Big Issues

• What is out there on the Internet?
• How do we get it?
• What can we do with it?
  – Can we solve old problems more easily?
  – Can we propose new problems?
Subject-specific Data

Photos of Coliseum (Snavely et al.)

Portraits of Bill Clinton
Much of Captured World is “Generic”
Generic Data

street scenes

Food plates

faces

pedestrians
The Internet as a Data Source

• Computer Vision data sets
  – Smaller in size,
  – Higher quality (annotated) photos

• Search engines
  – Sometimes inaccurate,
  – Not much meta data

• Photo sharing sites
  – Larger in size,
  – Higher quality photos,
  – Metadata (text and geotags)

• Social networks
  – Largely inaccessible
The Internet makes images ubiquitous

100 billion – Estimated number of photos on Facebook by mid-2011.

6 billion – Photos hosted on Flickr (August 2011).

4.5 million – Number of photos uploaded to Flickr each day.

60 – The average number of photos uploaded per second to Instagram.
How many images is enough?

- Number of possible images is huge
- Natural images make up only a small percentage of all possible
- If we have enough natural images do problems become easy?
  - If so, how many do we need?
Motivation

Space of all images

Subspace of monkeys

Parametric model of monkeys
Two kinds of Things in the World

Navier-Stokes Equation

$$\frac{\partial u}{\partial t} = -(u \cdot \nabla) u + v \nabla^2 u - \frac{1}{d} \nabla p + f$$

+ weather
+ location

+
Lots of data available
Non-parametric Approach

!!! HIGH DIMENSIONAL !!!
Subspace of natural images

!!! HIGH DIMENSIONAL !!!
Subspace of monkeys

Space of all images
“Unreasonable Effectiveness of Data”

Parts of our world can be explained by elegant mathematics
  • physics, chemistry, astronomy, etc.

But much cannot
  • psychology, economics, genetics, etc.

Enter The Data!
  • Great advances in several fields:
    – e.g. speech recognition, machine translation
    – Case study: Google

A.I. for the postmodern world:

- All questions have already been answered…many times, in many ways
- Google is dumb, the “intelligence” is in the data
How about *visual* data?

Text is simple:
- clean, segmented, compact, 1D

Visual data is much harder:
- Noisy, unsegmented, high entropy, 2D/3D

Quick Overview
- Comparing Images
- Uses of Visual Data
- The Dangers of Data
Computing visual distances is hard

\[ \text{CLIME} - \text{CRIME} = \text{Hamming distance of 1 letter} \]

\[ \begin{array}{cc}
\text{y} & \text{y} \\
\hline
& \\
\end{array} \\
\begin{array}{c}
\rightarrow \\
\hline
\rightarrow \\
\end{array} = \text{Euclidian distance of 5 units} \]

\[ \begin{array}{c}
\text{gray} \\
\end{array} = \text{Gray value distance of 50 values} \]

\[ \begin{array}{c}
\text{image} \\
\end{array} = ? \]
SSD says these are not similar
Thumbnail Collection Project

- Collect images for ALL objects
  - List obtained from WordNet
  - 75,378 non-abstract nouns in English

- Example first 20:

  a-bomb
  a-horizon
  a._conan_doyle
  a._e._burnside
  a._e._housman
  a._e._kennelly
  a.e.
  a_battery
  a_cappella_singing
  a_horizon

  a_kempis
  aalborg
  aalii
  aalost
  aalto
  aar
  aardvark
  aardwolf
  aare
  aare_river
Thumbnail Collection

- 7 different search engines

![Image Search Logos]
Dataset Statistics

• **Overall stats**
  - 79,302,017 images
  - 75,062 different words

• **Details**
  - Two formats: square & rectangular
  - Gathered at 4.5 images/second
  - Downloaded 97,245,098 images
  - 18% duplicate rate
  - Disk usage: ~ 700Gb
  - Collection time: ~ 9 months
Histogram Images/Word

Total, unique, non-uniform images: 79,302,017

Total number of words: 75,062

Mean # images per word: 1,056
Labeling Noise

- Manual labeling of 78 classes

- Best: Google & Altavista

- Worst: Cydral & Webshots
Suitable Image Representation

• Want minimal representation for task:
  – Classifying scene and dominant objects

• Compact representation has low storage requirements

• We blur & subsample to give low-res image (32x32 color)
Why such tiny images?

Small
• Easy to store in a reasonable amount of space
• Can process lots of them in a short amount of time

Humans can do recognition well at small scale
Human Labeling of Tiny Scenes

32x32
Human Labeling of Tiny Scenes

32x32

office

waiting area

dining room

dining room

wall-space

windows

desk

drawers

window

reception desk

Couch

chairs

table

light

plant

door

ceiling

picture

drawers

wall

center piece

table

chair

chair

floor

256x256

Context!
Image Segmentation (by humans)
Human Scene Recognition

The role of context in object recognition
A. Oliva, A. Torralba
Non-parametric Classifier

- Nearest-neighbors

- For each query, obtain sibling set (neighbors)

- 3 different types of distance metric

- Hand-designed, use whole image
Metric 1 - $D_{SSD}$

- Sum of squared differences (SSD)

\[ D^{2}_{SSD} = \sum_{x,y,c} \left( \text{Image 1} - \text{Image 2} \right)^{2} \]

To give invariance to illumination: Each image normalized to be zero mean, unit variance
Metric 2 - $D_{\text{Warp}}$

- SSD but allow small transformations

$$D^2_{\text{warp}} = \min_{\theta} \sum_{x,y,c}$$

Find min using gradient descent
Metric 3 - $D_{\text{Shift}}$

- As per Warping but also allow sub-window shifts

$$D_{\text{shift}}^2 = \min_{\text{Local sub-window}} \sum_{x,y,c}$$

- Quick since images are so small
Sibling Sets with Different Metrics

- Sibling set is 50 images
Approximate - $D_{SSD}$

- Exact distance metrics are too expensive to apply to all 79 million images.

- Use approximate scheme based on taking first $K=19$ principal components.

Apply $D_{SSD}$, $D_{warp}$ & $D_{shift}$ to these $M$ images @ 32x32.
How Does $D_{SSD}$ Relate to Semantic Distance?
Label Assignment

- Distance metrics give set of nearby images
- How to compute label?

Query | Grover | Cleveland | Linnet | Birdcage | Chiefs | Casing
-------|--------|-----------|--------|----------|--------|-------
![Query Image](image1) ![Grover Image](image2) ![Cleveland Image](image3) ![Linnet Image](image4) ![Birdcage Image](image5) ![Chiefs Image](image6) ![Casing Image](image7)

- Issues:
  - Labeling noise
  - Keywords can be very specific
    - e.g. yellowfin tuna
Wordnet – A Lexical Dictionary

http://wordnet.princeton.edu/

Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun aardvark

Sense 1
aardvark, ant bear, anteater, Orycteropus afer
  => placental, placental mammal, eutherian, eutherian mammal
  => mammal
    => vertebrate, craniate
      => chordate
        => animal, animate being, beast, brute, creature
          => organism, being
            => living thing, animate thing
              => object, physical object
                => entity

Each noun corresponds to a path in Wordnet tree
Vote for branches in Wordnet tree

Wordnet Voting Scheme

a) Input image

b) Neighbors

d) Wordnet voted branches

One image - one vote

Vote for branches in Wordnet tree
Wordnet Voting Scheme

- Input image
- Neighbors
- Ground truth
- Wordnet voted branches
Wordnet Voting

• Overcomes differences in level of semantic labeling:
  – e.g. “person” & “sir arthur conan doyle”

• Totally incorrect labels form hopefully uniform background noise

• Assumes semantic and visual consistency are closely related
Recognition Experiments
Person Recognition

- 23% of all images in dataset contain people

- Wide range of poses: not just frontal faces
Person Recognition – Test Set

- 1016 images from Altavista using “person” query
- High res and 32x32 available
- Disjoint from 79 million tiny images
Person Recognition

- Task: person in image or not?
Re-ranked Altavista Images

Original

Re-ranked
Scene Classification

- Test set: 1125 images randomly drawn from 79 million.
- Task: \{scene\} vs all other classes

Count the votes at the corresponding node of Wordnet tree for classification.
Object Classification

Extrapolation of how well it would do at Google scale dataset
Performance drops as classes become more specific
Automatic Colorization

Grayscale input
High resolution
Automatic Colorization

Grayscale input
High resolution

Grayscale
32x32 siblings
Automatic Colorization

Grayscale input
High resolution

Grayscale
32x32 siblings

Color siblings
high resolution
# Automatic Colorization

<table>
<thead>
<tr>
<th>Grayscale input High resolution</th>
<th><img src="image1" alt="Grayscale input High resolution" /></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grayscale 32x32 siblings</td>
<td><img src="image2" alt="Grayscale 32x32 siblings" /></td>
</tr>
<tr>
<td>Color siblings high resolution</td>
<td><img src="image3" alt="Color siblings high resolution" /></td>
</tr>
<tr>
<td>Average of color siblings</td>
<td><img src="image4" alt="Average of color siblings" /></td>
</tr>
</tbody>
</table>
Automatic Colorization

- Grayscale input
  - High resolution
- Grayscale
  - 32x32 siblings
- Color siblings
  - High resolution
- Average of color siblings
- Colorization of input using average
Automatic Colorization

Grayscale input
High resolution

Grayscale
32x32 siblings

Color siblings
high resolution

Average of
color siblings

Colorization of input
using average

Colorization of input
using specific siblings
Automatic Colorization Result

Grayscale input High resolution

Colorization of input using average
Automatic Orientation

- Look at mean distance to neighbors

Images at wrong orientation have neighbors further away
Automatic Orientation Examples

Average correlation to 50 closest neighbors
Automatic Orientation

- Many images have ambiguous orientation
- Look at top 25% by confidence:
- Examples of high and low confidence images:
Tiny Images Discussion

Why SSD?

Can we build a better image descriptor?
Image Representations: Histograms

global histogram

- Represent distribution of features
  - Color, texture, depth, ...
Image Representations: Histograms

Joint histogram
- Requires lots of data
- Loss of resolution to avoid empty bins

Marginal histogram
- Requires independent features
- More data/bin than joint histogram

Images from Dave Kauchak
Image Representations: Histograms

Adaptive binning

- Better data/bin distribution, fewer empty bins
- Can adapt available resolution to relative feature importance
Image Representations: Histograms

Clusters / Signatures
- “super-adaptive” binning
- Does not require discretization along any fixed axis

Images from Dave Kauchak
Issue: How to Compare Histograms?

Bin-by-bin comparison
Sensitive to bin size.
Could use wider bins …
… but at a loss of resolution

Cross-bin comparison
How much cross-bin influence is necessary/sufficient?
Red Car Retrievals (Color histograms)

\[ \chi^2(h_i, h_j) = \frac{1}{2} \sum_{m=1}^{K} \frac{[h_i(m) - h_j(m)]^2}{h_i(m) + h_j(m)} \]

Histogram matching distance
Capturing the "essence" of texture

...for real images

We don’t want an actual texture realization, we want a texture invariant.

What are the tools for capturing statistical properties of some signal?
Multi-scale filter decomposition

Filter bank

Input image
Filter response histograms
Heeger & Bergen’95

Start with a noise image as output

Main loop:

• Match pixel histogram of output image to input
• Decompose input and output images using multi-scale filter bank (Steerable Pyramid)
• Match sub-band histograms of input and output pyramids
• Reconstruct input and output images (collapse the pyramids)
Image Descriptors

- Blur + SSD
- Color / Texture histograms
- Gradients + Histogram (GIST, SIFT, HOG, etc)
- “Bag of Visual Words”
Scene Completion
Using Millions of Photographs

James Hays and Alexei A. Efros
Carnegie Mellon University
Efros and Leung result – no notion of semantics, also assumes necessary data is present elsewhere in the image
Scene Matching for Image Completion
Scene Completion Result
Challenges:

Computational costs of searching lots of images

Should fill in missing regions with semantically valid fragments

Scene Completion Result
The Algorithm

Input image
The Algorithm

Input image  Scene Descriptor

Compute a global description of the whole image
The Algorithm

Input image → Scene Descriptor → Image Collection

Compare to LOTS of images
The Algorithm

Input image ➔ Scene Descriptor ➔ Image Collection

Get top matches

200 matches

Hays and Efros, SIGGRAPH 2007
The Algorithm

Input image

Scene Descriptor

Image Collection

Compare more locally and merge pieces of matching images

Context matching + blending

Hays and Efros, SIGGRAPH 2007
The Algorithm

Input image → Scene Descriptor → Image Collection

20 completions (final results) → Context matching + blending → 200 matches

Hays and Efros, SIGGRAPH 2007
Data

We downloaded **2.3 Million** unique images from Flickr groups and keyword searches.

Groups: lonelyplanet, urban-fragments, ruraldecay ...
Keywords: outdoors, vacation, river...

Discard duplicates and small images

Hays and Efros, SIGGRAPH 2007
Scene Matching
Scene Descriptor

Compute oriented edge response at multiple scales (5 spatial scales, 6 orientations)
Scene Descriptor

Gist scene descriptor (Oliva and Torralba 2001)
“semantic” descriptor of image composition
aggregated edge responses over 4x4 windows
scenes tend to be semantically similar under this descriptor if very close

Hays and Efros, SIGGRAPH 2007
Scene Descriptor

Gist scene descriptor - with missing regions masked (weighted based on percentage of valid pixels)

Hays and Efros, SIGGRAPH 2007
Scene Descriptor

Color descriptor – color of the query image downsampled to 4x4

Distances calculated by SSD between query image descriptors & imgs in database

Total Dist = color dist + 2*gist dist

Gist scene descriptor (Oliva and Torralba 2001)
2 Million Flickr Images
... 200 total
Context Matching

Need to more precisely align matching scenes to local img context around missing region

local context = all pixels within 80 pixel radius of hole's boundary

Compute pixel-wise error of 200 best scene matches across all valid translations and 3 scales

Compute texture similarity of proposed fill-in to removed region

Hays and Efros, SIGGRAPH 2007
Result Ranking

We assign each of the 200 results a score which is the sum of:

- The scene matching distance
- The context matching distance (color + texture)
- The graph cut cost
Final result – blended between the two images along the cut to merge seamlessly
Pro – allows insertion of novel objects
... 200 scene matches

Hays and Efros, SIGGRAPH 2007
... 200 scene matches
... 200 scene matches
... 200 scene matches
... 200 scene matches
... 200 scene matches
Failures
Failures
Failures

Cause of failure – atypical scene caused lack of good matches
Failures
Failures
Failures
Failures

Hays and Efros, SIGGRAPH 2007
Failures

Cause of failure – fine scale texture mismatch

Hays and Efros, SIGGRAPH 2007
Failures
Failures
Failures

Cause of failure – no notion of “objects”
Evaluation
Real Image. This image has not been manipulated

or

Fake Image. This image has been manipulated
User Study Results - 20 Participants

![Graph showing the percentage of images marked as fake against maximum response time (seconds) for Criminisi et al., Our algorithm, and Real Photographs.]

Criminisi et al.
Our algorithm
Real Photographs

Hays and Efros, SIGGRAPH 2007
Why does it work?
10 nearest neighbors from a collection of 20,000 images
10 nearest neighbors from a collection of 2 million images
The Big Picture

Sky, Water, Hills, Beach,
Sunny, mid-day

Brute-force Image Understanding – insert semantically matching pieces by looking through millions of images.

Hays and Efros, SIGGRAPH 2007
Reading Assignment
