Object Recognition by Parts

- Object recognition started with line segments.

  - Roberts recognized objects from line segments and junctions.

  - This led to systems that extracted linear features.

  - CAD-model-based vision works well for industrial.

- An “appearance-based approach” was first developed for face recognition and later generalized up to a point.

- The new interest operators have led to a new kind of recognition by “parts” that can handle a variety of objects that were previously difficult or impossible.
Object Class Recognition
by Unsupervised Scale-Invariant Learning

R. Fergus, P. Perona, and A. Zisserman
Oxford University and Caltech

CVPR 2003
won the best student paper award
Goal:

• Enable Computers to Recognize Different Categories of Objects in Images.
Approach

• An object is a random constellation of parts (from Burl, Weber and Perona, 1998).

• The parts are detected by an interest operator (Kadir’s).

• The parts can be recognized by appearance.

• Objects may vary greatly in scale.

• The constellation of parts for a given object is learned from training images
Components

• Model
  – Generative Probabilistic Model including Location, Scale, and Appearance of Parts

• Learning
  – Estimate Parameters Via EM Algorithm

• Recognition
  – Evaluate Image Using Model and Threshold
Model: Constellation Of Parts

Fischler & Elschlager, 1973

Yuille, 91
Brunelli & Poggio, 93
Lades, v.d. Malsburg et al. 93
Cootes, Lanitis, Taylor et al. 95
Amit & Geman, 95, 99
Perona et al. 95, 96, 98, 00
Parts Selected by Interest Operator
Kadir and Brady's Interest Operator.
Finds Maxima in Entropy Over Scale and Location
Representation of Appearance

121 dimensions was too big, so they used PCA to reduce to 10-15.
Learning a Model

• An object class is represented by a generative model with $P$ parts and a set of parameters $\theta$.

• Once the model has been learned, a decision procedure must determine if a new image contains an instance of the object class or not.

• Suppose the new image has $N$ interesting features with locations $X$, scales $S$ and appearances $A$. 
Generative Probabilistic Model

Top-Down Formulation

Bayesian Decision Rule

\[ R = \frac{p(\text{Object}|X, S, A)}{p(\text{No object}|X, S, A)} \]

\[ = \frac{p(X, S, A|\text{Object}) p(\text{Object})}{p(X, S, A|\text{No object}) p(\text{No object})} \]

\[ \approx \frac{p(X, S, A|\theta) p(\text{Object})}{p(X, S, A|\theta_{bg}) p(\text{No object})} \]

\[ p(X, S, A|\theta) = \sum_{h \in H} p(X, S, A, h|\theta) = \]

\[ \sum_{h \in H} \left( \underbrace{p(A|X, S, h, \theta)}_{\text{Appearance}} \underbrace{p(X|S, h, \theta)}_{\text{Shape}} \underbrace{p(S|h, \theta)}_{\text{Rel. Scale}} \underbrace{p(h|\theta)}_{\text{Other}} \right) \]

R is the likelihood ratio.

\( \theta \) is the maximum likelihood value of the parameters of the object and \( \theta_{bg} \) of the background.

h is the hypothesis as to which P of the N features in the image are the object, implemented as a vector of length P with values from 0 to N indicating which image feature corresponds to each object feature.

H is the set of all hypotheses; Its size is \( O(N^P) \).
The appearance (A) of each part p has a Gaussian density with mean $c_p$ and covariance $V_p$.

Background model has mean $c_{bg}$ and covariance $V_{bg}$.

The vector $d$ of length $P$ has a 1 for visible parts and 0 for occluded parts. $d_p$ stands for $d[p]$.

$$\frac{p(A|X, S, h, \theta)}{p(A|X, S, h, \theta_{bg})} = \prod_{p=1}^{P} \left( \frac{G(A(h_p)|c_p, V_p)}{G(A(h_p)|c_{bg}, V_{bg})} \right)^{d_p}$$
Shape as Location

Object shape is represented by a joint Gaussian density of the locations ($X$) of features within a hypothesis transformed into a scale-invariant space.

\[
\frac{p(X|S, h, \theta)}{p(X|S, h, \theta_{bg})} = G(X(h)|\mu, \Sigma) \alpha^f
\]

$\alpha$ is the area of the image

$f$ is number of foreground features in the hypothesis.
The relative scale of each part is modeled by a Gaussian density with mean $t_p$ and covariance $U_p$.

$$\frac{p(S|h, \theta)}{p(S|h, \theta_{bg})} = \prod_{p=1}^{P} G(S(h_p)|t_p, U_p)^{d_p} r^f$$

$r$ is a range. $f$ is number of visible features.
Occlusion and Part Statistics

\[
\frac{p(\mathbf{h}|\theta)}{p(\mathbf{h}|\theta_{bg})} = \frac{p_{\text{Poiss}}(n|M)}{p_{\text{Poiss}}(N|M)} \frac{1}{nC_r(N,f)} p(d|\theta)
\]

- First term: Poisson distribution (mean M) models the number of features in the background.
- Second term: (constant) 1/(number of combinations of \( f_t \) features out of a total of \( N_t \))
- Third term: gives probability for possible occlusion patterns.
Learning

• Train Model Parameters Using EM:
  • Optimize Parameters
  • Optimize Assignments
  • Repeat Until Convergence

\[ \theta = \{ \mu, \Sigma, c, V, M, p(d|\theta), t, U \} \]

\[ \hat{\theta}_{ML} = \arg \max_{\theta} p(X, S, A | \theta) \]
Recognition

Make This:

\[
R = \frac{p(\text{Object}|X, S, A)}{p(\text{No object}|X, S, A)} \\
= \frac{p(X, S, A|\text{Object}) \cdot p(\text{Object})}{p(X, S, A|\text{No object}) \cdot p(\text{No object})} \\
\approx \frac{p(X, S, A|\theta) \cdot p(\text{Object})}{p(X, S, A|\theta_{bg}) \cdot p(\text{No object})}
\]

Greater Than Threshold
RESULTS

- Initially tested on the Caltech-4 data set
  - motorbikes
  - faces
  - airplanes
  - cars

- Now there is a much bigger data set: the Caltech-101
  http://www.vision.caltech.edu/archive.html
Equal error rate: 7.5%

Motorbikes

Motorbike shape model
Background Images
Equal error rate: 4.6%
Equal error rate: 9.8%

Airplanes
Scale-Invariant Cats

Equal error rate: 10.0%

Part 1 – Det: 8e–22

Part 2 – Det: 2e–22

Part 3 – Det: 5e–22

Part 4 – Det: 2e–22

Part 5 – Det: 1e–22

Part 6 – Det: 4e–21

Background – Det: 2e–18

Spotted cat shape model
Scale-Invariant cars

Equal error rate: 9.7%
Robustness of Algorithm

![Graphs showing the robustness of the algorithm for Face dataset and Motorbike dataset. The x-axis represents the percentage of training images containing the object, and the y-axis represents the percentage of correct predictions. The graph for Face dataset shows a steep increase in correct predictions as the percentage of training images increases, reaching nearly 100% by 100%. The graph for Motorbike dataset also shows an increase, but with a slower rate, reaching around 85% by 6 parts.](image)
# Accuracy

Initial Pre-Scaled Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ours</th>
<th>Others</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorbikes</td>
<td>92.5</td>
<td>84</td>
<td>[17]</td>
</tr>
<tr>
<td>Faces</td>
<td>96.4</td>
<td>94</td>
<td>[19]</td>
</tr>
<tr>
<td>Airplanes</td>
<td>90.2</td>
<td>68</td>
<td>[17]</td>
</tr>
<tr>
<td>Cars(Side)</td>
<td>88.5</td>
<td>79</td>
<td>[1]</td>
</tr>
</tbody>
</table>
ROC equal error rates

Scale-Invariant Learning and Recognition:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total size of dataset</th>
<th>Object size range (pixels)</th>
<th>Pre-scaled performance</th>
<th>Unscaled performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorbikes</td>
<td>800</td>
<td>200-480</td>
<td>95.0</td>
<td>93.3</td>
</tr>
<tr>
<td>Airplanes</td>
<td>800</td>
<td>200-500</td>
<td>94.0</td>
<td>93.0</td>
</tr>
<tr>
<td>Cars (Rear)</td>
<td>800</td>
<td>100-550</td>
<td>84.8</td>
<td>90.3</td>
</tr>
</tbody>
</table>