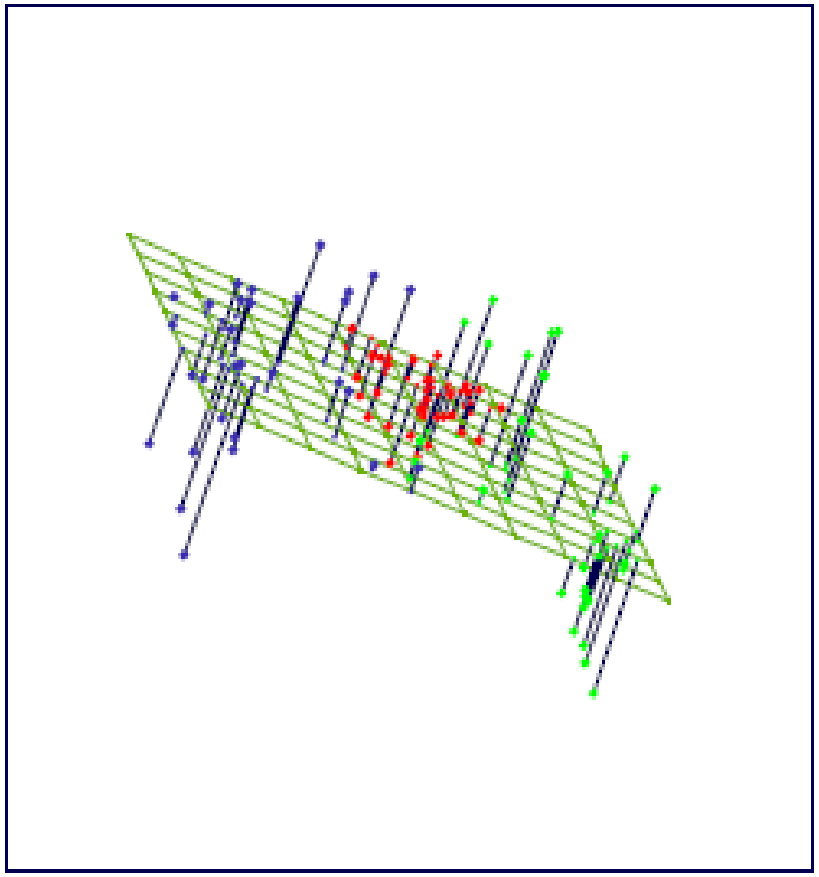
$$\mu = \begin{pmatrix} 0.2 & 0.4 & 0.4 & 0.4 & 0.2 \\ 0.1 & 0.5 & 0.6 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.9 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.8 & 0.4 & 0.1 \\ 0.2 & 0.3 & 0.3 & 0.3 & 0.2 \end{pmatrix}$$

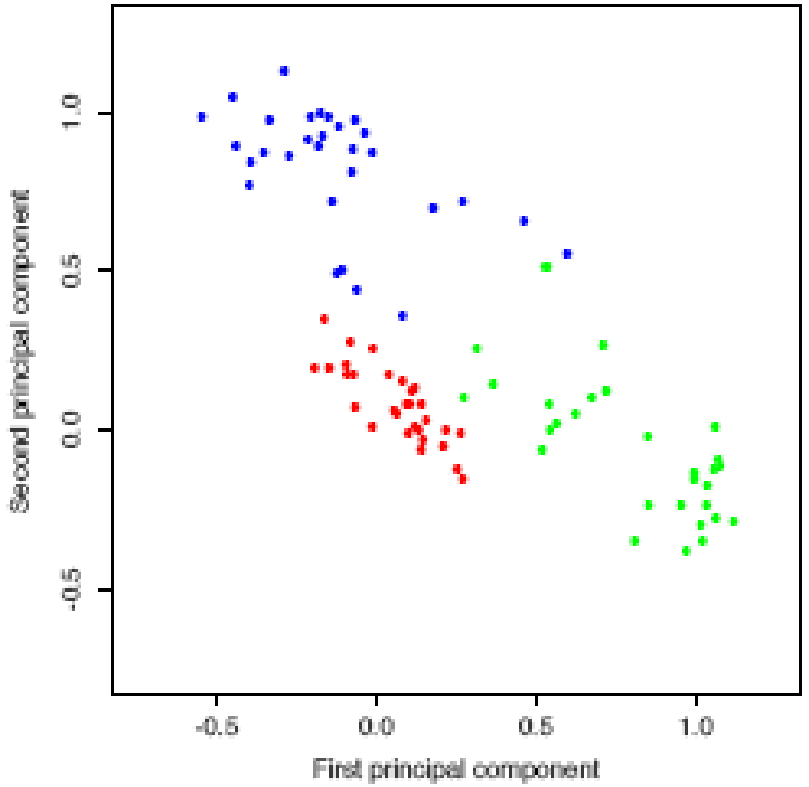
But can reduce dimensions!

- The intuition:



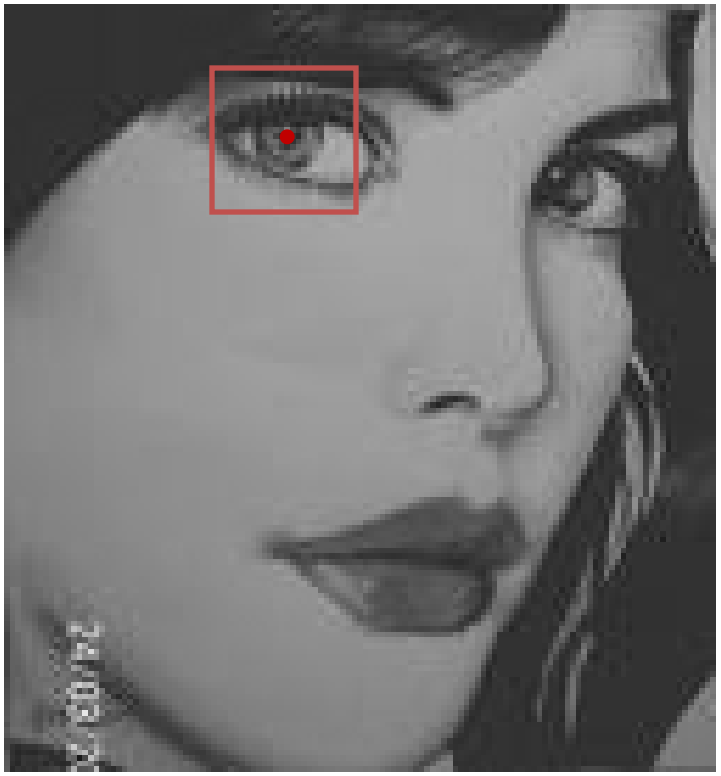
More intuition-building:



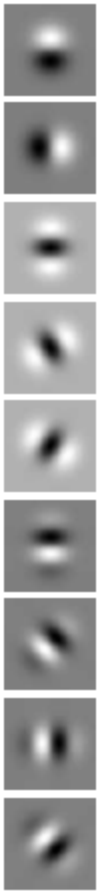


COS424 notes

Applied to natural data



$\otimes$

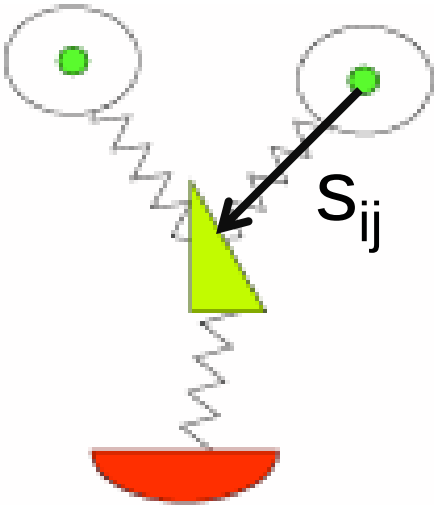


$= \begin{pmatrix} 0.2 \\ 0.1 \\ 0.7 \\ 0.9 \\ 0.4 \\ 0.6 \\ 0.2 \\ 0.5 \\ 0.6 \end{pmatrix} = \alpha(l_i)$

$$u_i = (\mu_i, \Sigma_i)$$

$$p(I \mid l_i, u_i) \propto \mathrm{N}(\alpha(l_i), u_i, \Sigma_i)$$

Characterizing springs



$$c_{ij} = (s_{ij}, \Sigma_{ij})$$

$$p(l_i, l_i \mid c_{ij}) = \mathrm{N}(l_i - l_j, s_{ij}, \Sigma_{ij})$$

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$

Seek models to define:

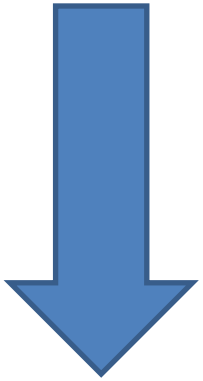
- Location
- Appearance
- Connections
- $p(I|l_i, u_i)$
- $p(l_i, l_j|c_{ij})$

$$l_i = (x, y)$$
$$u_i = (\mu_i, \Sigma_i)$$
$$c_{ij} = (s_{ij}, \Sigma_{ij})$$

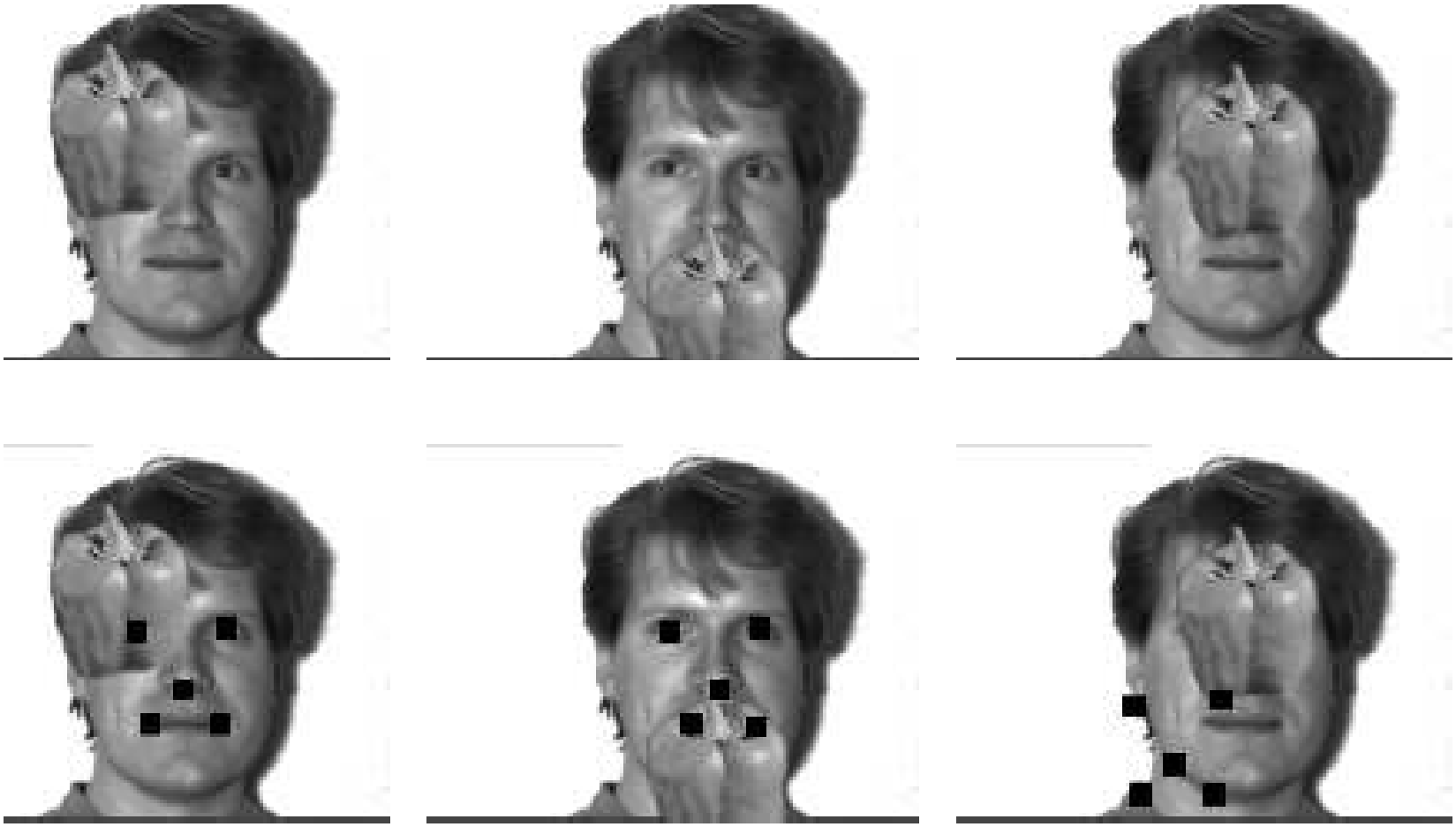
$$p(I \mid l_i, u_i) \propto \mathrm{N}(\alpha(l_i), u_i, \Sigma_i)$$
$$p(l_i, l_i \mid c_{ij}) = \mathrm{N}(l_i - l_j, s_{ij}, \Sigma_{ij})$$

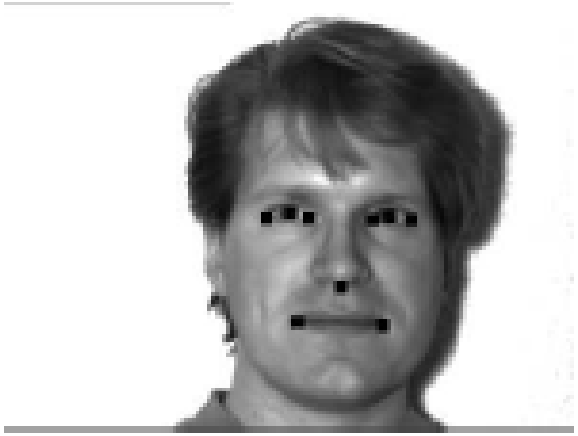
Some Math:

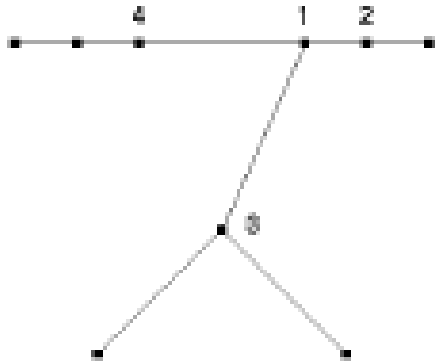
$$p(l_i, l_i \mid c_{ij}) = \mathrm{N}(l_i - l_j, s_{ij}, \Sigma_{ij})$$

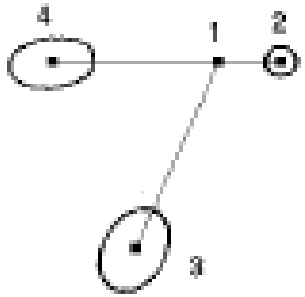


$$p(l_i, l_j | c_{ij}) = \mathcal{N}(T_{ij}(l_i) - T_{ji}(l_j), 0, D_{ij})$$

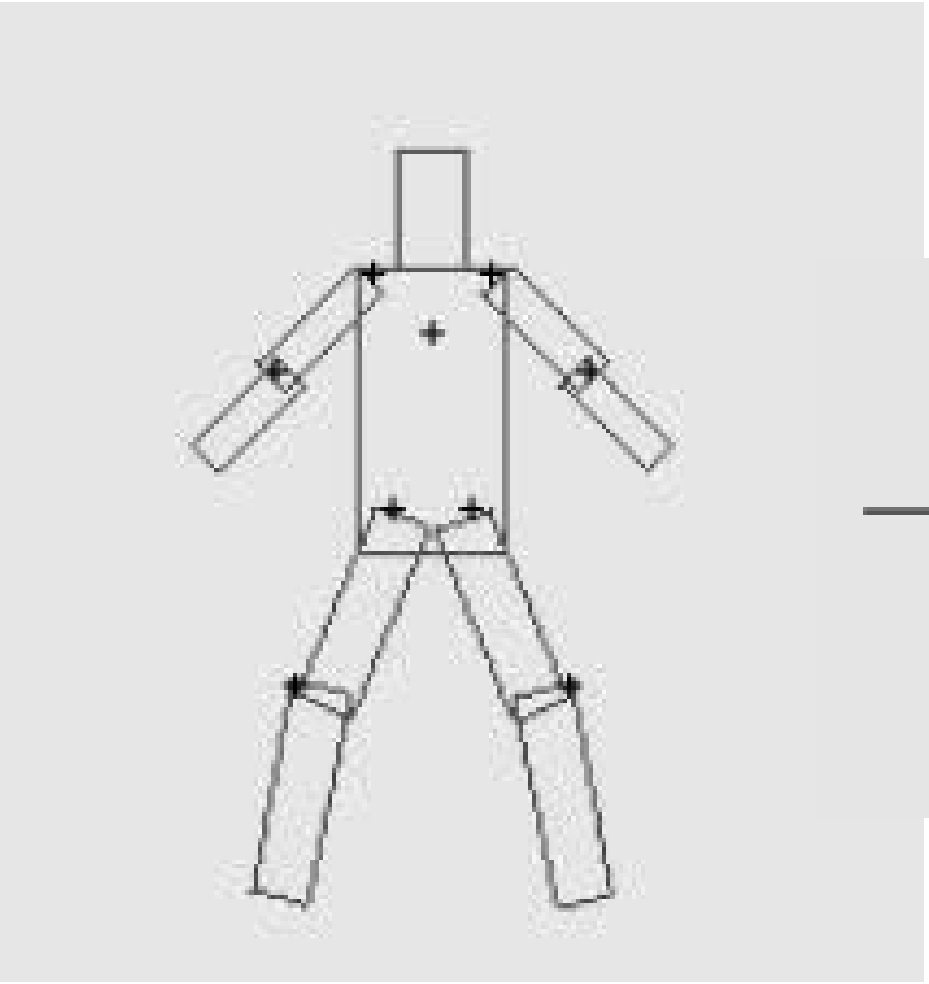


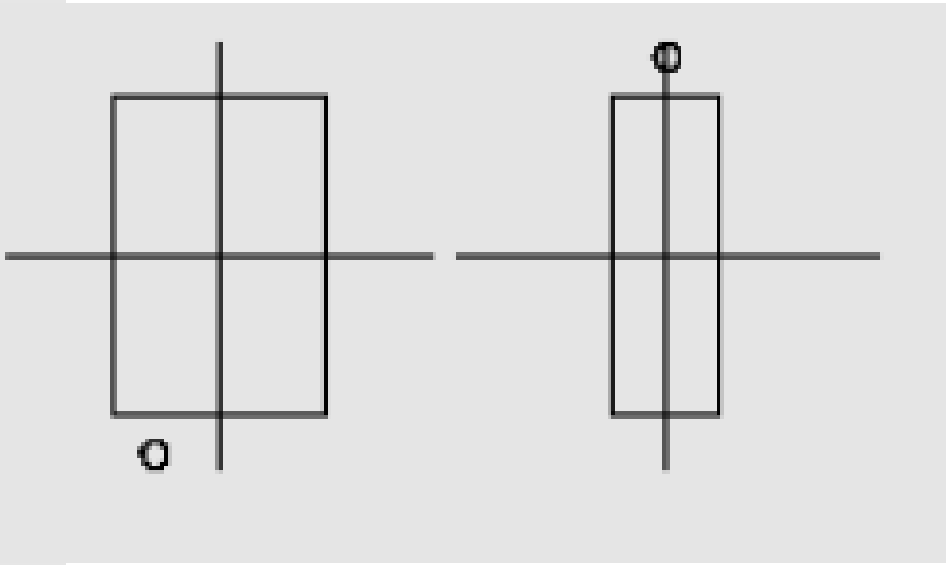



4
1
2
3


4
1
2
3

Articulated Models - Humans





$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$

Seek models to define:

- Location
- Appearance
- Connections
- $p(I|l_i, u_i)$
- $p(l_i, l_j|c_{ij})$

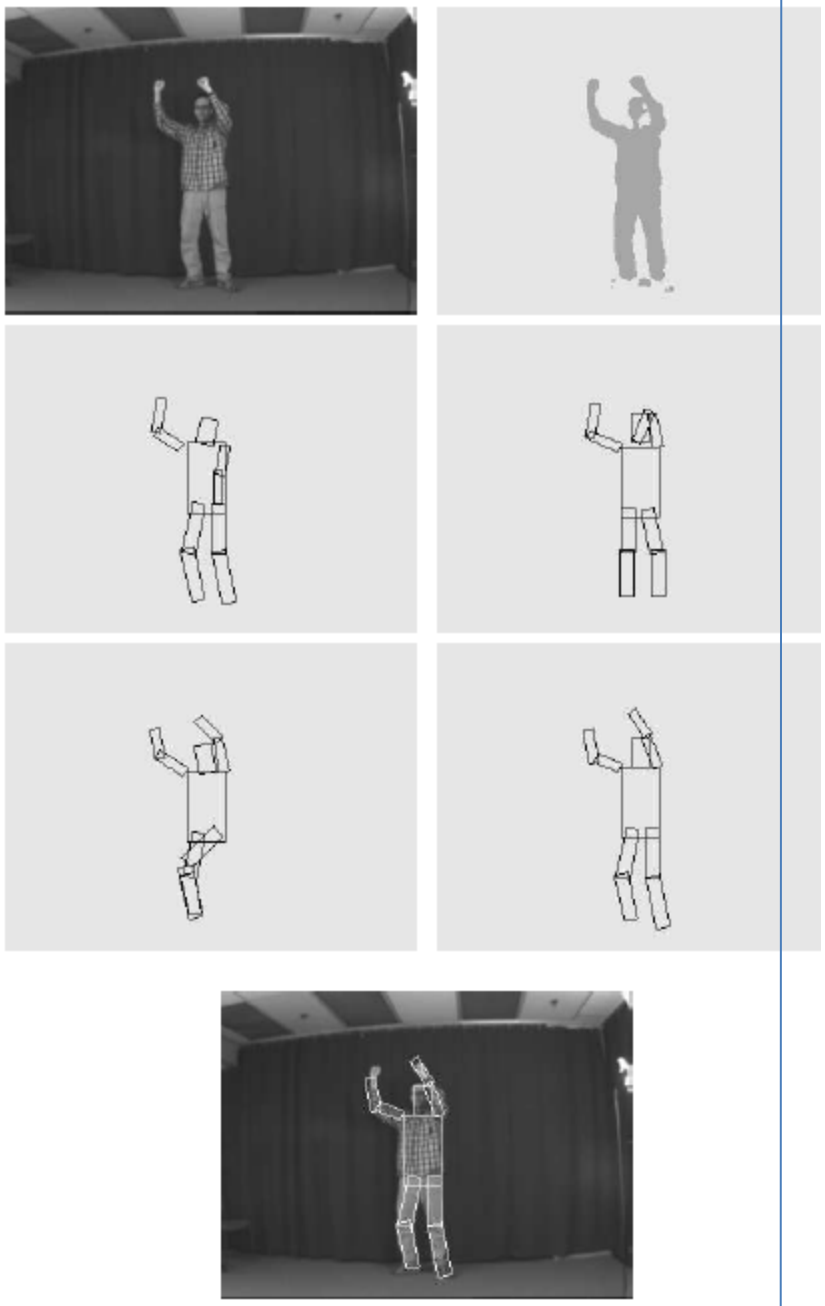
$$l_i = (x, y, s, \theta)$$

$$u_i = (q_1, q_2)$$

$$c_{ij} = (x_{ij}, y_{ij}, x_{ji}, y_{ji}, \sigma_x, \sigma_y, \sigma_s, \theta_{ij}, k)$$

$$p(I \mid l_i, u_i) = {q_1}^{n_1} (1-q_1)^{n'_1} {q_2}^{n_2} (1-q_2)^{n'_2} (0.5)^{T-A_1-A_2}$$

$$p(l_i, l_i \mid c_{ij}) \propto \mathrm{N}(T_{ji}(l_i) - T_{ij}(l_j), 0, D_{ij})$$



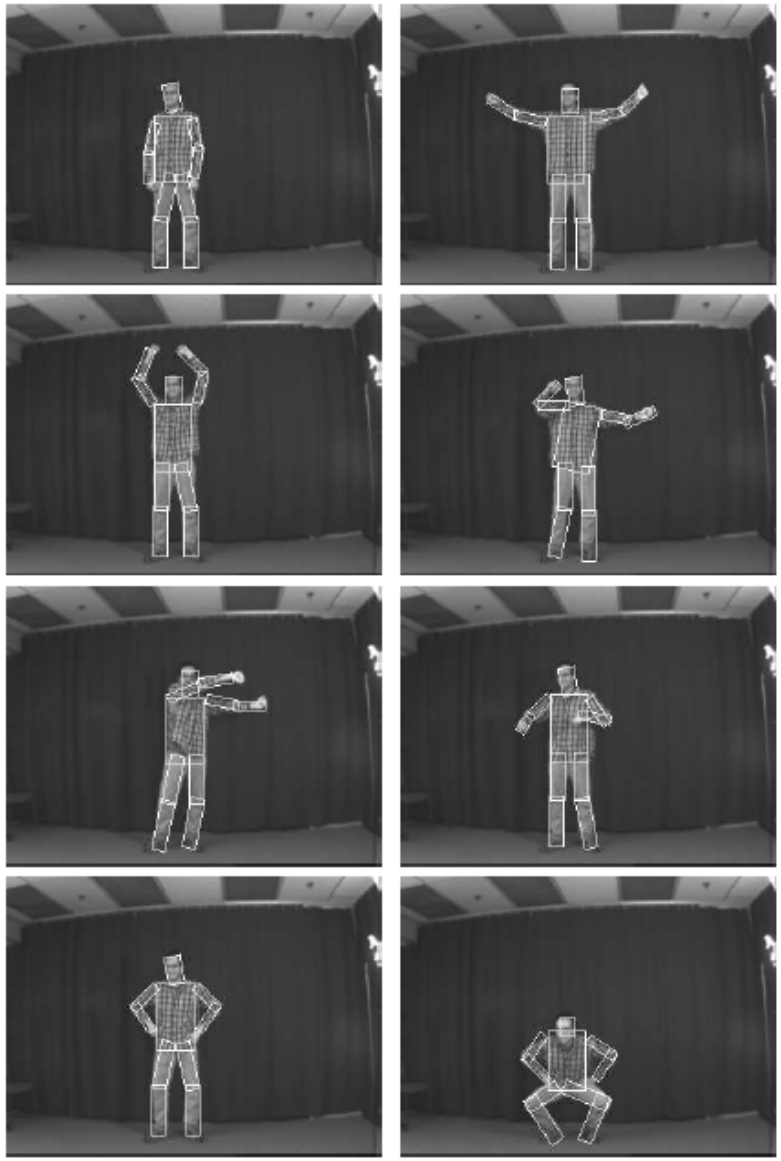


Figure 13: Matching results (sampling 200 times).

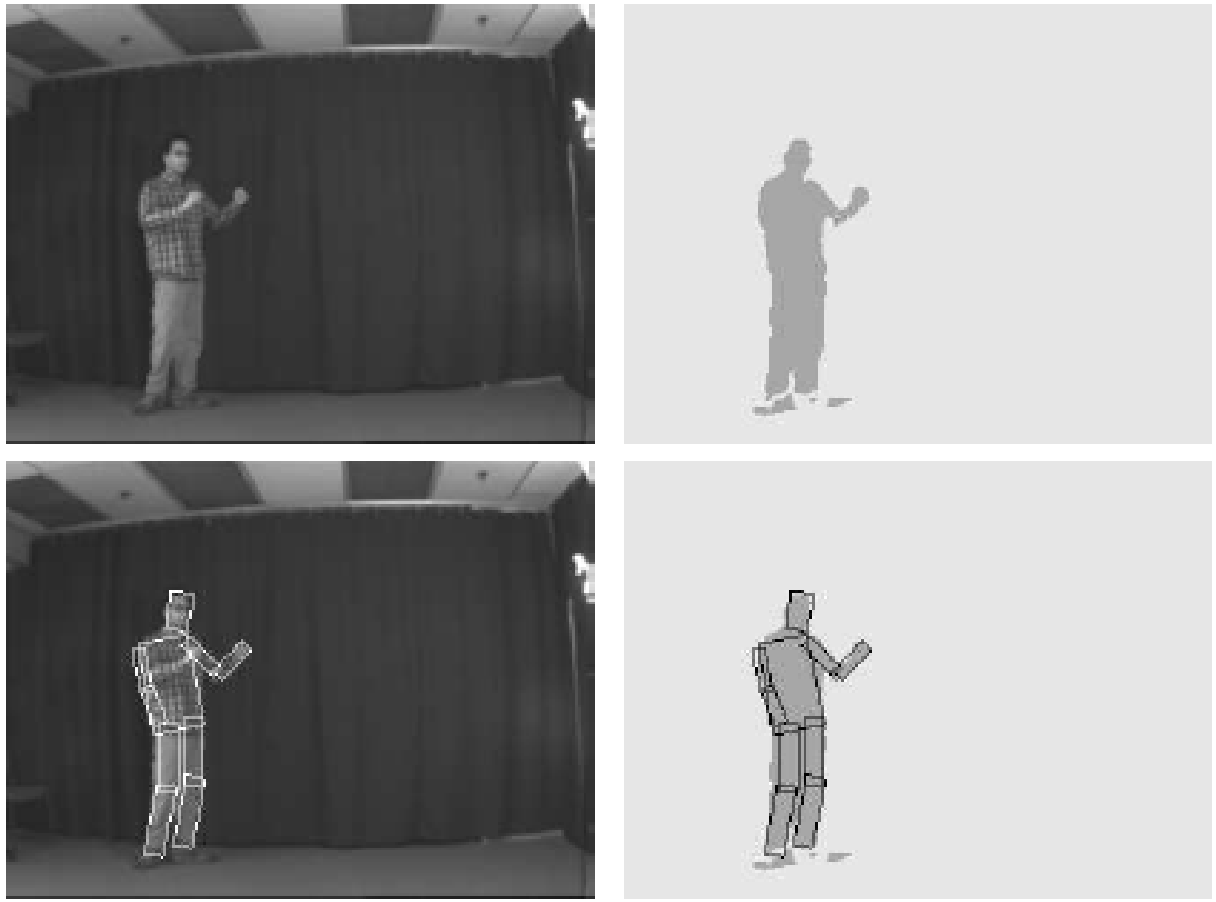


Figure 14: In this case, the binary image doesn't provide enough information to estimate the position of one arm.

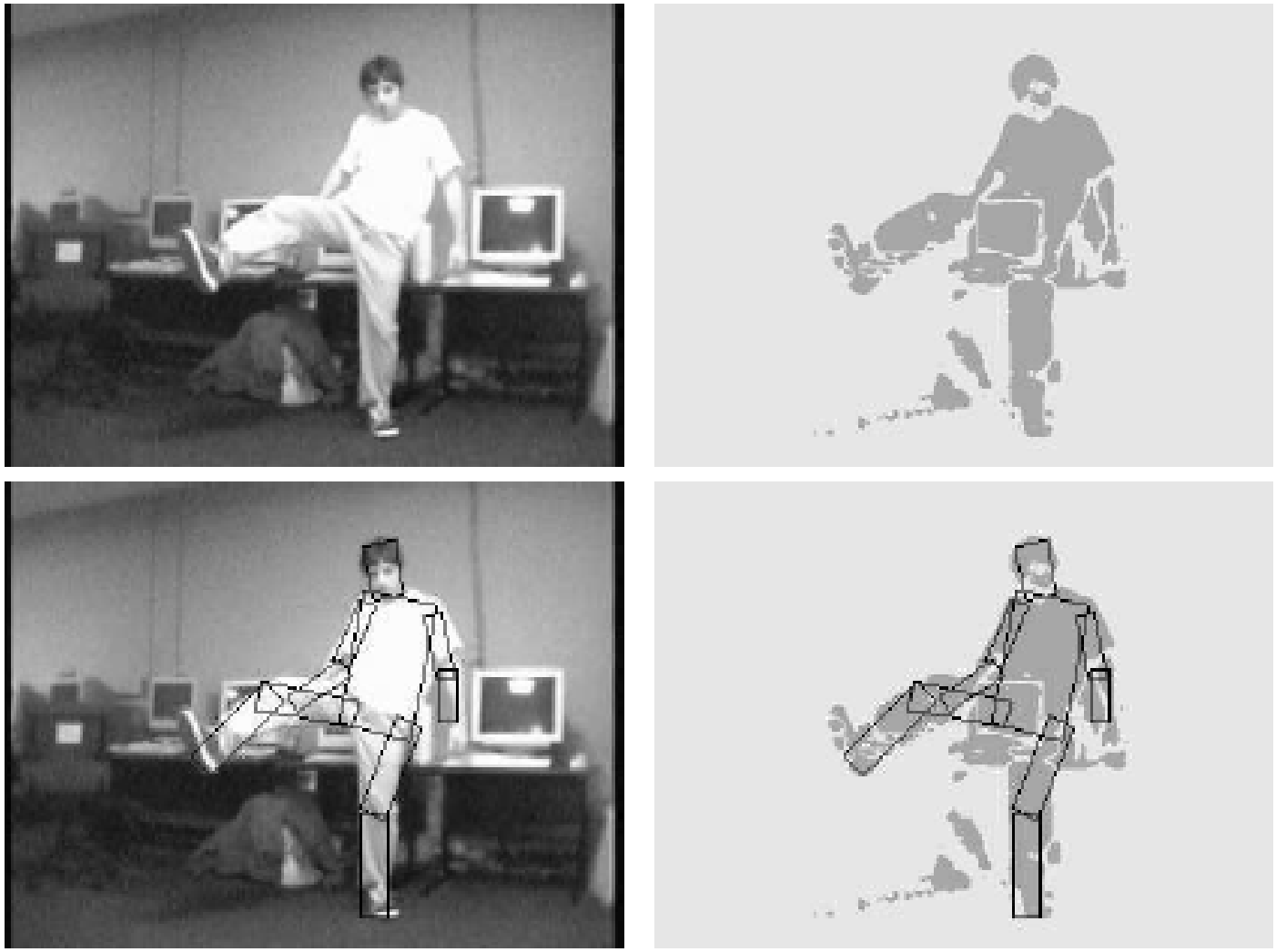
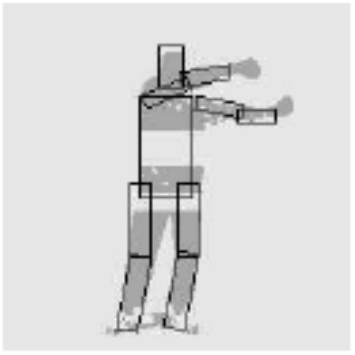
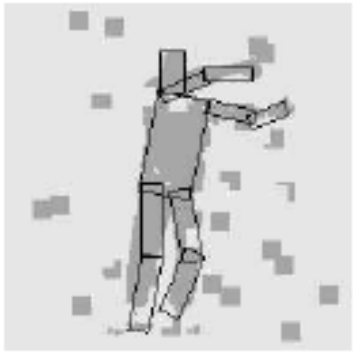
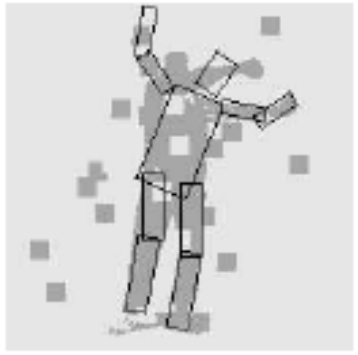


Figure 15: This example illustrates how our method works well with noisy images.

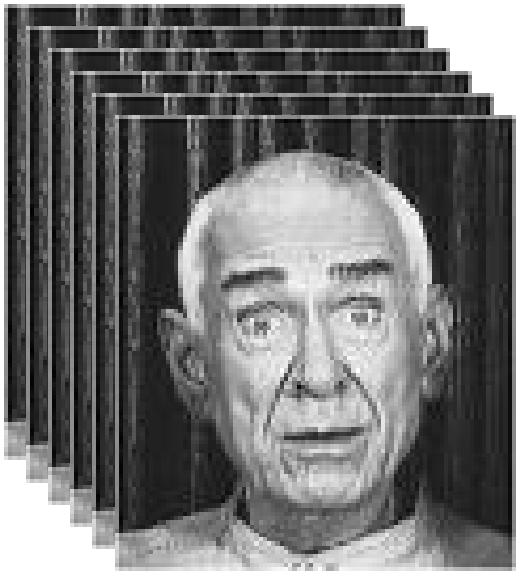




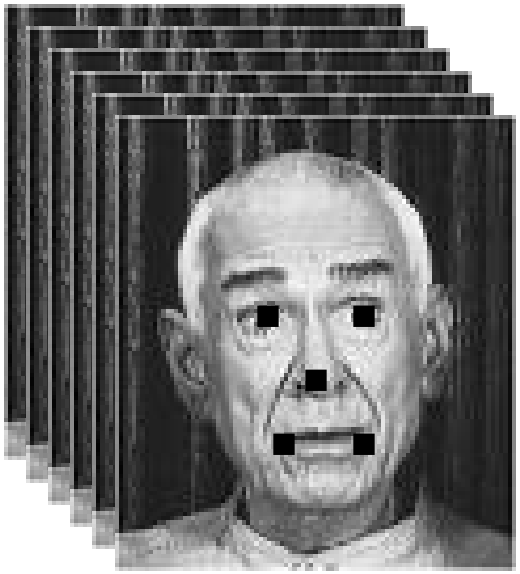


- Restriction on d_{ij} allows linear running time for Finding L*
- Efficient ways to sample from the Posterior $p(L|I, \theta)$

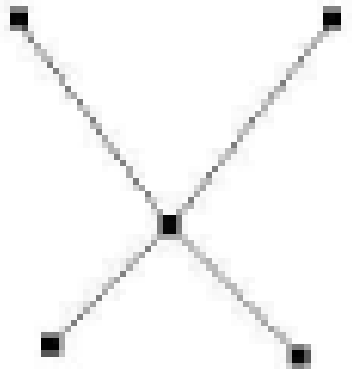
Learning



$$I = \{I^1, I^2, ..\}$$



$$L = \{L^1, L^2, ..\}$$



$$\Theta = \{u, c\}$$

Learning the Model

$$p(I^1, \ldots, I^m, L^1, \ldots, L^m|\theta) = \prod_{k=1}^{m} p(I^k, L^k|\theta),$$

$$\theta^* = \arg\max_{\theta} \prod_{k=1}^{m} p(I^k|L^k, \theta) \prod_{k=1}^{m} p(L^k|\theta).$$

Learning Appearance and Connections

$$u_i^* = \arg\max_{u_i} \prod_{k=1}^{m} p(I^k | l_i^k, u_i).$$

$$c_{ij}^* = \arg\max_{c_{ij}} \prod_{k=1}^{m} p(l_i^k, l_j^k | c_{ij}).$$

$$q(v_i, v_j) = \prod_{k=1}^{m} p(l_i^k, l_j^k | c_{ij}^*).$$

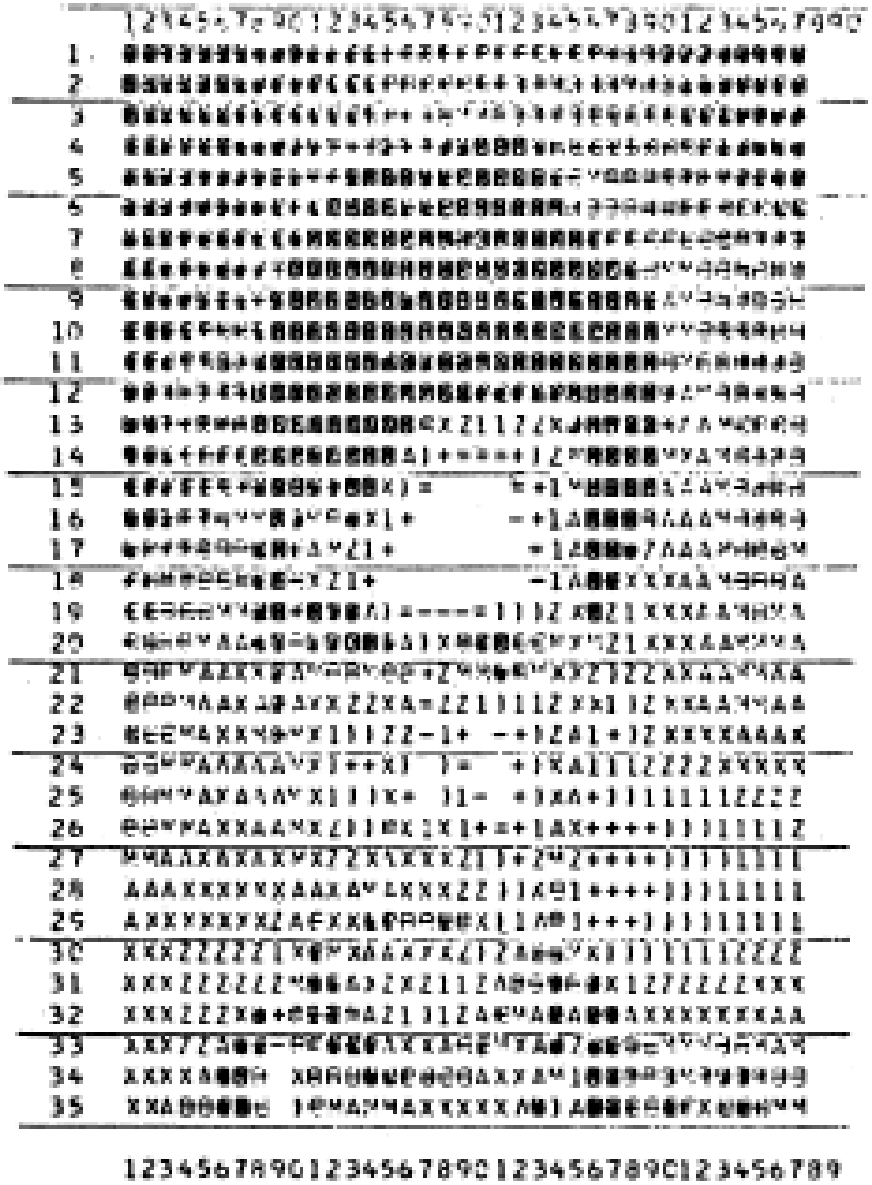
$$E^* = \arg\max_{E} \prod_{(v_i,v_j)\in E} q(v_i, v_j) = \arg\min_{E} \sum_{(v_i,v_j)\in E} -\log q(v_i, v_j).$$

General Framework

Felzenswab Talk, 2007

Key Points

- Pictorial models bring context to recognition
- Robust to noise, scale, lighting effects
- Generalized structure
- Heavily dependent on prior training of model
- Require robust definitions of (u,v)



Original picture.

HAIR WAS LOCATED AT (11, 21)
L/EDGE WAS LOCATED AT (25, 11)
R/EDGE WAS LOCATED AT (25, 24)
L/EYE WAS LOCATED AT (21, 15)
R/EYE WAS LOCATED AT (21, 21)
NOSE WAS LOCATED AT (26, 18)
MOUTH WAS LOCATED AT (29, 17)

Fischler and Eschlager (1972)