Object Recognition

16-385 Computer Vision (Kris Kitani)

Carnegie Mellon University

Henderson and Davis. Shape recognition using hierarchical Constraint Analysis. 1979
What do we mean by ‘object recognition’?
Is this a street light?
(Verification / classification)
Where are the people?
(Detection)
Is that Potala palace? (Identification)
What's in the scene?
(semantic segmentation)

- Sky
- Mountain
- Trees
- Building
- Vendors
- People
- Ground
What type of scene is it?
(Scene categorization)

Outdoor
Marketplace
City
Challenges
(Object Recognition)
Viewpoint variation
Illumination variation
Scale variation
Deformation
Intra-class variation
Common approaches
Common approaches: object recognition

Feature Matching

Spatial reasoning

Window classification
Feature matching
What object do these parts belong to?
Some local features are very informative. An object can be described as a collection of local features (bag-of-features). Properties include:

- Deals well with occlusion
- Scale invariant
- Rotation invariant

Are the positions of the parts important?
Pros

- Simple
- Efficient algorithms
- Robust to deformations

Cons

- No spatial reasoning
Common approaches: object recognition

- Feature Matching
- Spatial reasoning
- Window classification
Spatial reasoning
The position of every part depends on the positions of all the other parts

Many parts, many dependencies!
1. Extract features

2. Match features

3. Spatial verification
1. Extract features
2. Match features
3. Spatial verification
1. Extract features
2. Match features
3. Spatial verification

an old idea...
Scene

Structural (grammatical) description
Coded Chromosome

\[ v_T = \{ \left( \begin{array}{c} \circlearrowleft \, a, \\ \nearrow \, b, \\ \nearrow \, c, \\ \searrow \, d \end{array} \right) \} \]

\[ x = \text{cddabbbdabbbabbcbbabbbbd} \]

Substructures of Coded Chromosome

\[ S_1 = \{ [b[[[a]b]b]b]b; [b[b[b[a]]b]b]b; [b[b[[[a]b]b]b]b]b; [b[b[b[a]]b]] \} \]
The Representation and Matching of Pictorial Structures
MARTIN A. FISCHLER and ROBERT A. ELSSLAGER

Abstract—The primary problem dealt with in this paper is the following. Given some description of a visual object, find that object in an actual photograph. Part of the solution to this problem is the specification of a descriptive scheme, and a metric on which to base the decision of “goodness” of matching or detection.

We offer a combined descriptive scheme and decision metric which is general, intuitively satisfying, and which has led to promising experimental results. We also present an algorithm which takes the above descriptions, together with a matrix representing the intensities of the actual photograph, and then finds the described object in the matrix. The algorithm uses a procedure similar to dynamic programming in order to cut down on the vast amount of computation otherwise necessary.

One desirable feature of the approach is its generality. A new programming system does not need to be written for every new description; instead, one just specifies descriptions in terms of a certain set of primitives and parameters.

Description for left edge of face

VALUE(X)=(E+F+G+H)-(A+B+C+D)

Note: VALUE(X) is the value assigned to the L(EV)A corresponding to the location X as a function of the intensities of locations A through H in the sensed scene.
A more modern probabilistic approach…

think of locations as random variables (RV)

\[ \mathbf{L} = \{ L_1, L_2, \ldots, L_M \} \]
A more modern probabilistic approach…

think of locations as random variables (RV)

vector of RVs: set of part locations

\[ \mathbf{L} = \{ L_1, L_2, \ldots, L_M \} \]

What are the dimensions of R.V. \( \mathbf{L} \)?

How many possible combinations of part locations?
A more modern probabilistic approach…

think of locations as random variables (RV)

vector of RVs: set of part locations

\[ \mathbf{L} = \{ L_1, L_2, \ldots, L_M \} \]

What are the dimensions of R.V. \( \mathbf{L} \)?

\[ L_m = [x \ y] \]

How many possible combinations of part locations?
A more modern probabilistic approach…

think of locations as random variables (RV)

vector of RVs: set of part locations

\[ L = \{ L_1, L_2, \ldots, L_M \} \]

What are the dimensions of R.V. \( L \)?

\[ L_m = [x \ y] \]

How many possible combinations of part locations?

\[ N^M \]
Most likely set of locations $L$ is found by maximizing:

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

**Posterior**

**Likelihood:** How likely it is to observe image $I$ given that the $M$ parts are at locations $L$ (scaled output of a classifier)

**Prior:** Spatial prior controls the geometric configuration of the parts

What kind of prior can we formulate?
Given any collection of selfie images, where would you expect the nose to be?

What would be an appropriate prior?

\[ P(L_{\text{nose}}) = ? \]
A simple factorized model

\[ p(L) = \prod_{m} p(L_m) \]

Break up the joint probability into smaller (independent) terms
Independent locations

Each feature is allowed to move independently

\[ p(L) = \prod_{m} p(L_m) \]

Does not model the relative location of parts at all
Tree structure
(star model)

Represemt the location of all the parts relative to a single reference part

Assumes that one reference part is defined (who will decide this?)

\[ p(L) = p(L_{\text{root}}) \prod_{m=1}^{M-1} p(L_m | L_{\text{root}}) \]
Fully connected
(constellation model)

\[ p(L) = p(l_1, \ldots, l_N) \]

Explicitly represents the joint distribution of locations

**Good model:**
Models relative location of parts
BUT Intractable for moderate number of parts
Pros

• Retains spatial constraints
• Robust to deformations

Cons

• Computationally expensive
• Generalization to large inter-class variation (e.g., modeling chairs)
Feature Matching

Spatial reasoning

Window classification
Window-based
Template Matching

1. get image window
2. extract features
3. classify

When does this work and when does it fail?

How many templates do you need?
Per-exemplar

find the ‘nearest’ exemplar, inherit its label
Template Matching

1. get image window (or region proposals)
2. extract features
3. compare to template

Do this part with one big classifier ‘end to end learning’
Convolutional Neural Networks

Convolution

Image patch (raw pixels values)

response of one ‘filter’

A 96 x 96 image convolved with 400 filters (features) of size 8 x 8 generates about 3 million values ($89^2 \times 400$)

Pooling

Image patch (raw pixels values)

max/min response over a region

response of one ‘filter’

Pooling aggregates statistics and lowers the dimension of convolution
630 million connections
60 millions parameters to learn

Krizhevsky, A., Sutskever, I. and Hinton, G. E.
ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012.
Pros

- Retains spatial constraints
- Efficient test time performance

Cons

- Many many possible windows to evaluate
- Requires large amounts of data
- Sometimes (very) slow to train
How to write an effective CV resume
Deep Learning
+1-DEEP-LEARNING  deeplearning@deeplearning  http://deeplearning


**Education**

<table>
<thead>
<tr>
<th>Deep Learning</th>
<th>Deep Learning</th>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Learning</td>
<td>Deep Learning</td>
<td>Deep Learning</td>
</tr>
</tbody>
</table>

**Experience**

<table>
<thead>
<tr>
<th>Deep Learning</th>
<th>Deep Learning</th>
</tr>
</thead>
</table>

**Deep Learning in another country**

<table>
<thead>
<tr>
<th>Deep Learning</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>· Deep Learning</td>
</tr>
</tbody>
</table>

**Publications**


2. **Shallow Learning... Nawww... Deep Learning bruh**  Under submission while Deep Learning

**Patent**

1. **System and Method for Deep Learning**  Deep Learning, Deep Learning , Deep Learning , Deep Learning