What is Internet Vision?

- Vast majority of data on Internet is in form of images/video
- Lots of unique applications of Computer Vision in this setting
- Also a very useful tool for vision researchers
  - Get labels for images
The Internet as source of labor

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find HITs Now

Slide credit: N. Snavely
Mechanical Turk – Demographics

- United States: 76.25%
- India: 8.03%
- United Kingdom: 3.34%
- Canada: 2.34%

Age distribution

Motivation
LABELING IMAGES WITH WORDS

MARTHA STEWART
FLOWERS
SUPER EVIL

STILL AN OPEN PROBLEM

Slides courtesy Luis von Ahn
IMAGE SEARCH ON THE WEB

USES FILENAMES AND HTML TEXT
THE ESP GAME

TWO-PLAYER ONLINE GAME

PARTNERS DON’T KNOW EACH OTHER AND CAN’T COMMUNICATE

OBJECT OF THE GAME:
TYPE THE SAME WORD

THE ONLY THING IN COMMON IS AN IMAGE
THE ESP GAME

PLAYER 1

GUESSING: CAR

GUESSING: HAT

GUESSING: KID

SUCCESS!
YOU AGREE ON CAR

PLAYER 2

GUESSING: BOY

GUESSING: CAR

SUCCESS!
YOU AGREE ON CAR
THE ESP GAME IS FUN

4.1 MILLION LABELS WITH 23,000 PLAYERS

THERE ARE MANY PEOPLE THAT PLAY OVER 20 HOURS A WEEK
SAMPLE LABELS

- Beach
- Chairs
- Sea
- People
- Man
- Woman
- Plant
- Ocean
- Talking
- Water
- Porch
REVEALING IMAGES

GUESSER

BAESH
GUESS

REVEALER

CAR

BAESH
PARTNER’S GUESS
Photo Collections

• Phototourism / Photosynth
  – Snavely, Szeliski and Seitz (Siggraph 2006)
Scene exploration

Path planning
Mapping the World’s Photos (35 million)

Slide credit: N. Snavely

[Crandal, Backstrom, Huttenlocher, Kleinberg, WWW ‘09]
Mapping the World’s Photos

[Slide credit: N. Snavely]

[Crandall, Backstrom, Huttenlocher, Kleinberg, WWW ‘09]
"Priors for Large Photo Collections and What They Reveal about Cameras,"
S. Kuthirummal, A. Agarwala, D. B Goldman, and S. K. Nayar,
European Conference on Computer Vision, 2010
Leveraging Huge Data

• What if we had millions or billions of images?
  – Facebook has $O(10^{10})$ images (10 Billion)
  – Roughly a lifetime of visual experience (5 glances/sec)

• What kind of new algorithms could we apply?
  – Brute Force methods
Scene Completion Using Millions of Photographs

James Hays and Alexei A. Efros
Carnegie Mellon University
Efros and Leung result
Criminisi et al. result
Criminisi et al. result
Scene Matching for Image Completion
The Algorithm

Input image → Scene Descriptor → Image Collection

20 completions → Context matching + blending → 200 matches
Data

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

Gist scene descriptor
(Oliva and Torralba 2001)
Scene Descriptor

Gist scene descriptor
(Oliva and Torralba 2001)
Scene Descriptor

Gist scene descriptor
(Oliva and Torralba 2001)
Context Matching
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
Top 20 Results
... 200 scene matches
... 200 scene matches
... 200 scene matches
... 200 scene matches
... 200 scene matches
Failures
Failures
Failures
Failures
Failures
Failures
Failures
Failures
Failures
Failures
Failures
Evaluation
Criminisi et al.          Scene Completion

Original Images          Criminisi et al.          Scene Completion
Single result          Each result selected from 20
Each result selected from 20

Original Images

Criminisi et al.

Scene Completion

Single result
Real Image. This image has not been manipulated

or

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

![Graph showing the performance of Criminisi et al. and Our algorithm compared to Real Photographs over maximum response time (seconds).]
Why does it work?
10 nearest neighbors from a collection of 20,000 images
10 nearest neighbors from a collection of 2 million images
Database of 70 Million 32x32 images

The Small Picture

Image Collection

Pixels

Pixels + Semantics
Hybrid Solution?

Pixels

Image Collection

Semantics
The Big Picture

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

Brute-force Image Understanding
80 Million Tiny Images

Antonio Torralba
Rob Fergus
William T. Freeman
Admin

- HW4 due on Thursday 12th May

- This is a hard deadline!

- The TA has to grade the assignment by Saturday so I can turn in grades
Overview

• Non-parametric approach to category-level recognition

• Dataset of 80 million images from Internet

• Use very low resolution images (32x32 color)
Overview

• Use simple algorithms: nearest neighbors
Motivation

Space of all images

Subspace of monkeys

Parametric model of monkeys
Non-parametric Approach

Subspace of natural images

Subspace of monkeys

Space of all images

Query image

!!! HIGH DIMENSIONAL !!!

!!! HIGH DIMENSIONAL !!!
Non-parametric Approach

Subspace of natural images

Subspace of monkeys

Query image

Space of all images

!!! HIGH DIMENSIONAL !!!
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)

$$D^2_{ssd} = \sum_{x,y,c} (\text{Image 1} - \text{Image 2})^2$$

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

- SSD but allow small transformations

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

Find min using gradient descent

Translation:

Horizontal flip:

Scalings:

Transformations $\theta$
Metric 3 - $D_{shift}$

- As per Warping but also allow sub-window shifts

\[
D_{shift}^2 = \sum_{x,y,c} [\text{Image 1}] - [\text{Transformed Image}]^2
\]

Start with warped version of image 2, as per $D_{warp}$
Metric 3 - $D_{shift}$

- As per Warping but also allow sub-window shifts

$$D_{shift}^2 = \sum_{x,y,c} \left[ \text{Transformed} \theta \right]^2$$

Start with warped version of image 2, as per $D_{warp}$
Metric 3 - $D_{shift}$

- As per Warping but also allow sub-window shifts

$$D^2_{shift} = \sum_{x,y,c} \left[ \begin{array}{c} \text{warped image 2} \\ \text{original image 2} \end{array} \right]^2$$

Start with warped version of image 2, as per $D_{warp}$
Metric 3 - $D_{shift}$

- As per Warping but also allow sub-window shifts

$$D_{shift}^2 = \min_{\text{Local sub-window}} \sum_{x,y,c}$$
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
Metric 3 - $D_{shift}$

- As per Warping but also allow sub-window shifts

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

Tried various sizes of sub-window

$\Rightarrow$ 1x1 (i.e. single pixel) worked best
Comparison of metrics

Target | SSD | Warping | Pixel shifting
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

Query image

Project into $K$ dimensional PCA basis

Take $M$ neighbors using L2 norm

Apply $D_{SSD}$, $D_{warp}$ & $D_{shift}$ to these $M$ images @ 32x32
Exact SSD vs Approximate SSD

Using $N=50$ neighbors

@ $p(\text{overlap}) = 0.8$

Overlap between approximate set & 50 true neighbors

# of neighbors in PCA space

$10^0$ $10^1$ $10^2$ $10^3$ $10^4$

$10^1$ $10^2$ $10^3$ $10^4$

# of neighbors in PCA space

@ $p(\text{overlap}) = 0.8$

$10^1$ $10^2$ $10^3$ $10^4$

# of true neighbors

$10^0$ $10^1$ $10^2$

Color codes:
- Cyan: 7,900
- Magenta: 79,000
- Yellow: 790,000
- Black: 7,900,000
- Red: 79,000,000
Quality of Sibling Set using $D_{\text{shift}}$

Size of dataset

- Target
- 7,900
- 790,000
- 79,000,000

$10^5$
$10^6$
$10^8$
Exploring the Sub-Space of Natural Images
How Many Images Are There?

Note: $D_1 = D_{SSD}$
Examples

Normalized correlation scores:

<table>
<thead>
<tr>
<th>term</th>
<th>score1</th>
<th>score2</th>
<th>score3</th>
<th>score4</th>
<th>score5</th>
<th>score6</th>
<th>score7</th>
<th>score8</th>
<th>score9</th>
</tr>
</thead>
<tbody>
<tr>
<td>skagerak</td>
<td>0.94</td>
<td>0.74</td>
<td>0.74</td>
<td>0.72</td>
<td>0.70</td>
<td>0.65</td>
<td>0.60</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>katmandu</td>
<td>0.93</td>
<td>0.92</td>
<td>0.91</td>
<td>0.90</td>
<td>0.85</td>
<td>0.80</td>
<td>0.75</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>noether</td>
<td>0.93</td>
<td>0.92</td>
<td>0.91</td>
<td>0.90</td>
<td>0.85</td>
<td>0.80</td>
<td>0.75</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>
How Many Images Are There?

Note: $D_1 = D_{SSD}$
How Does $D_{ssd}$ Relate to Semantic Distance?
Label Assignment

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

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

• Convert graph structure into tree by taking most common meaning
Wordnet Voting Scheme

a) Input image

b) Neighbors

c) Ground truth

d) Wordnet voted branches

One image – one vote
Classification at Multiple Semantic Levels

Votes:

- Animal: 6
- Person: 33
- Plant: 5
- Device: 3
- Administrative: 4
- Others: 22

Votes:

- Living: 44
- Artifact: 9
- Land: 3
- Region: 7
- Others: 10

1. d) Wordnet voted branches
Wordnet Voting Scheme

a) Input image

b) Neighbors

c) Ground truth

d) 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
Semantic vs Visual Hierarchy
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?
Person Recognition

- Subset where face >20% of image
Re-ranked Altavista Images

Original

Re-ranked
Object Classification

- Organism 0.75
- Artifact 0.71
- Location 0.81
- Food 0.73
- Geological formation 0.88
- Body of water 0.77
- Drug 0.75

- Person 0.87
- Animal 0.75
- Plant life 0.79
- Vehicle 0.80
- Mountain 0.86
- River 0.86

- Insect 0.82
- Bird 0.70
- Fish 0.86
- Car 0.85
- Flower 0.70
Object Classification

# images: 7,900

- Geological formation (32)
- Fish (29)
- Plant life (335)
- Flower (58)

---

Detection rate vs. false alarm rate:
- Geological formation
- Fish
- Plant life
- Flower

Precision vs. Recall:
- Geological formation
- Fish
- Plant life
- Flower
Other Applications
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

Grayscale input
High resolution

Grayscale
32x32 siblings

Color siblings
high resolution

Average of
color siblings
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
Automatic Orientation

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

• Hayes & Efros, Scene Completion using Millions of photographs, SIGGRAPH 2007.
• Barnard et al., Matching words and pictures. JMLR, 2003.
• Shakhnarovich et al. Fast pose estimation with parameter sensitive hashing, ICCV 2003
Conclusions

Model  Data

Few Data  Huge amounts of Data
Complex Model  No Model

• Can get good results simple algorithms & lots of data