Big Visual Data – Part 1

Today’s Agenda

• 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.)

Much of Captured World is “Generic”

Portraits of Bill Clinton
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.

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

Subspace of natural images

Subspace of monkeys

Space of all images

Parametric model of monkeys

Two kinds of Things in the World

Navier-Stokes Equation

\[ \frac{\partial \mathbf{u}}{\partial t} = - (\mathbf{u} \cdot \nabla) \mathbf{u} + \nu \nabla^2 \mathbf{u} - \frac{1}{ho} \nabla p + \mathbf{f} \]

+ weather
+ location

Lots of data available

Non-parametric Approach

!!! HIGH DIMENSIONAL !!!

Subspace of natural images

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

[Halevy, Norvig, Pereira 2009]
http://www.youtube.com/watch?v=yvDCzhbjYWs

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

CLIME = CRIME = Hamming distance of 1 letter

= Euclidean distance of 5 units

= Gray value distance of 50 values

= ?
SSD says these are not similar

Tiny Images

80 million tiny images: A large dataset for non-parametric object and scene recognition, Antonio Torralba, Rob Fergus and William T. Freeman, PAMI 2008.

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_burmaide
  - a_e_housman
  - a_e_kennelly
  - a.e.
  - a_battery
  - a_cappella_singing
  - a_horizon
  - a_kempis
  - a_lborg
  - aails
  - aalost
  - aaita
  - aair
  - aardvark
  - aardwol
  - aare
  - aare_river

Thumbnail Collection

- 7 different search engines
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

Aardvark Images

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

Context!
Image Segmentation (by humans)

Non-parametric Classifier

- Nearest-neighbors
- For each query, obtain sibling set (neighbors)
- 3 different types of distance metric
- Hand-designed, use whole image

Human Scene Recognition

Metric 1 - $D_{SSD}$

- Sum of squared differences (SSD)

$$D^2_{SSD} = \sum (x, y, c)^2$$

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

- SSD but allow small transformations

\[ D_{warp}^2 = \min_\theta \sum_{x,y,c} \left( \text{Transformation:} \right) \]

Find min using gradient descent

Metric 3 - $D_{Shift}$

- As per Warping but also allow sub-window shifts

\[ D_{shift}^2 = \min_{\text{local sub-window}} \sum_{x,y,c} \left( \text{Image:} \right) \]

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
How Does $D_{SSD}$ Relate to Semantic Distance?

![Graph showing the relationship between pixelwise correlation and probability for same category for different subjects.](image)
Label Assignment

- Distance metrics give set of nearby images
- How to compute label?
  - Query
  - Grover Cleveland
  - Linnet
  - Birdcages
  - Chiefs
  - Casing

Issues:
- Labeling noise
- Keywords can be very specific
  - e.g. yellowfin tuna

Wordnet – A Lexical Dictionary

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

Sense 1
aardvark, ant bear, ant bear, 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

- Convert graph structure into tree by taking most common meaning

Wordnet Voting Scheme

a) Input image

b) Neighbors

One image – one vote
Vote for branches in Wordnet tree
Wordnet Voting Scheme

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

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

**Automatic Colorization**

Performance drops as classes become more specific

**Automatic Colorization**
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 Examples

Average correlation to 50 closest neighbors

Automatic Orientation

- Look at mean distance to neighbors

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

Images at wrong orientation have neighbors further away
Tiny Images Discussion

- Why SSD?
- Can we build a better image descriptor?

Image Representations: Histograms

- Global histogram
  - Represent distribution of features
  - Color, texture, depth, ...

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

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

- 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

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)

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

Image Descriptors

- Blur + SSD
- Color / Texture histograms
- Gradients + Histogram (GIST, SIFT, HOG, etc.)
- “Bag of Visual Words”

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)
Scene Completion Using Millions of Photographs

James Hays and Alexei A. Efros
Carnegie Mellon University

Diffusion Result
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

The Algorithm

1. Input image
2. Scene Descriptor
3. Compute a global description of the whole image
4. Compare to LOTS of images

Scene Completion Result
The Algorithm

- Input image
- Scene Descriptor
- Image Collection
- Get top matches
- 200 matches

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
Scene Matching

Scene Descriptor

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

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

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

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
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

Top 20 Results
Pro – allows insertion of novel objects
… 200 scene matches
… 200 scene matches

Failures
Failures

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

Cause of failure – fine scale texture mismatch
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

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

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

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