CS395T: Visual Recognition and Search Leveraging Internet Data

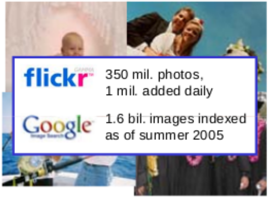
Birgi Tamersoy

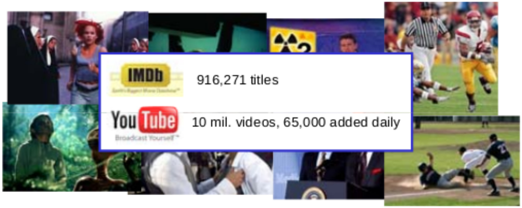
March 27, 2009

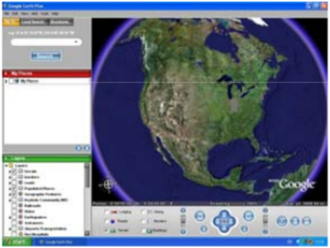
Theme I

L. Lazebnik

Theme II







L. Lazebnik

Theme III

Situated search
Yeh et al., MIT

kooaba

MSR Lincoln

K. Grauman

Outline

Scene Segmentation Using the Wisdom of Crowds by I. Simon and S.M. Seitz

World-scale Mining of Objects and Events from Community Photo Collections by T. Quack, B. Leibe and L. Van Gool

80 Million Tiny Images: a Large Dataset for Non-parametric Object and Scene Recognition by A. Torralba, R. Fergus and W.T. Freeman

Introduction [Wisdom of Crowds]

Goal

Given a set of images of a static scene, identify and segment the interesting objects in the scene.

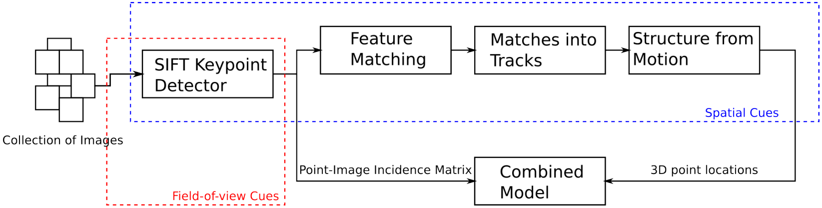
Observations

- The *distribution* of photos in a collection holds valuable semantic information.
- Interesting objects will be frequently photographed.
- Detecting interesting features is straightforward, but identifying interesting objects is more challenging.
- Features on the same object will appear together in many photos.

Field-of-view cue

Co-occurrence information is used to group features into objects.

Big Picture



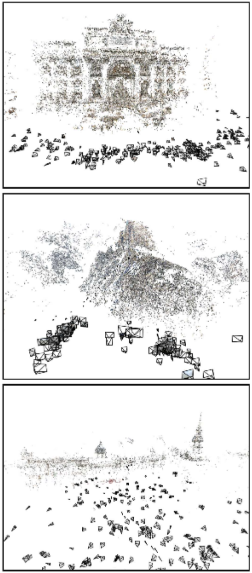
Collection of Images
SIFT Keypoint Detector
Feature Matching
Matches into Tracks
Structure from Motion
Spatial Cues
Point-Image Incidence Matrix
Combined Model
3D point locations
Field-of-view Cues

Spatial Cues I

Algorithm

1. Find feature points in each image using SIFT keypoint detector.
2. For each pair of images, match the detected feature points.
3. Robustly estimate a fundamental matrix for the pair using RANSAC (RAndom SAmple Consensus) and remove the outliers.
4. Organize the matches into tracks.
   - A track is a connected set of of matching keypoints across multiple images.
5. Recover camera parameters and a 3D location for each track.

Spatial Cues II



Snavely et al.

- A single 3D Gaussian distribution per object to enforce spatial cues.
- A mixture of Gaussians to model the spatial cues from multiple objects.

$$P(C, X|\pi, \mu, \Sigma) = \prod_j P(c_j|\pi)P(x_j|c_j, \mu, \Sigma)$$

- A class variable c_j is associated with each point x_i. Drawn from a multinomial distribution.
- Point locations are drawn from 3D Gaussians, where the point class determines which Gaussian to use.

Field-of-view Cues

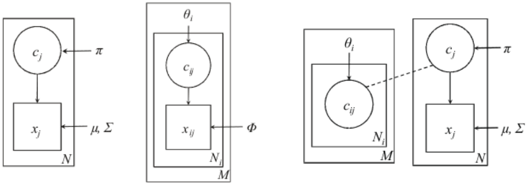
pLSA

Co-occurrence information is modeled by Probabilistic Latent Semantic Analysis (pLSA).

$$P(C, X|\theta, \phi) = \prod_{i} \prod_{j|x_j \in V_i} P(c_{ij}|\theta_i) P(x_{ij}|c_{ij}, \phi)$$

- A class variable c_{ij} for each point-image incidence.
- In original pLSA, x_{ij} would be the number of times word j appears in document i.

Combined Model



Simon and Seitz

$$P(C, X|\theta, \pi, \mu, \Sigma) = \left(\prod_i \prod_{j|x_j \in V_i} P(c_{ij}|\theta_i)\right) \times \left(\prod_j P(c_j|\pi) P(x_j|c_j, \mu, \Sigma)\right)$$

- This joint density is locally maximized using the EM algorithm.

Evaluation I

- For each test scene, the ground truth clusterings C^* are manually created.
- Three different models, mixture of Gaussians, pLSA and the combined model, are all tested.
- Computed clusterings are evaluated using Meila's Variation of Information (VI) metric:

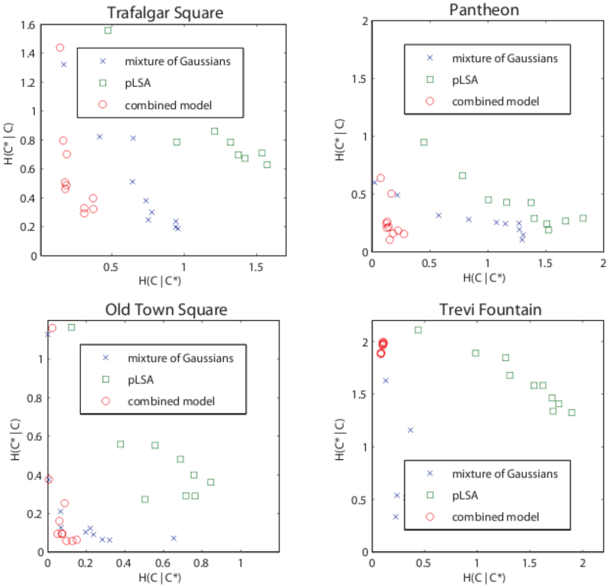
$$VI(C, C^*) = H(C|C^*) + H(C^*|C)$$

- The two terms represent the conditional entropies; information lost and gained between the two clusterings.

	Trafalgar	Pantheon	Hagia Sophia	Trevi	Prague	Navona
mixture of Gaussians	1.15	1.36	0.63	**0.81**	0.35	0.68
pLSA	2.07	1.70	0.64	3.12	1.13	1.46
combined model	**0.69**	**0.38**	**0.53**	2.07	**0.20**	**0.45**

Simon and Seitz

Evaluation II



Simon and Seitz

Importance Viewer

- Interesting objects appear in many photos.
- Penalize objects for size for not to reward the large background objects.

$$imp(c) = \alpha \frac{1}{|\Sigma_c|} \sum_i \theta_i(c)$$

Simon and Seitz

Region Labeling

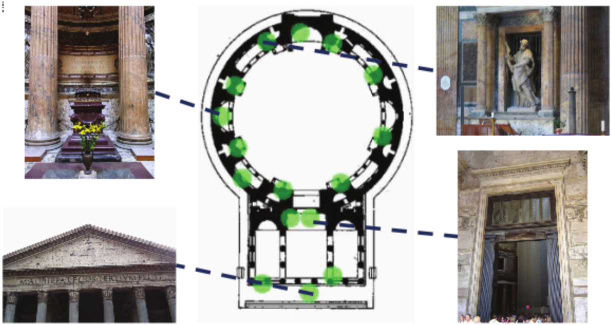
- Image tags in the Internet are very noisy.
- Accurate tags could be computed by examining tag-cluster co-occurrence statistic.
- Score of each cluster c tag t pair is given by:

$$score(c,t) = P(c,t)(\log P(c,t) - \log P(c)P(t))$$

Simon and Seitz

Interactive Map Viewer

- After the scene is segmented, the scene points are manually aligned with an overhead view.



Simon and Seitz

Summary

- Field-of-view cue is introduced.
- Field-of-view cues are incorporated with spatial cues to identify the interesting objects of a scene.
- Source of the information: distribution of photos, ie. wisdom of crowds.

Introduction [World-scale Object Mining]

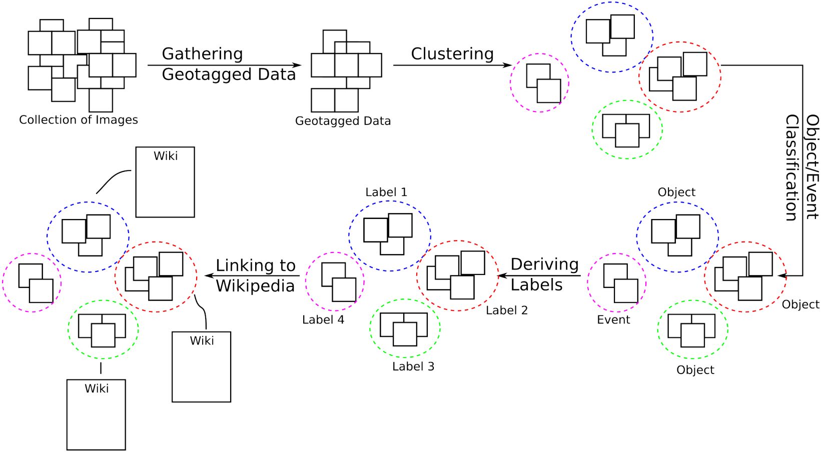
Goal

Automated collection of a high quality image database with correct annotations.

Observations

- Large databases of visual data is available from community photo collections.
- More and more images are "geotagging".
- Geotags and textual tags are sparse and noisy.

Big Picture



Collection of Images
Gathering
Geotagged Data
Geotagged Data
Clustering
Object/Event
Classification
Object
Object
Event
Object
Deriving
Labels
Label 1
Label 2
Label 4
Label 3
Linking to
Wikipedia
Wiki
Wiki
Wiki

Gathering the Data

Quack et. al.

- Earth's surface is divided into tiles.
- High overlap between tiles.
- 70.000 tiles are processed (52.000 containing no images at all).

Photo Clustering

1. Dissimilarity matrices are computed for several modalities:
   - Visual cues.
   - Textual cues.
   - (User/timestamps cues.)
2. A hierarchical clustering step is used to create clusters of photos for the same object or event.

Visual Features and Similarity I

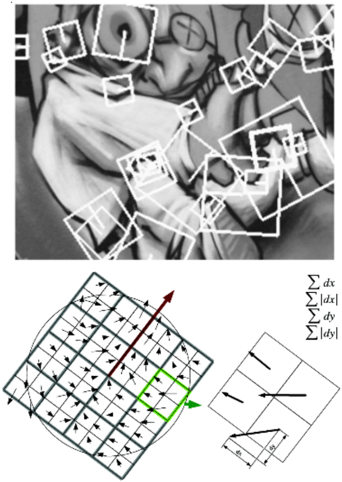
1. Extract SURF features from each photo of the tile.
2. For each pair of images find the matching features.
3. Estimate homography H between the two images using RANSAC.
4. Create the distance matrix using the number of "inlier" feature matches I_{ij} for each image pair:

$$d_{ij} = \begin{cases} \frac{I_{ij}}{I_{max}} & \text{if } I_{ij} \geq 10 \\ \infty & \text{if } I_{ij} < 10 \end{cases}$$

Visual Features and Similarity II

Speeded-Up Robust Features by Bay et. al.

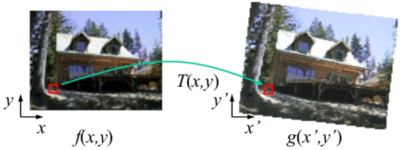
- Scale- and rotation-invariant detector and descriptor.
- At each step integral images are used to get very fast detections and descriptions.
- A box filter approximation of the Hessian matrix is used as the underlying filter.
- The 64-dimensional SURF descriptor describes the distribution of the intensity content within the interest point neighborhood.



Bay et. al.

Visual Features and Similarity III

Homography

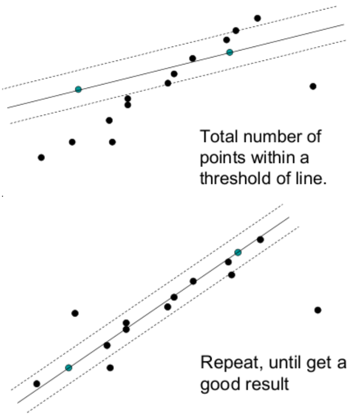


K. Grauman

$$p' = Hp$$

$$\begin{bmatrix} wx' \\ wy' \\ w \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

RANdom SAmple Consensus



K. Grauman

Text Features and Similarity

1. Three meta-data (tags, title and description) are combined to form a single text per image.
2. Image specific stop lists are applied.
3. Pairwise text similarities are computed to create the distance matrix.

Term weighting

$$w_{i,j} = L_{i,j} * G_i * N_j$$

$$\begin{aligned} L_{i,j} &= \frac{\log tf_{i,f} + 1}{\sum_j (\log(tf_{i,f} + 1)} \\ G_i &= \log \frac{D - d_i}{d_i} \\ N_j &= \frac{U_j}{1 + 0.0115 * U_j} \end{aligned}$$

where U_j is the number of unique terms in image j.

Clustering

- Hierarchical agglomerative clustering is applied to the computed distance matrices with the following cut-off distances:

	Visual	Text
Single-link	0.985	0.989
Complete-link	0.99	0.99
Average-link	0.99	0.99

Quack et. al

- Three different linkage methods are employed in order to capture different visual properties:

$$\begin{aligned}
\text{single-link}: \quad d_{AB} &= \min_{i \in A, j \in B} d_{ij} \\
\text{complete-link}: \quad d_{AB} &= \max_{i \in A, j \in B} d_{ij} \\
\text{average-link}: \quad d_{AB} &= \frac{1}{n_i n_j} \sum_{i \in A, j \in B} d_{ij}
\end{aligned}$$

Classification into Objects and Events

- Two features are extracted by using only the meta-data of the images in a tile:
  - Number of unique days the photos in a cluster were taken at.
  - The number of different users who "contributed" photos to this cluster divided by the cluster size.

- An individual ID3 decision tree is trained for each class.
  - 88% precision for objects and 94% precision for events.

Quack et. al

Labeling the Objects

- "Correct" labels of a cluster are found using frequent itemset mining.
- Top 15 itemsets are kept per cluster.

Frequent Itemset Mining

Let $I = \{i_1 \cdots i_p\}$ be a set of p words. The text of each image in the tile is a subset of I, $T \subseteq I$. The text of all images in a tile forms the database D. The goal is to find an itemset $A \subseteq T$, which has relatively high support:

$$supp(A) = \frac{|\{T \in D | A \subseteq T\}|}{|D|} \in [0, 1]$$

Linking to Wikipedia

1. Each itemset is used as a query to Google (search is limited to Wikipedia articles.
2. Images in the article are compared with the images in the cluster.
3. If there is a match, the query is kept as a label, otherwise it is rejected.

Experiments

- 70.000 tiles covering approximately 700 square kilometers.
- Over 220.000 images.
- Over 20.000.000 similarities (only 1 million being greater than 0).
- At the end, 73.000 images could be assigned to a cluster.

Object Clusters

Quack et. al

Event Clusters

Quack et. al

Linkage Methods

Single-link

Complete-link
Quack et. al

Summary

- World surface is divided into tiles.
- Images belonging to a tile are identified using geotags.
- These images are clustered.
- Clusters are classified as objects or events.
- Object labels are determined, and additional information from the Internet is linked to these objects.
- FULLY UNSUPERVISED!!!

Introduction [80 Million Tiny Images]

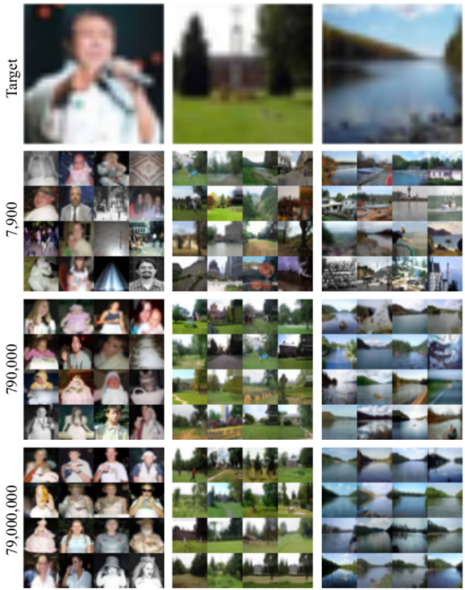
Goal

Creating an image database that densely populates the manifold of natural images, allowing the use of non-parametric approaches.

Observations

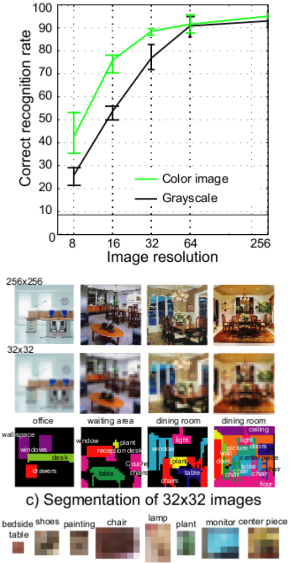
- Billions of images are available on the Internet.
- Human vision system has a remarkable tolerance to degradations in image resolutions.
- Visual world is very regular limiting the space of possible images significantly.

Big Picture



Torralba et. al

Low Dimensional Image Representation



d) Cropped objects

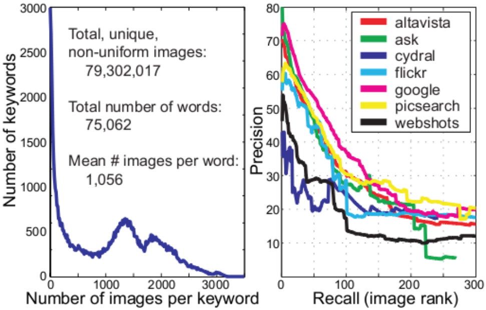
Torralba et. al

- 32×32 color images contain enough information for scene recognition, object detection and segmentation (for humans).
- Two advantages of low resolution representation:
  - Intrinsic dimensionality of the manifold gets much smaller.
  - Storing and efficient indexing of vast amounts of data points becomes feasible.
- It is important that information is not lost, while the dimensionality is reduced.

A Large Dataset of 32×32 Images I

1. 75.062 non-abstract nouns are extracted from Wordnet.
2. 7 independent search engines are searched for all of the images belonging to one of these categories.
3. In 8 mounts 97.245.098 images are collected.
4. Duplicates and uniform images are eliminated to form the final dataset of 79.302.017 images.

A Large Dataset of 32×32 Images II



Torralba et. al

Keywords

Around 10% of keywords have very few images. Mean number of images per word: 1.056.

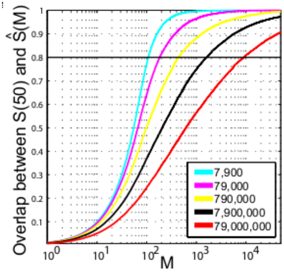
Labeling Noise

The dataset is not cleaned up. Often visual content is unrelated to the query word.

Dataset Statistics

Dataset Size

Distance between two images can be approximated using few principal components.



Similarity Measures

$$D_{ssd}^2 = \sum_{x,y,c} (I_1(x,y,c) - I_2(x,y,c))^2$$

$$D_{warp}^2 = \sum_{x,y,c} (I_1(x,y,c) - T_\theta[I_2(x,y,c)])^2$$

$$D_{shift}^2 = \sum_{x,y,c} (I_1(x,y,c) - \hat{I}_2(x+D_x, y+D_y, c)])^2$$

Wordnet Voting Scheme in Recognition I

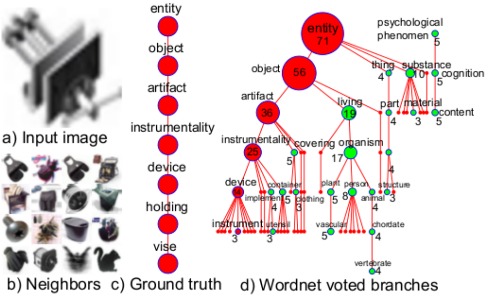
Recognition

Rather than relying on complex matching schemes, let the data do the work.

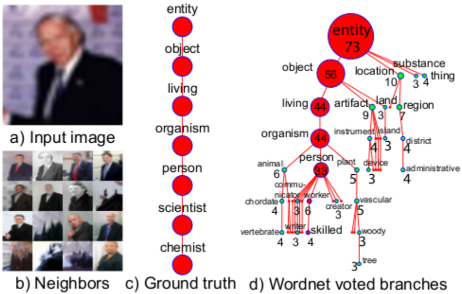
Wordnet Voting Scheme for Labeling Noise

- Given a query image find the nearest neighbors using some similarity measure.
- Each neighbor votes for its branch in the Wordnet tree.
- Classification at a specific level is done with respect to the votes.

Wordnet Voting Scheme in Recognition II



b) Neighbors c) Ground truth d) Wordnet voted branches

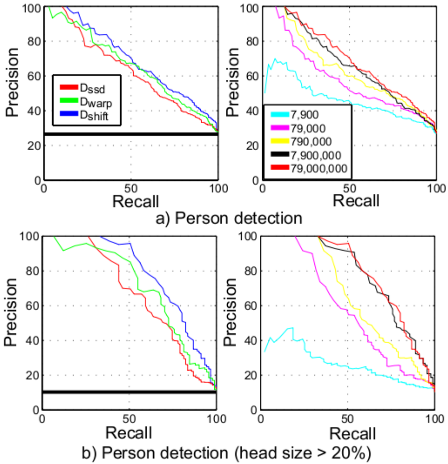


b) Neighbors c) Ground truth d) Wordnet voted branches

Torralba et. al

Person Detection I

- 23% of the images contain people in it.
- Hence, the corresponding region in the manifold is covered very densely.



Torralba et. al

Person Detection II

Torralba et. al

Person Localization

- Segment the input image using normalized cuts (10 segments).
- Query the dataset using cropped continuous segments.

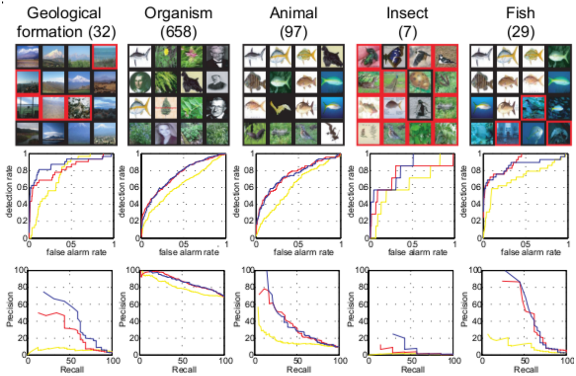
Torralba et. al

Scene Recognition

- Scene images will have a high count accumulated at the "location" node of the Wordnet tree.

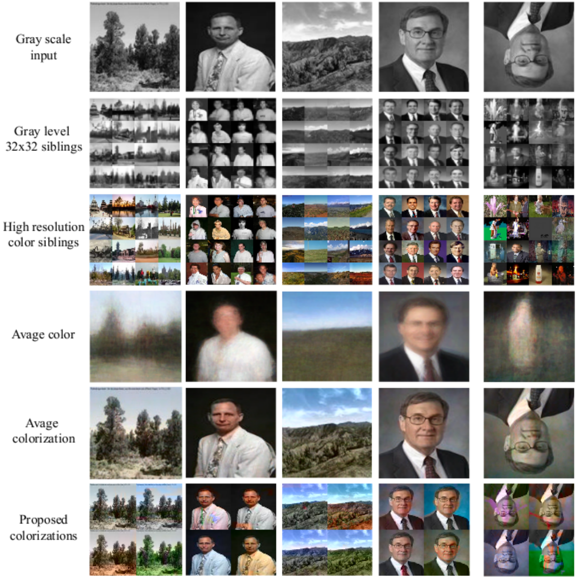
Torralba et. al

Automatic Image Annotation



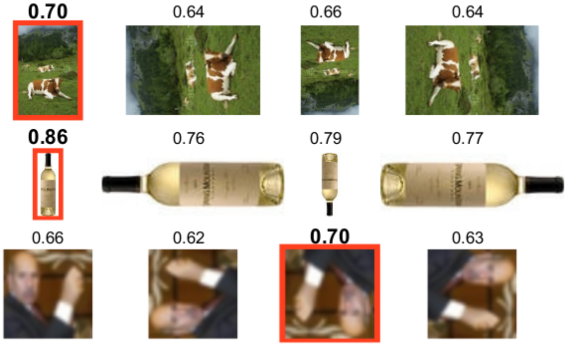
Torralba et. al

Image Colorization



Torralba et. al

Detecting Image Orientation



Torralba et. al

Summary

- Data should do the work, not us!!!
- 32×32 color images are enough for most of the computer vision tasks.
- Covering the manifold of natural images densely, so that for every query image there will be a semantically very similar image in the database.

Final Word

- Wisdom of Crowds: importance viewer, region labeling, interactive map viewer.
- World-scale Mining of Objects: recognition, automatic annotation.
- 80 Million Tiny Images: detection, recognition, localization, automatic annotation, etc.

Dataset Size

1. Wisdom of Crowds
2. World-scale Mining of Objects
3. 80 Million Tiny Images

Complexity

1. 80 Million Tiny Images
2. World-scale Mining of Objects
3. Wisdom of Crowds

References

- Scene Segmentation Using the Wisdom of Crowds by I. Simon and S. M. Seitz
- Photo Tourism: Exploring Photo Collections in 3D by N. Snavely, S. M. Seitz and R. Szeliski
- Computing Clusterings - an Information Based Distance by M. Meila
- World-scale Mining of Objects and Events from Community Photo Collections by T. Quack, B. Leibe and L. Van Gool
- Speeded-Up Robust Features (SURF) by H. Bay, A. Ess, T. Tuytelaars and L. Van Gool
- 80 million tiny images: a large dataset for non-parametric object and scene recognition by A. Torralba, R. Fergus and W. T. Freeman
- Dr. Kristen Grauman's CS378 (Fall 2008) and CS395T (Spring 2009) lecture slides