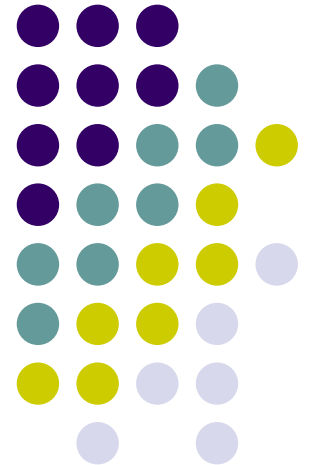


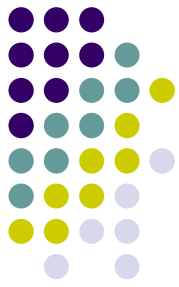
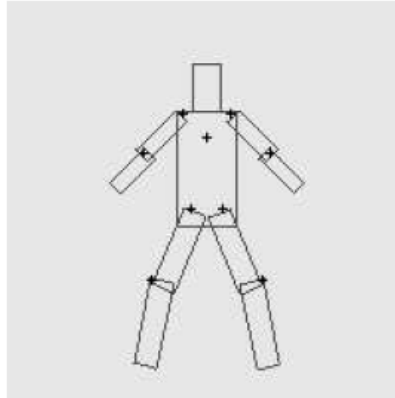
# Pictorial Structures for Object Recognition

Felzenszwalb and Huttenlocher

Presented by Stephen Krotosky

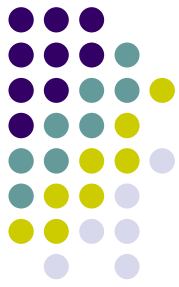


# Pictorial Structures

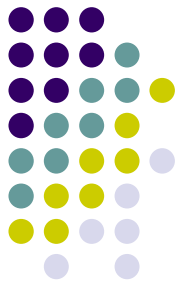


- Introduced by Fischler and Elschlager in 1973
- Objects are modeled by a collection of parts in a deformable configuration
- Best match is found by minimizing function that measures both individual match costs and connection costs

# Pictorial Structures Shortcomings Addressed in Paper



1. Resulting energy minimization problem is hard to solve efficiently
  - Develop efficient algorithm for solving when connections are acyclic and of certain non-general type
2. The model has many parameters
  - Method for learning all model parameters, even interpart connections
3. It is often desirable to find more than a single best match
  - Develop techniques for finding multiple good hypotheses rather than a single best estimate



# Pictorial Structures Definitions

- A collection of parts with connections between certain pairs.
- Best match depends on how well each part matches at its location and how well the locations agree with the deformable model
- Matching a pictorial structure does not involve making any decisions about location of individual parts
- Find a **global minimum** of energy function **without any initialization!**

Undirected Graph:

$$\mathbf{G} = (\mathbf{V}, \mathbf{E})$$

$\mathbf{V} = \{v_1, \dots, v_n\}$ , correspond to the  $n$  parts

$(v_i, v_j) \in \mathbf{E}$ , for each pair of connected points

$\mathbf{L} = \{l_1, \dots, l_n\}$ , configuration where  $l_i$  corresponds to  $v_i$ .

Can simply be location or more complicated parameterization

$m_i(l_i)$  is the degree of mismatch when  $v_i$  is placed at  $l_i$

$d_{ij}(l_i, l_j)$  is the function measuring the degree of deformation of the model

when  $v_i$  is placed at  $l_i$  and  $v_j$  is at  $l_j$

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

# Efficient Algorithm Requirements



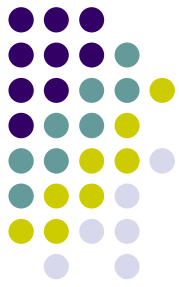
- Graph G be acyclic, i.e. is a tree
  - Can correspond to skeletal structure
  - Star-graphs
  - Enables best match to be computed in polynomial time using Viterbi algorithm
- Second restriction reduces search to essentially linear
  - $d$  is a Mahalanobis distance between transformed locations

$$d_{i,j}(l_i, l_j) = \left( T_{ij}(l_i) - T_{ji}(l_j) \right)^T M_{ij}^{-1} \left( T_{ij}(l_i) - T_{ji}(l_j) \right)$$

$T_{ij}(l_i)$  and  $T_{ji}(l_j)$  are 1 to 1 and are represented as positions on a grid

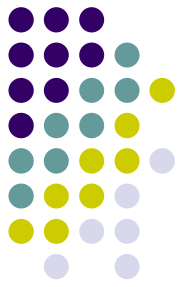
They represent the ideal relative locations of the parts  $v_i$  and  $v_j$

The distance between the locations, weighted by  $M_{ij}^{-1}$  measures the deformation of a "spring" connecting the two parts



# Statistical Framework

- Let's view pictorial structures energy minimization problem in terms of statistical estimation
- Will help gain insight into
  - How to learn pictorial structures training examples
  - How to find multiple good matches



# Statistical Framework

- $\Theta$  – parameters of object model
- $I$  – is the image
- $L$  – is a configuration
- $p(I|L, \Theta)$  is the likelihood of seeing a particular image given that an object is at some location
- $p(L|\Theta)$  is the prior probability that an object is at a given location. Encodes information about the *relative* position of parts, which can be informative and general
- $p(L|I, \Theta)$  is the of the configuration given the model parameters and the image
- Using Bayes' rule:  $p(L|I, \theta) \propto p(I|L, \theta)p(L|\theta)$

# Statistical Framework

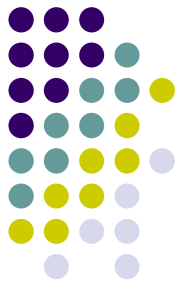


- Things that can be characterized using framework:
  - MAP estimation: equivalent to energy estimate

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

- Sampling from the posterior: provides a natural way to find many good matches. Useful when model is imprecise
- Model estimation: allows us to find  $\Theta$  that specifies a good model for the object

# Statistical Framework



$\theta = (u, E, c)$  are the model parameters

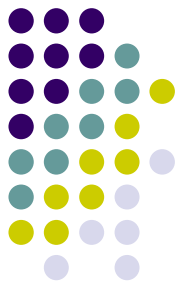
$u = \{u_1, \dots, u_n\}$  are the appearance parameters

$E$  is the set of edges that indicates which parts are connected

$c = \{c_{i,j} \mid (v_i, v_j) \in E\}$  are connection parameters

$$p(I|L, \theta) = p(I|L, u) \propto \prod_{i=1}^n p(I|l_i, u_i).$$

- The likelihood of seeing an image given a particular configuration is the product of individual likelihoods
- Good when no occlusion or overlap
- If occurs, a good estimate can be obtained by sampling from the posterior and using an independent method to select a best match



# Statistical Framework

$$p(L|\theta) = p(L|E, c) = \prod_{(v_i, v_j) \in E} p(l_i, l_j | c_{ij}).$$

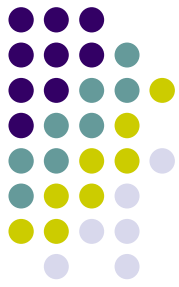
- Prior distribution is defined as a tree-structured Markov random field where no preference is given to the absolute location of each part
- The likelihood of a configuration given an Image and model parameters and is the same as original equation when multiplied by -log

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

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

$$p(l_i, l_j | c_{ij}) \propto \mathcal{N}(T_{ij}(l_i) - T_{ji}(l_j), 0, D_{ij}), \quad (6)$$

where  $T_{ij}$ ,  $T_{ji}$ , and  $D_{ij}$  are the connection parameters encoded by  $c_{ij}$ . These parameters correspond to the ones in equation (2) where  $D_{ij} = M_{ij}/2$  is a diagonal covariance matrix.



# Learning Model Parameters

- Given independent example Images  $\{I^1, \dots, I^m\}$  and corresponding  $\{L^1, \dots, L^m\}$
- Want to estimate model parameters  $\theta$
- Using maximum likelihood estimate:

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

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

- The first part of the equation only depends on the appearance parameters
- The second part depends only on the set of connections and the connection parameters
- Can be solved independently and implies that any kind of part model can be used as long as maximum likelihood can be computed for an individual part

# Estimating Appearance Parameters

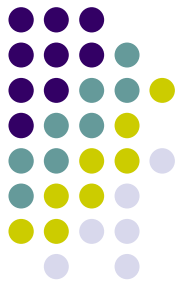


$$u^* = \arg \max_u \prod_{k=1}^m p(I^k | L^k, u).$$

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

Looking at the right hand side we see that to find  $u^*$  we can independently solve for the  $u_i^*$ ,

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



# Estimating the Dependencies

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

- We can first estimate each possible connection before knowing a specific  $E$

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

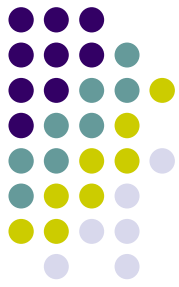
- This maximum likelihood of the joint distribution of  $l_i$  and  $l_j$  depends on the specific representation and it varies with different modeling schemes
- We can however, define a generic quality measure:

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

- These quality measures can be used to estimate  $E$  as follows:

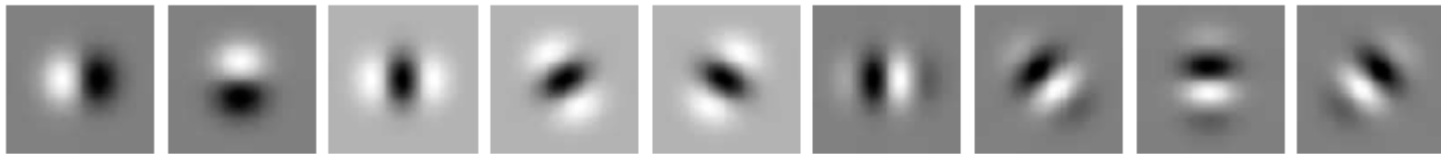
$$E^* = \arg \max_E \prod_{(v_i, v_j) \in E} q(v_i, v_j) = \arg \min_E \sum_{(v_i, v_j) \in E} -\log q(v_i, v_j).$$

- Since  $E$  is a tree, we can solve by finding the minimum spanning tree of the graph. This is a well known problem and can be efficiently solved for



# Iconic Models

- So far, haven't specified how objects are represented.
- Need to choose a specific model
- This type of model has been popular in the context of face detection





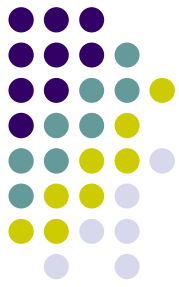
# Model Parts

- Based on the response of Gaussian derivative filters of different orders, orientations and scales (27)
- Vector is normalized and called the iconic index at that position
- Model each as a Gaussian with diagonal covariance matrix
  - Assumes that filters are independent
  - Makes it possible to estimate with a small number of samples

$$p(I|l_i, u_i) \propto \mathcal{N}(\alpha(l_i), \mu_i, \Sigma_i),$$

- $\alpha(l_i)$  is the iconic index at location  $l_i$  in the image
- Can easily estimate the maximum likelihood parameters of this distribution
- Could also use other methods to represent image patches (e.g. eigenspaces)

# Spatial Relations



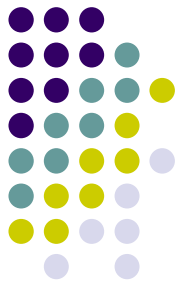
- Connections modeled by a collection of springs between parts
- Characterized by ideal relative location  $s_{ij}$  and full covariance matrix  $\Sigma_{ij}$
- $c_{ij} = (s_{ij}, \Sigma_{ij})$
- Model as Gaussian Distribution:

$$p(l_i, l_j | c_{ij}) = \mathcal{N}(l_i - l_j, s_{ij}, \Sigma_{ij}).$$

- Ideally, location of  $v_i$  is location of  $v_j$  shifted by  $s_{ij}$
- Since models are deformable, locations can vary depending on cost associated with covariance matrix
- To show specific Mahalanobis distance write  $\Sigma = UDU^T$

$$T_{ij}(l_i) = U_{ij}^T(l_i - s_{ij}), \quad \text{and} \quad T_{ji}(l_j) = U_{ij}^T(l_j),$$

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



# Experiments

- Five parts: eyes, nose, corners of mouth
- 20 training images from Yale face database
- More training examples made from scaling and rotating by small amounts
  - Builds invariance into model
- No information about which parts should be connected is given. Model is learned from training data
- No “knobs” to turn
- Few seconds to learn model
- Less than 1 second to find face in image

# Training Results

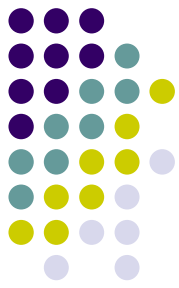
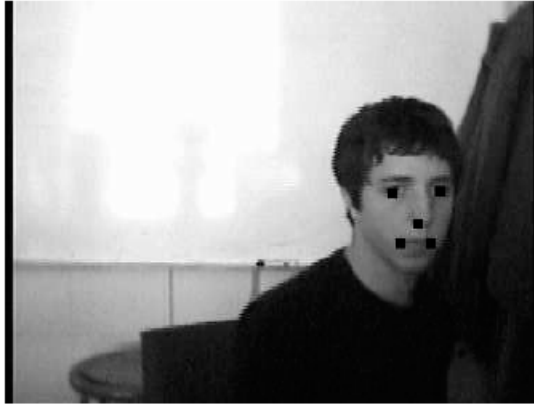


Figure 3: Three examples from the first training set showing the locations of the labeled features and the structure of the learned model.

# Experimental Results



# Occlusions

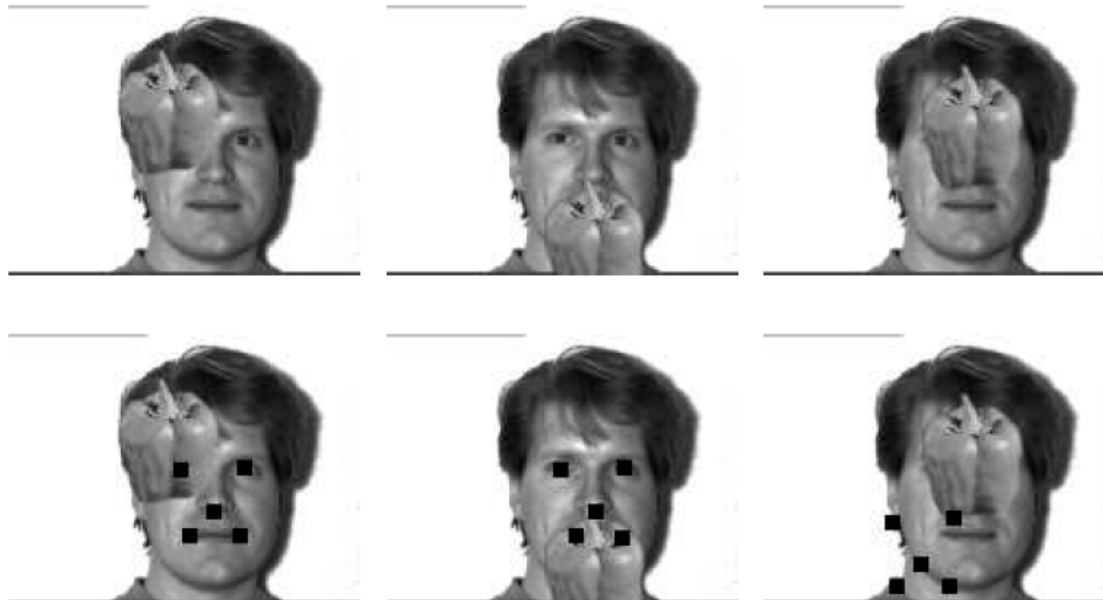
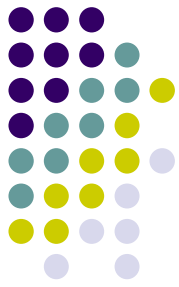


Figure 5: Matching results on occluded faces. The top row shows some input images and the bottom row shows the corresponding matching results. The MAP estimate was a good match when the faces had up to two of five parts occluded and incorrect when three parts were occluded.



# Summary

- Provides statistical framework for representing visual appearance and deformable rigid parts of objects
- Efficient algorithms for finding the best global match
- Statistical framework provides method of learning pictorial structure from example images
  - Most prior work uses manually constructed models.