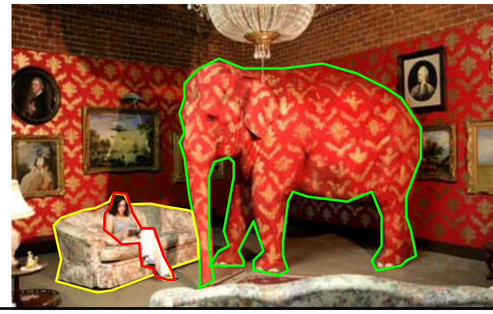
Large Image Databases and Small Codes for Object Recognition

Rob Fergus (NYU)
Antonio Torralba (MIT)
Yair Weiss (Hebrew U.)
William T. Freeman (MIT)

Object Recognition

Pixels → Description of scene contents



Banksy, 2006

Internet contains billions of images

Amazing resource

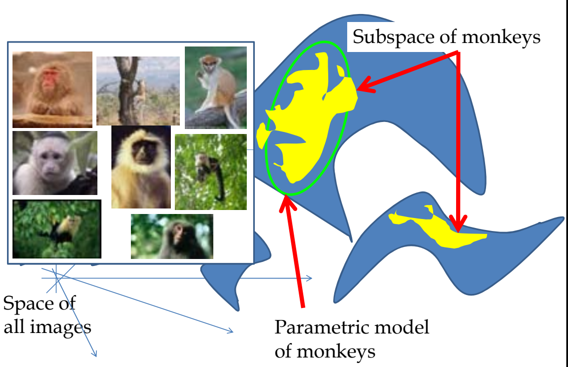
Maybe we can use it for recognition?

But so much data

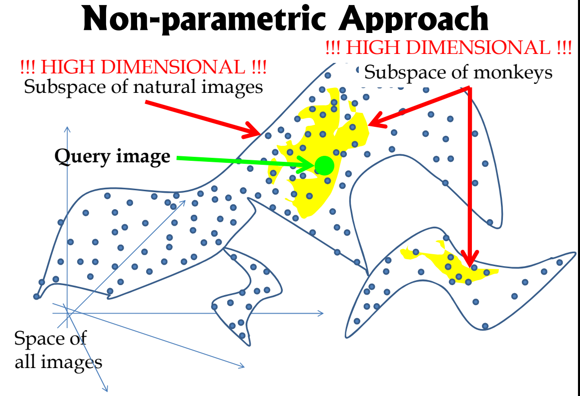
How can we search fast?

1. Big Data

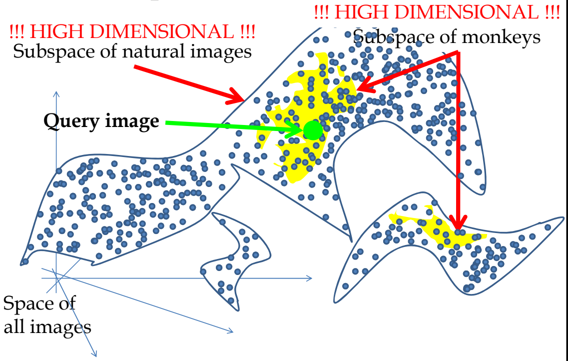
Parametric models



Non-parametric Approach



Non-parametric Approach



Large Collection of Internet Images

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

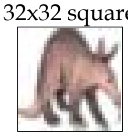
Thumbnail Collection

- 7 different search engines

Dataset Statistics

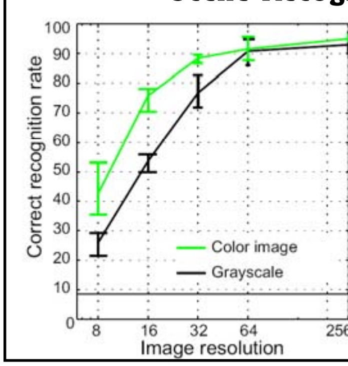
- Overall stats
  - 79,302,017 images
  - 75,062 different words
- Details
  - Two formats: square & rectangular
  - Gathered at 4.5 images/second
  - Downloaded 97,245,098 images
  - 18% duplicate rate
  - Disk usage: ~ 700Gb
  - Collection time: ~ 9 months

32x32 square



32xN rectangular

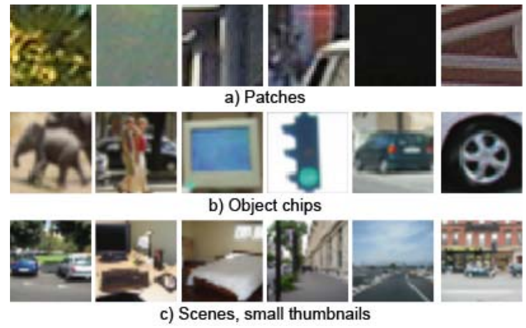
Human Performance at Scene Recognition



The role of context in object recognition
A. Oliva, A. Torralba
Trends in Cognitive Sciences, in press.
December 2007.

Human Labeling of Tiny Scenes

Image Patches vs Tiny Images



Recognition Approach

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)

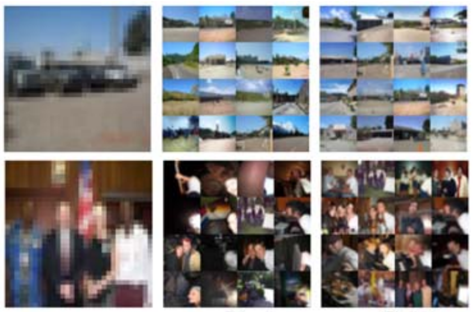
$$D_{ssd}^2 = \sum_{x,y,c} \left[\text{Image 1} - \text{Image 2} \right]^2$$

To give invariance to illumination:
Each image normalized to
be zero mean, unit variance

Comparison of metrics

Target | SSD Metric 1 | Warping Metric 2 | Pixel shifting Metric 3

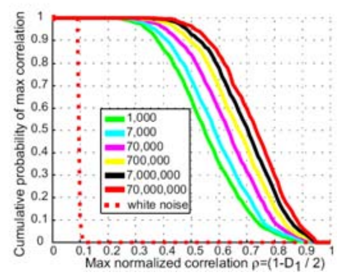
Sibling Sets with Different Metrics



Quality of Sibling Set using D_{shift}

Exploring the Sub-Space of Natural Images

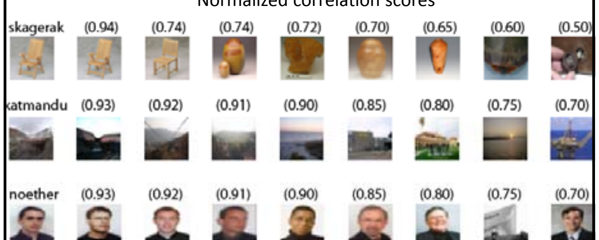
How Many Images Are There?



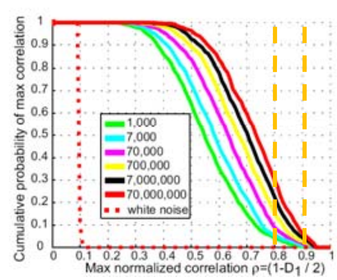
Note: $D_1 = D_{SSD}$

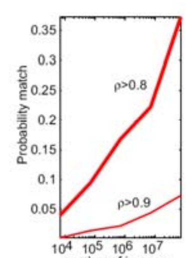
Examples

Normalized correlation scores



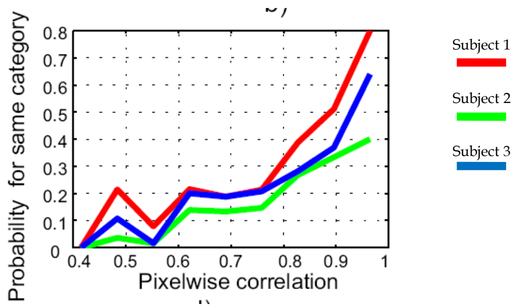
How Many Images Are There?





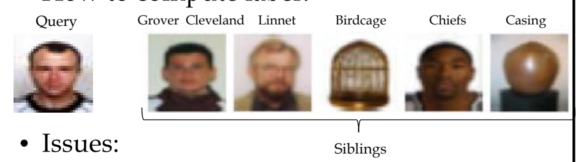
Note: $D_1 = D_{SSD}$

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

Wordnet – a Lexical Dictionary

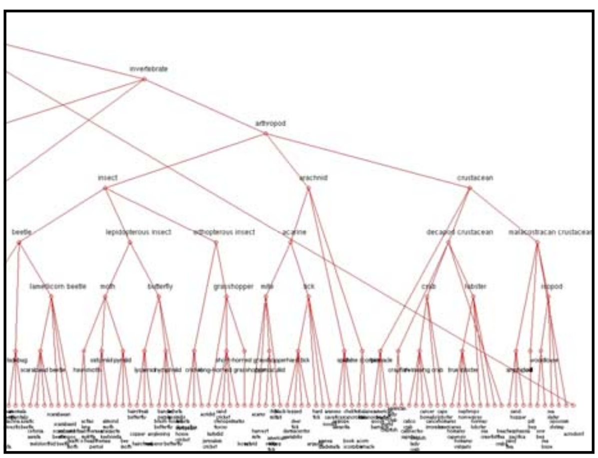
http://wordnet.princeton.edu/

```
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
```

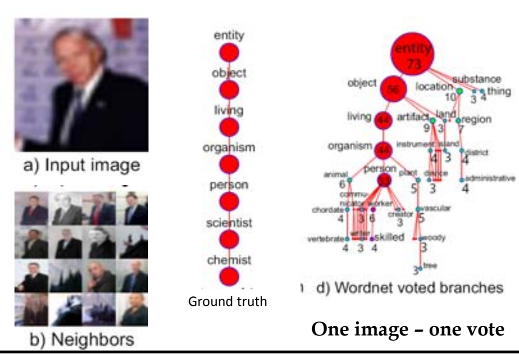
Wordnet Hierarchy

```
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

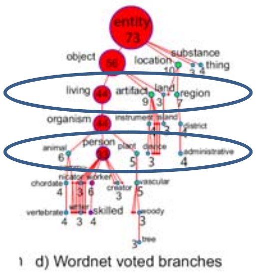


Wordnet Voting Scheme



One image – one vote

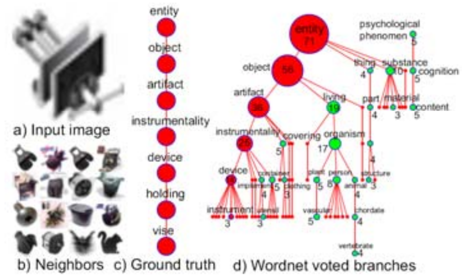
Classification at Multiple Semantic Levels



Votes:

Living / Animal	6	44
Artifact / Person	33	9
Land / Plant		5 / 3
Region / Device	3 / 7	
Others / Administrative	4 / 10	
Others	22	

Wordnet Voting Scheme



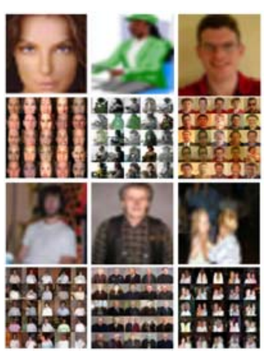
Wordnet Voting

- 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

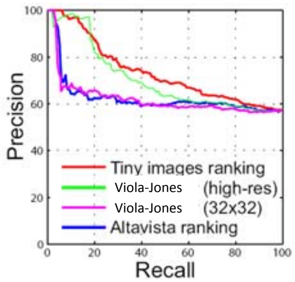


Person Recognition – Test Set

- 1016 images from Altavista using "person" query
- High res and 32x32 available
- Disjoint from 79 million tiny images

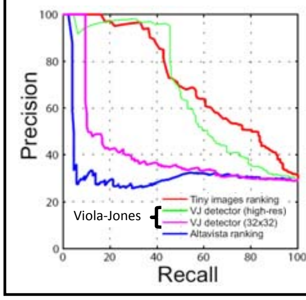
Person Recognition

- Task: person in image or not?



Person Recognition

- Subset where face >20% of image

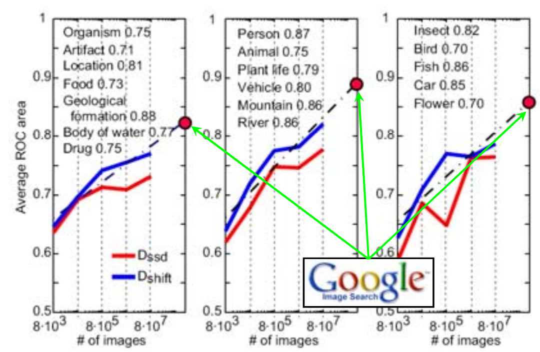


Re-ranked Altavista Images

Original

Re-ranked

Object Classification



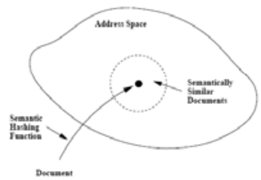
So far….

- Surprising performance from non-parametric methods
- But so slow…..
- ~ 1 Minute to find neighbors in 80 million
  - Essentially a brute force scheme

2. Small codes

Learning to retrieve quickly

- Semantic Hashing
  - Salakhutdinov & Hinton, SIGIR 2007
  - Text documents
- Non-linear dimensionality reduction of data to binary codes
- Preserve semantic distance
- Hamming ball search
  - Hamming distance → # different bits
  - Direct memory lookup via bit flips
  - Lookup time independent of # data points



Compact Binary Codes

- Google has few billion images (10^9)
- Big PC has ~10 Gbytes (10^{11} bits)

→ Budget of 10^2 bits/image

- 1 Megapixel image is 10^7 bits
- 32x32 color image is 10^4 bits

→ Need serious dimensionality reduction!!

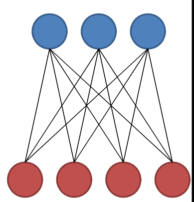
Restricted Boltzmann Machine (RBM) architecture

- Network of binary stochastic units
- Hinton & Salakhutdinov, Nature 2006

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i\in\text{visible}} b_i v_i - \sum_{j\in\text{hidden}} b_j h_j - \sum_{i,j} v_i h_i w_{ij}$$

Parameters: Weights w Biases b

$$p(\mathbf{v}) = \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{v},\mathbf{h})}}{\sum_{\mathbf{u},\mathbf{g}} e^{-E(\mathbf{u},\mathbf{g})}}$$



RBM architecture

- Network of binary stochastic units
- Hinton & Salakhutdinov, Nature 2006

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i\in\text{visible}} b_i v_i - \sum_{j\in\text{hidden}} b_j h_j - \sum_{i,j} v_i h_i w_{ij}$$

Parameters: Weights w Biases b

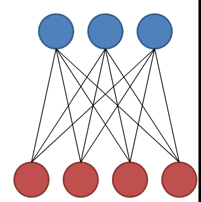
Convenient conditional distributions:

$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i w_{ij} v_i)$$

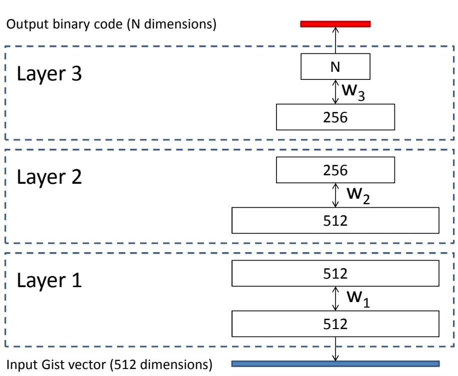
$$p(v_i = 1|\mathbf{h}) = \sigma(b_i + \sum_j w_{ij} h_j)$$

$\sigma(x) = 1/(1+e^{-x})$, the logistic function

Learn weights and biases using Contrastive Divergence



Multi-Layer RBM architecture

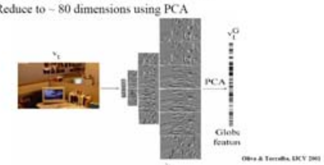


Input to RBM: Gist vectors

- Difficult to train directly on pixels
- Use GIST descriptor instead

Feature vector for an image: the "gist" of the scene

- Compute 12 x 30 = 360 dim. feature vector
- Or use steerable filter bank, 6 orientations, 4 scales, averaged over 4x4 regions = 384 dim. feature vector
- Reduce to ~ 80 dimensions using PCA

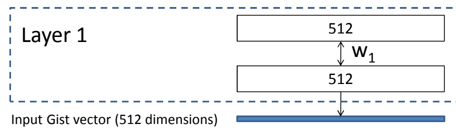


Training RBM models

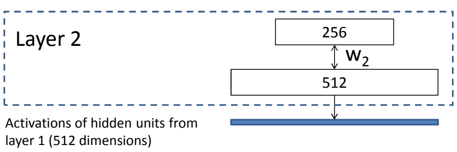
- Two phases

1. Pre-training
   - Unsupervised
   - Use Contrastive Divergence to learn weights and biases
   - Gets parameters to right ballpark
2. Fine-tuning
   - Supervised
   - No longer stochastic
   - Backpropagate error to update parameters
   - Moves parameters to local minimum

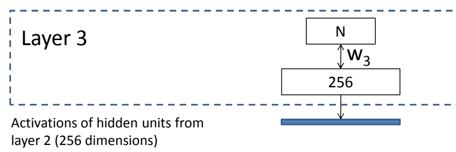
Greedy pre-training (Unsupervised)



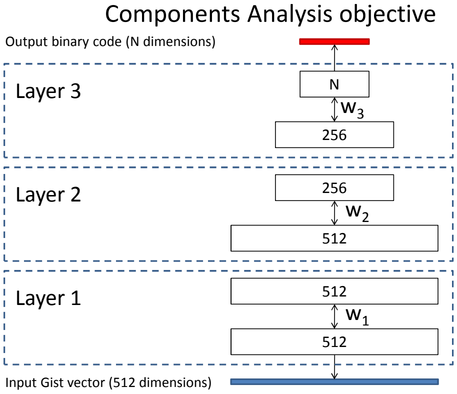
Greedy pre-training (Unsupervised)



Greedy pre-training (Unsupervised)



Backpropagation using Neighborhood Components Analysis objective



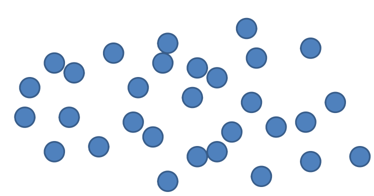
Neighborhood Components Analysis

- Goldberger, Roweis, Salakhutdinov & Hinton, NIPS 2004

$$O_{\mathrm{NCA}} = \sum_{k=1}^{K} \sum_{l:c^k=c^l} p_{kl} \qquad p_{kl} = \frac{e^{-\|f(\mathbf{x}^k|W)-f(\mathbf{x}^l|W)\|^2}}{\sum_{m\neq l} e^{-\|f(\mathbf{x}^m|W)-f(\mathbf{x}^l|W)\|^2}}$$

Output of RBM

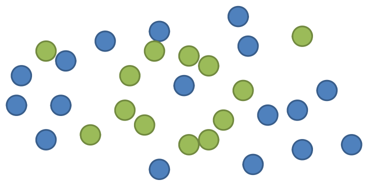
W are RBM weights



Neighborhood Components Analysis

- Goldberger, Roweis, Salakhutdinov & Hinton, NIPS 2004

$$O_{\mathrm{NCA}} = \sum_{k=1}^{K} \sum_{l:c^k=c^l} p_{kl} \qquad p_{kl} = \frac{e^{-||f(\mathbf{x}^k|W)-f(\mathbf{x}^l|W)||^2}}{\sum_{m \neq l} e^{-||f(\mathbf{x}^m|W)-f(\mathbf{x}^l|W)||^2}}$$

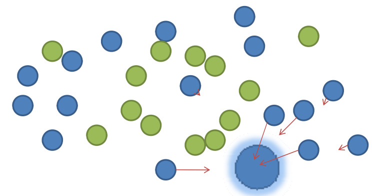


Assume K=2 classes

Neighborhood Components Analysis

- Goldberger, Roweis, Salakhutdinov & Hinton, NIPS 2004

$$O_{\mathrm{NCA}} = \sum_{k=1}^{K} \sum_{l:c^k=c^l} p_{kl} \qquad p_{kl} = \frac{e^{-||f(\mathbf{x}^k|W)-f(\mathbf{x}^l|W)||^2}}{\sum_{m \neq l} e^{-||f(\mathbf{x}^m|W)-f(\mathbf{x}^l|W)||^2}}$$

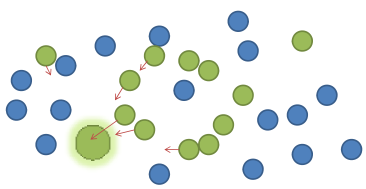


Pulls nearby points
OF SAME CLASS
closer

Neighborhood Components Analysis

- Goldberger, Roweis, Salakhutdinov & Hinton, NIPS 2004

$$O_{\mathrm{NCA}} = \sum_{k=1}^{K} \sum_{l:c^k=c^l} p_{kl} \qquad p_{kl} = \frac{e^{-||f(\mathbf{x}^k|W)-f(\mathbf{x}^l|W)||^2}}{\sum_{m \neq l} e^{-||f(\mathbf{x}^m|W)-f(\mathbf{x}^l|W)||^2}}$$



Pulls nearby points
OF SAME CLASS
closer

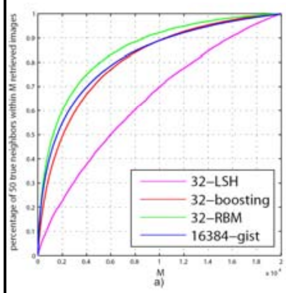
Goal is to preserve neighborhood structure of original, high dimensional, space

Two test datasets

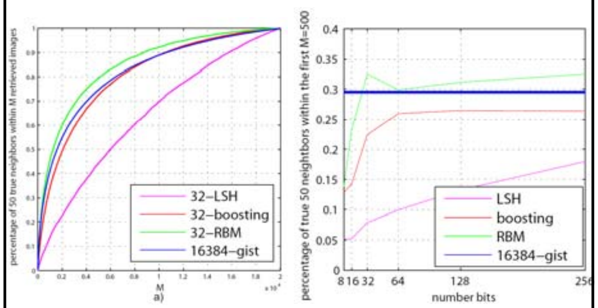
- LabelMe
  - 22,000 images
  - Ground truth segmentations for all
  - Can define distance btw. images using these segmentations
- Web data
  - 12.9 million images
  - Subset of 80 million images
  - No labels, so use L2 distance btw. GIST vectors as ground truth

Retrieval Experiments

LabelMe Retrieval

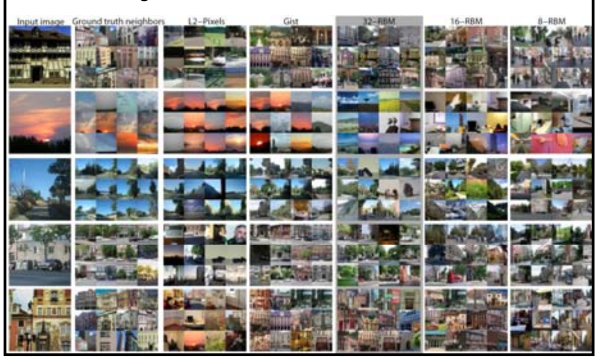


LabelMe Retrieval

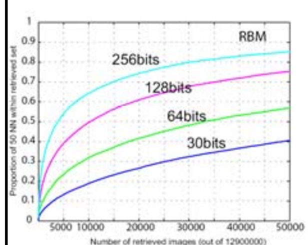


Examples of LabelMe retrieval

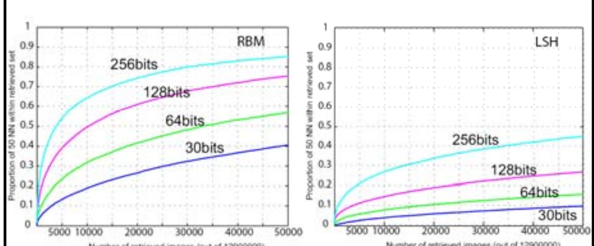
- 12 closest neighbors under different distance metrics



Web images retrieval



Web images retrieval



Examples of Web retrieval

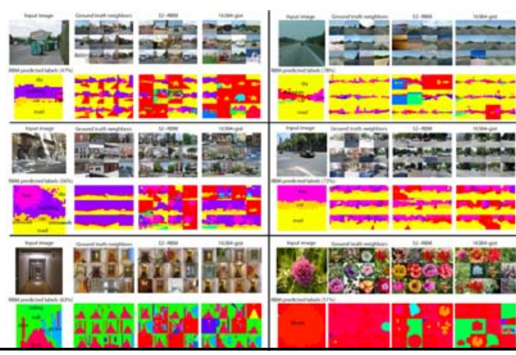
- 12 neighbors using different distance metrics

Retrieval Timings

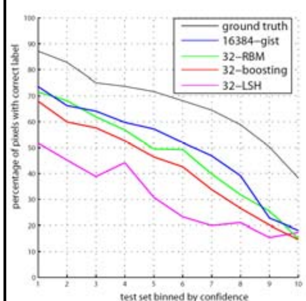
Dataset	LabelMe	Web
# images	2×10^4	1.29×10^7
Gist vector dim.	512	384
Method	Time (s)	Time (s)
Spill tree - Gist vector	1.05	-
Brute force - Gist vector	0.38	-
Brute force - 30 bit binary	4.3×10^{-4}	0.146
" - 30 bit binary, M/T	2.7×10^{-4}	0.074
Brute force - 256 bit binary	1.4×10^{-3}	0.75
" - 256 bit binary, M/T	4.7×10^{-4}	0.23
Hashing - 30 bit binary	6×10^{-6}	6×10^{-6}

Recognition Experiments

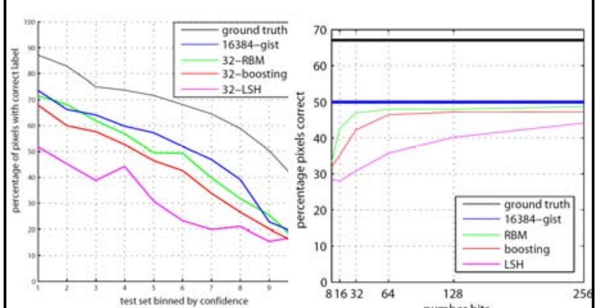
LabelMe Recognition examples



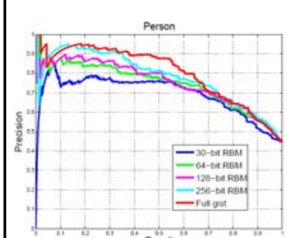
LabelMe Recognition



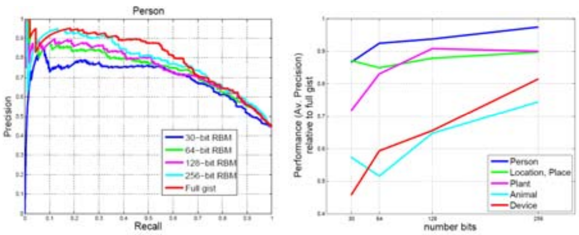
LabelMe Recognition



Web dataset Recognition



Web dataset Recognition



Conclusions

- Can do interesting things with lots of data
  - What would happen with Google's ~ 2 billion images?
- Possible to build compact codes for retrieval
  - Much room for improvement