Spatial Priors for Part-based Recognition using Statistical Models

David Crandall
Cornell Univ.

Pedro Felzenszwalb
Univ. of Chicago

Daniel Huttenlocher
Cornell Univ.

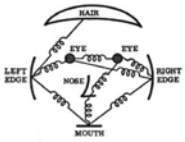
Presented by Elden Yu

Part based recognition

- Combination of appearance-based and geometrical models
  - Each part represents local visual properties
  - Spatial configuration captured by statistical model or spring-like connections
- Pictorial structures, Constellation of parts

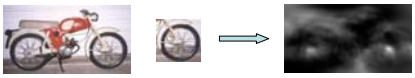
P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. IJCV, 61(1), 2005

R. Fergus, P. Perona and A. Zisserman. Object class recognition by unsupevised scale invariant learning. CVPR, 2003



Main points

- Recognition without feature detection
  - Single overall inference problem
  - Parts have a match quality at each location



- Family of geometrical models
  - Represent using undirected graphical models
  - Choice has strong computational consequences

Statistical models

- An object with n parts $V = \{v_1, ..., v_n\}$
- Location of the object given by a configuration of its parts $L = \{l_1, ..., l_n\}$
- Geometrical model $P(L)$
  - A family of undirected graphs
- Appearance model $P(I \mid L)$
  - Each part is modeled by a template

The detection problem

- To decide if the image has an instance of the object or not

$$q = \frac{p_M(I \mid \omega_1)}{p_M(I \mid \omega_0)}$$

$$p_M(I \mid \omega_1) = \sum_L p_M(L)\, p_M(I \mid L)$$

$$q = \frac{p_M(I \mid \omega_1)}{p_M(I \mid \omega_0)} = \sum_L p_M(L) \prod_{v_i \in V} g_i(I, l_i)$$

The localization problem

- Assume the object is present, find its location

$$L^* = \arg\max_L p_M(L \mid I)$$

$$p_M(L \mid I) \propto P(I \mid L) P(L)$$

Appearance

Spatial prior

The learning problem

- The MLE of the model parameters M=(S,A) given a set of labeled training images

$$S^* = \arg\max_S \prod_i p_M(L_i)$$

$$A^* = \arg\max_A \prod_i p_M(I_i \mid L_i)$$

The appearance model

- Each part v_i is a template Γ_i
  - Apply edge detector on the image
  - Categorize the edge orientations at each pixel $I(p)$
    - 0 means no edge at the pixel
    - A value in {1,2,…r} indicates a predefined direction
  - Compute the orientation histogram $f_i(p)[u]$
- Compute the orientation histogram for the background $b[u]$
- The complete set of parameters for the appearance model is

$$A = ((\Gamma_1, f_1),...(\Gamma_n, f_n), b)$$

The appearance model

- For a background image

$$p_M(I \mid \omega_0) = \prod_p b[I(p)]$$

- For a image without overlapping parts

$$p_M(I \mid L) = p_M(I \mid \omega_0) \prod_{v_i \in V} g_i(I, l_i)$$

$$g_i(I, l_i) = \prod_{p \in \Gamma} \frac{f_i(p)[I(p + l_i)]}{b[I(p + l_i)]}$$

- The key is to assume independence among parts

Common spatial priors

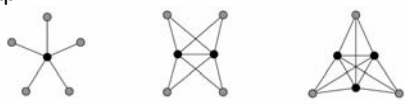
- The simplest form is to assume no spatial dependence between parts
  - Easy to compute
  - Hard to capture relative spatial information

$$p_M(L) = \prod_{v_i \in V} p_M(l_i)$$

- The other extreme is to assume no independence among parts
  - Hard to make inference
  - Normally feature detection is applied to reduce the search space

k-fans

- Family between a star graph and the complete graph



1-fan 2-fan 3-fan

- Each k-fan has k nodes that are completed connected with each other; while other nodes connect to each of the k node only.

Why is *k*-fans useful?

- Let $R = \{v_1,...,v_k\}$ be the reference parts in a *k*-fan, and $\bar{R} = V - R$ be the remaining parts; denote $l_R = \{l_1,...,l_k\}$ as a particular configuration of the reference parts
- The spatial priors defined by a k-fan can be written as

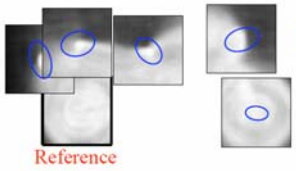
$$p_M(L) = p_M(l_R) \prod_{v_i \in \bar{R}} p_M(l_i \mid l_R)$$

Benefits of k-fans

- The localization and detection problems for models with k-fans as spatial priors can be solved in $o(nh^{k+1})$ time, where n is number of parts, and h is the number of locations
- K controls the complexity of inference with the model.
  - K=0 means no dependency between parts
  - K=n-1 means no independence between parts

Motorbike model

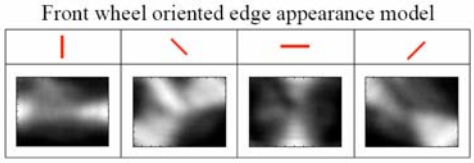
Part appearance defined by probability of an edge



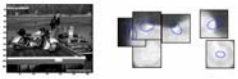
Ellipses show spatial uncertainty of non-reference parts

Reference

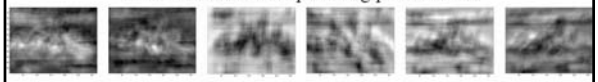
Front wheel oriented edge appearance model



Processing steps



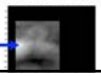
Match costs for each part - log probabilities



DT of non-reference match costs - correct form of blurring

Sum shifted DTs and reference match cost

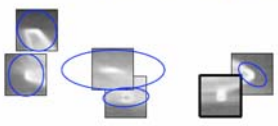
Find best location for reference



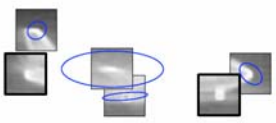
Localization with 1-fans

Airplane model

As k increases, the geometrical model becomes more precise

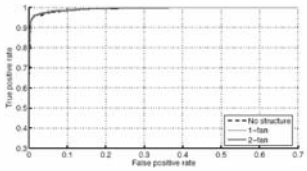


1-fan

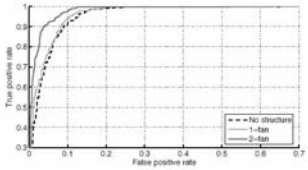


2-fan

ROC curves



For motorbikes 1-fan is as good as 2-fan



For airplanes 2-fan is better than 1-fan

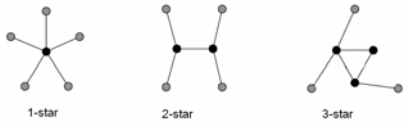
Comparable to other methods on the Caltech dataset

Discussion

- Detection times:
  - 0.1 sec for 1-fan vs. 3.3 sec for 2-fan
- Small amount of geometry can buy a lot
  - Appropriate amount depends on object class
  - Trade-off between model structure and computational complexity
- Recognition without feature detection combines bottom-up and top-down constraints
  - Each part is detected in the context of the others

Star-skeleton representation

- Each star point is a point that is visible to all points in the same polygon



- A simple polygon can be decomposed into a set of star polygons