Taking Computer Vision Into The Wild

Neeraj Kumar

October 4, 2011
CSE 590V – Fall 2011
University of Washington
A Joke

Q. What is computer vision?

A. If it doesn’t work (in the wild), it’s computer vision.

(I’m only half-joking)
Instant Object Recognition Paper*
1. Design new algorithm
   - Fixed set of training examples
   - Fixed set of classes/objects

2. Pick dataset(s) to evaluate on

3. Repeat until conference deadline:
   a. Train classifiers
      - Training examples only have one object, often in center of image
   b. Evaluate on test set
      - Fixed test set, usually from same overall dataset as training
   c. Tune parameters and tweak algorithm
      - MTurk filtering, pruning responses, long training times, ...

4. Brag about results with ROC curves
   - How does it do on real data? New classes?

*Just add grad students
Object Recognition Paper

1. User proposes new object class
2. System gathers images from flickr
3. Repeat until convergence:
   a. Choose windows to label
   b. Get labels from MTurk
   c. Improve classifier (detector)
4. Also evaluate on Pascal VOC
   - How does it compare to state of the art?

[S. Vijayanarasimhan & K. Grauman – Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds (CVPR 2011)]
Object Representation

Deformable Parts: Root + Parts + Context

P=6 parts, from bootstrap set

C=3 context windows, excluding object candidate, defined to the left, right, above
### Features: Sparse Max Pooling

<table>
<thead>
<tr>
<th></th>
<th>Bag of Words</th>
<th>Sparse Max Pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base features</td>
<td>SIFT</td>
<td>SIFT</td>
</tr>
<tr>
<td>Build vocabulary tree</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Quantize features</td>
<td>Nearest neighbor, hard decision</td>
<td>Weighted nearest neighbors, sparse coded</td>
</tr>
<tr>
<td>Aggregate features</td>
<td>Spatial pyramid</td>
<td>Max pooling</td>
</tr>
</tbody>
</table>

---

[Y.-L. Boureau, F. Bach, Y. LeCun, J. Ponce – Learning Mid-level Features for Recognition (CVPR 2010)]

[J. Yang, K. Yu, Y. Gong, T. Huang – Linear Spatial Pyramid Matching Sparse Coding for Image Classification (CVPR 2009)]
How to Generate Root Windows?

100,000s of possible locations, aspect ratios, sizes

X

1000s of images

= too many possibilities!
Jumping Windows

• Build lookup table of how frequently given feature in a grid cell predicts bounding box
• Use lookup table to vote for candidate windows in query image a la generalized Hough transform
Pick Examples via Hyperplane Hashing

• Want to label “hard” examples near the hyperplane boundary
• But hyperplane keeps changing, so have to recompute distances...

• Instead, hash all unlabeled examples into table
• At run-time, hash current hyperplane to get index into table, to pick examples close to it

[P. Jain, S. Vijayanarasimhan & K. Grauman – Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning (NIPS 2010)]
Comparison on Pascal VOC

- Comparable to state-of-the-art, better on few classes
- Many fewer annotations required!
- Training time is 15mins vs 7 hours (LSVM) vs 1 week (SP+MKL)
Online Live Learning for Pascal

- Comparable to state-of-the-art, better on fewer classes
- But using flickr data vs. Pascal data, and automatically
Sample Results

Correct

Incorrect
Lessons Learned

• It is possible to leave the sandbox
  • And still do well on sandbox evaluations
• Sparse max pooling with a part model works well
• Linear SVMs can be competitive with these features
• Jumping windows is MUCH faster than sliding
• Picking examples to get labeled is a big win
• Linear SVMs also allow for fast hyperplane hashing
Limitations

“Hell is other people”

With apologies to Jean-Paul Sartre
Object Recognition Paper

1. User proposes new class
2. System gathers images from flickr
3. Repeat until convergence:
   a. Choose windows to label
   b. Get labels from MTurk
   c. Improve classifier (detector)
4. Also evaluate on Pascal VOC
Solving Real Problems for Users

Users want to do stuff

It doesn’t work well enough

Users express their displeasure

*With apologies to John Gabriel

Object Recognition Paper
1. User proposes new class
2. System gathers images from flickr
3. Repeat until convergence:
   a. Choose windows to label
   b. Get labels from MTurk
   c. Improve classifier (detector)
4. Also evaluate on Pascal VOC
...And Never The Twain Shall Meet?

Pascal VOC Results from Previous Paper
Unsolved Vision Problems

Simplify Problem!
leafsnap

Columbia University
University of Maryland
Smithsonian Institution

Available on the App Store
Available on the iPad App Store
Place leaf on a white background.

Center leaf in the frame.
Easier Segmentation for Leafsnap
Plants vs Birds

2d

- Doesn’t move
- Okay to pluck from tree
- Mostly single color
- Very few parts
- Adequately described by boundary
- Relatively easy to segment

3d

- Moves
- Not okay to pluck from tree
- Many colors
- Many parts
- Not well described by boundary
- Hard to segment
Human-Computer Cooperation

Red!
Top-right!
Uh, it’s pointy?
Bottom-left!

Where is it?
Okay.

What color is it?
Where’s the beak?
Describe its beak
Where’s the tail?

20 Questions

Is the beak cone-shaped? yes
Is the upper-tail brown? yes
Is the breast solid colored? no
Is the breast striped? yes
Is the throat white? yes
The bird is a Henslow’s Sparrow

http://20q.net/
Information Gain for 20Q

Pick most informative question to ask next

\[
I(c; u_i | x, U^{t-1}) = \mathbb{E}_u \left[ \text{KL} \left( p(c | x, u_i \cup U^{t-1}) \| p(c | x, U^{t-1}) \right) \right]
\]

\[
= \sum_{u_i \in A \times V} p(u_i | x, U^{t-1}) \left( H(c | x, u_i \cup U^{t-1}) - H(c | x, U^{t-1}) \right)
\]

Expected information gain of class c, given image & previous responses
Probability of getting response \( u_i \), given image & previous responses
Entropy of class c, given image and possible new response \( u_i \)
Entropy of class c right now
Answers make distribution peakier
Incorporating Computer Vision

Probability of class $c$, given image and any set of responses

Bayes’ rule

Assume variations in user responses are NOT image-dependent

Probabilities affect entropies!
Incorporating Computer Vision...

...leads to different questions

**Western Grebe**

- **w/ vision:**
  - Q #1: Is the throat white? yes (Def.)
- **w/o vision:**
  - Q #1: Is the shape perching-like? no (Def.)

**Rose-breasted Grosbeak**

- Only CV
- CV + Q #1: Is the crown black? yes (Def.)

**Yellow-headed Blackbird**

- Rose-breasted Grosbeak
Ask for User Confidences
Modeling User Responses is Effective!
Birds-200 Dataset

http://www.vision.caltech.edu/visipedia/CUB-200.html
Results
Results

With fewer questions, CV does better
With more questions, humans do better
Lessons Learned

• Computer vision is not (yet) good enough for users
  • But users can meet vision halfway
• Minimizing user effort is key!
• Users are not to be trusted (fully)
• Adding vision improves recognition
• For fine-scale categorization, attributes do better than 1-vs-all classifiers if there are enough of them

<table>
<thead>
<tr>
<th>Classifier</th>
<th>200 (1-vs-all)</th>
<th>288 attr.</th>
<th>100 attr.</th>
<th>50 attr.</th>
<th>20 attr.</th>
<th>10 attr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg # Questions</td>
<td>6.43</td>
<td>6.72</td>
<td>7.01</td>
<td>7.67</td>
<td>8.81</td>
<td>9.52</td>
</tr>
</tbody>
</table>
Limitations

• Real system still requires much human effort
• Only birds
• Collecting and labeling data
  • Crowdsourcing?
  • Experts?
• Building usable system
  • Minimizing
Visipedia

http://www.vision.caltech.edu/visipedia/