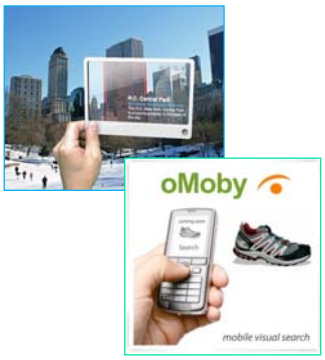
EE 6882
Visual Search Engine
Lec. 1: Introduction

tinyeye, photo copy search



mobile search

Web image search

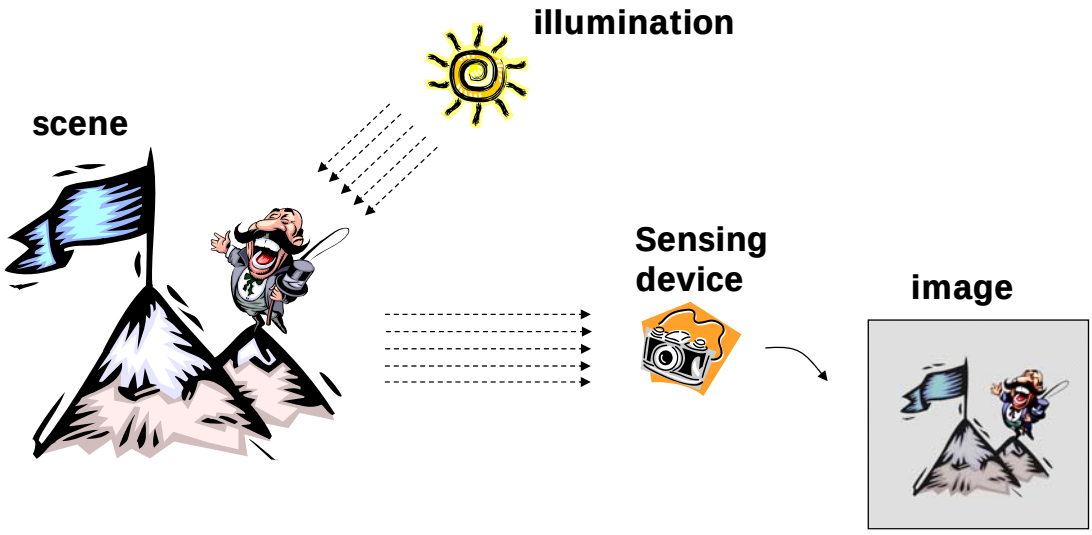
Jan. 23 2012

Demos: Google Image Google Goggles photo copy search

Topics of Interest

- How is visual information represented?
- How are images matched?
  - How to handle distortion and occlusion?
- How to handle gigantic database?
  - 36 billions photos uploaded to Facebook per year
- Possibility of semantic image tagging?
  - How to combine multimodal information?
- How to design search interfaces for multimedia?
  - For different purposes: information, entertainment, networking
- How to present multimedia search results?
  - Summarization and augmented reality

Visual Information Generation



Visual Representation and Features

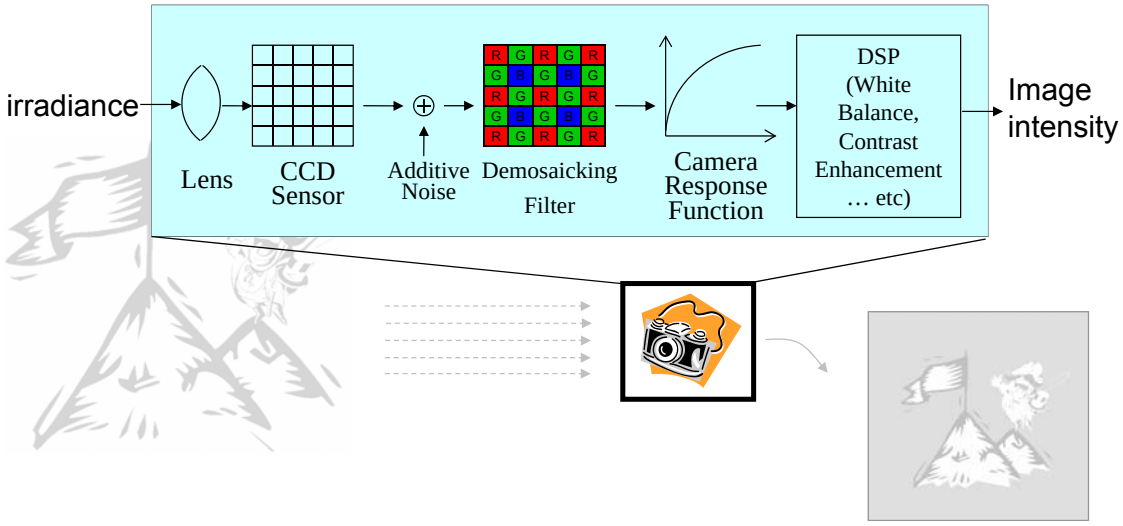
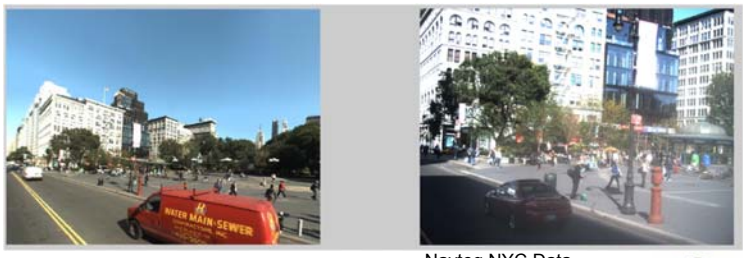


Image quality not always perfect

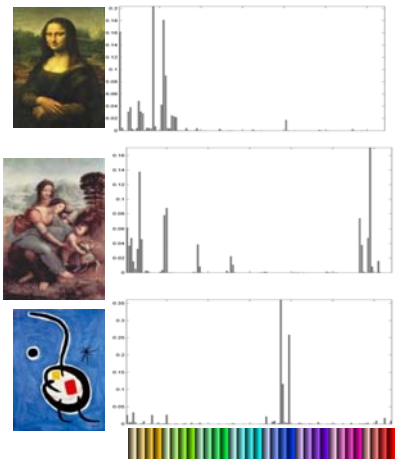
- Image quality variations
  - Exposure
  - Shadow
  - Distance
  - Obstruction
  - Blur
  - Weather
  - Day/Night



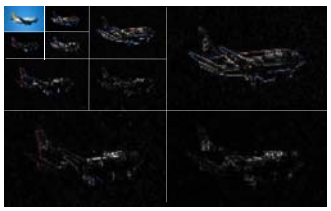
Navteq NYC Data

Visual Representation: Global Features

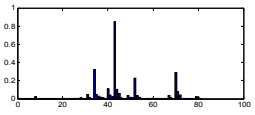
Color



Texture



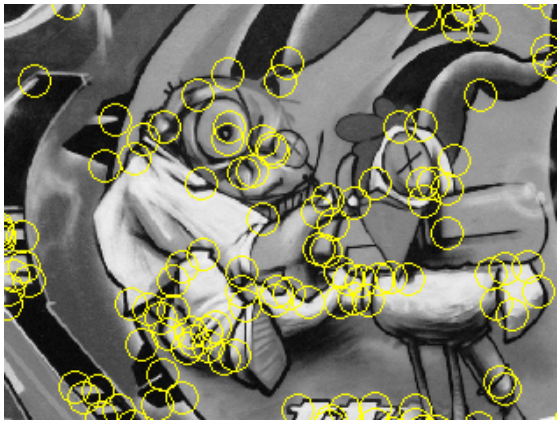
energy in filter banks



Shape

http://www.cs.princeton.edu/gfx/proj/shape/

Local Features: Keypoint Localization



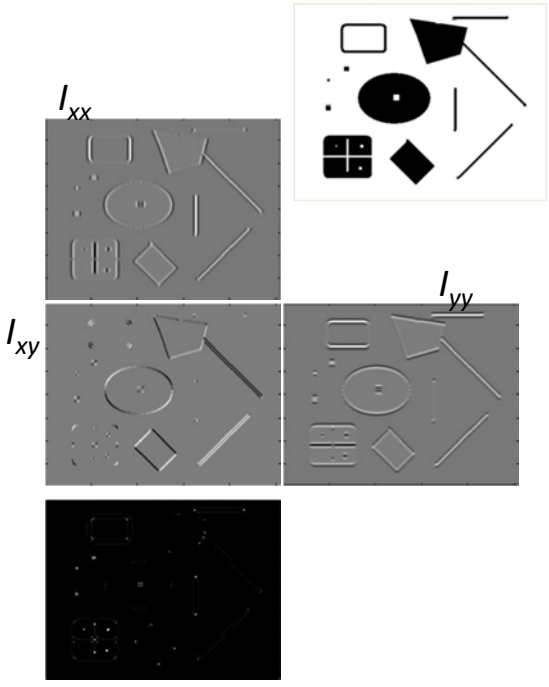
- Keypoint properties:
  - Interesting content
  - Precise localization
  - Repeatable detection under variations of scale, rotation, etc

(Slide of K. Grauman)

Example: Hessian Detector [Beaudet78]

- Hessian determinant

$$Hessian(I) = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix}$$



$$\det(Hessian(I)) = I_{xx}I_{yy} - I_{xy}^2$$

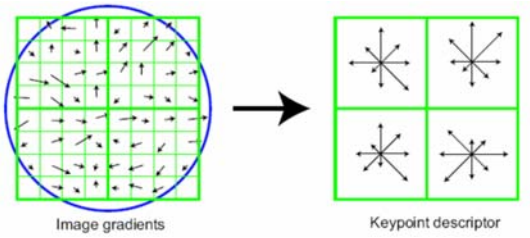
In Matlab:

$$I_{xx}.*I_{yy} - (I_{xy})\text{^}2$$

(Slide of K. Grauman)

Local Appearance Descriptor (SIFT)

Compute gradient
in a local patch



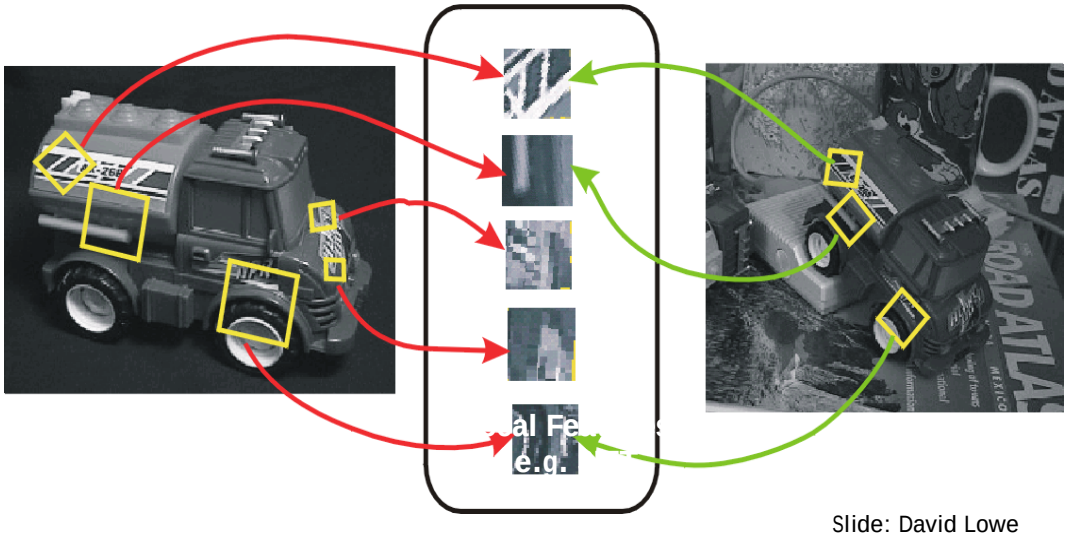
Histogram of oriented gradients over local grids

- e.g., 2x4, or 4x4 grids and 8 directions
  -> 4x4x8=128 dimensions
- Scale invariant

 [Lowe, ICCV 1999]

Image representation

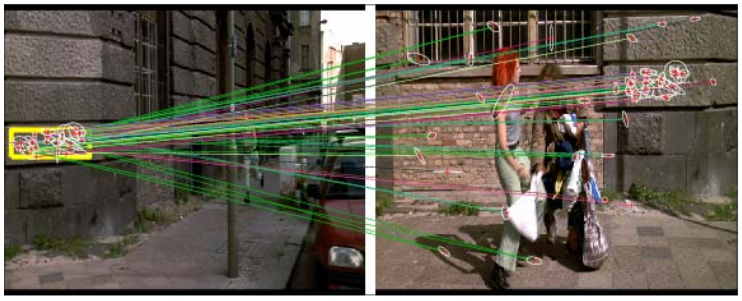
• Image content is transformed into local features that are invariant to geometric and photometric transformations



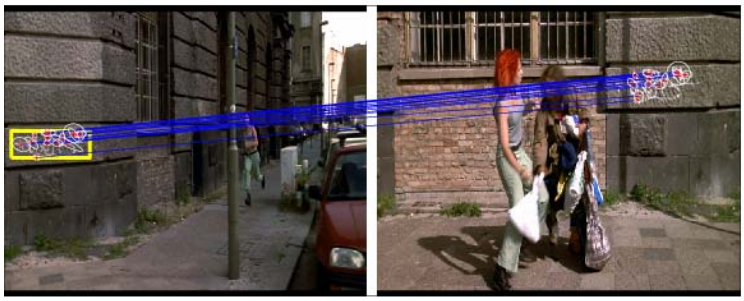
Slide: David Lowe

Example

Initial matches

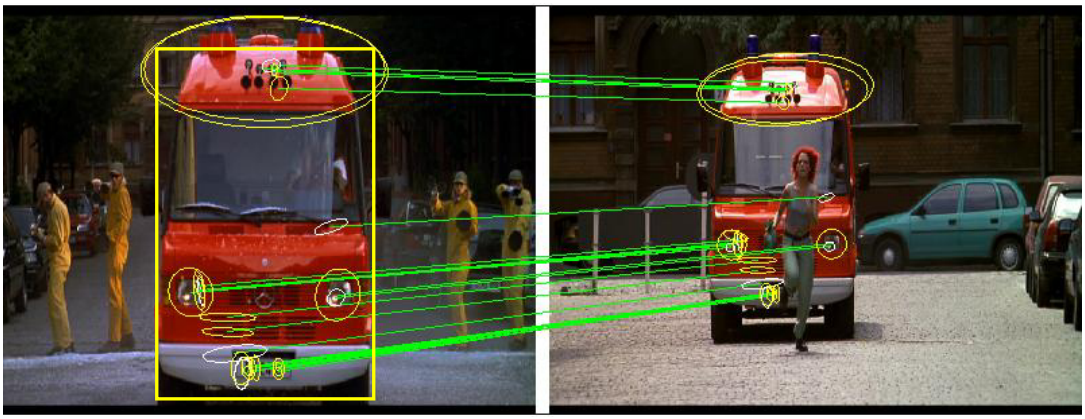


Spatial consistency required



Slide credit: J. Sivic

Match regions between frames using SIFT descriptors and spatial consistency



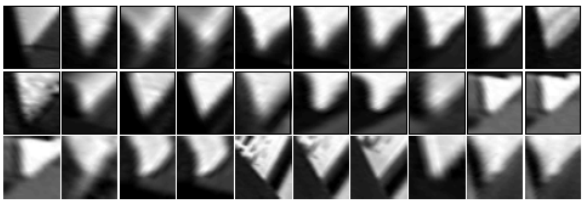
Multiple regions overcome problem of partial occlusion

Shape adapted regions

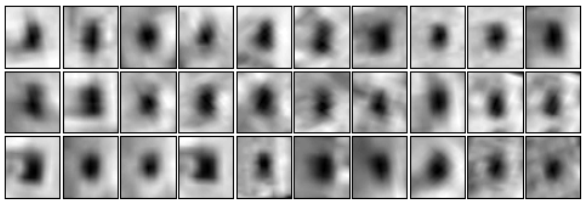
Maximally stable regions

Slide credit: J. Sivic

Clustering of Image Patch Patterns

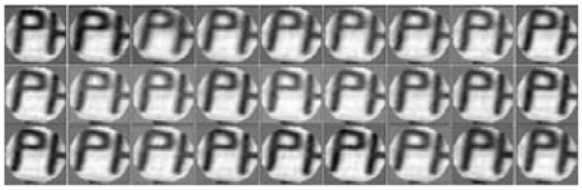


Corners



Blobs

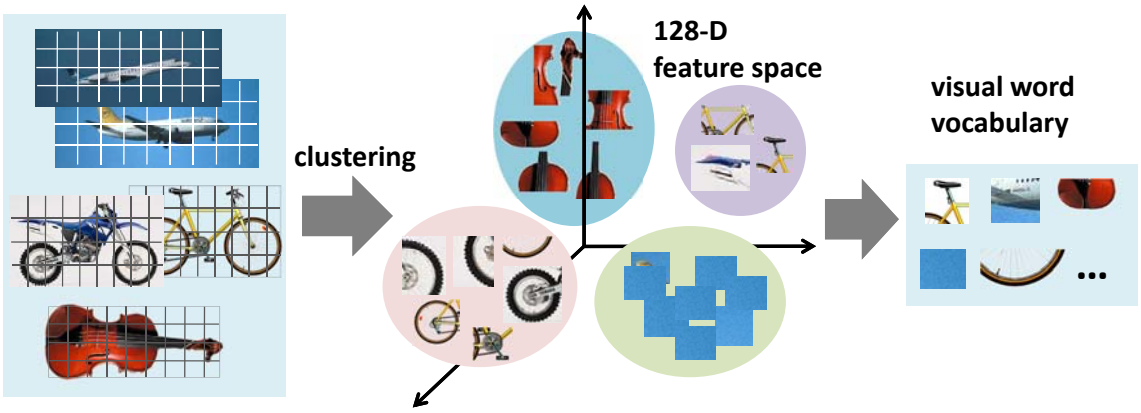
eyes



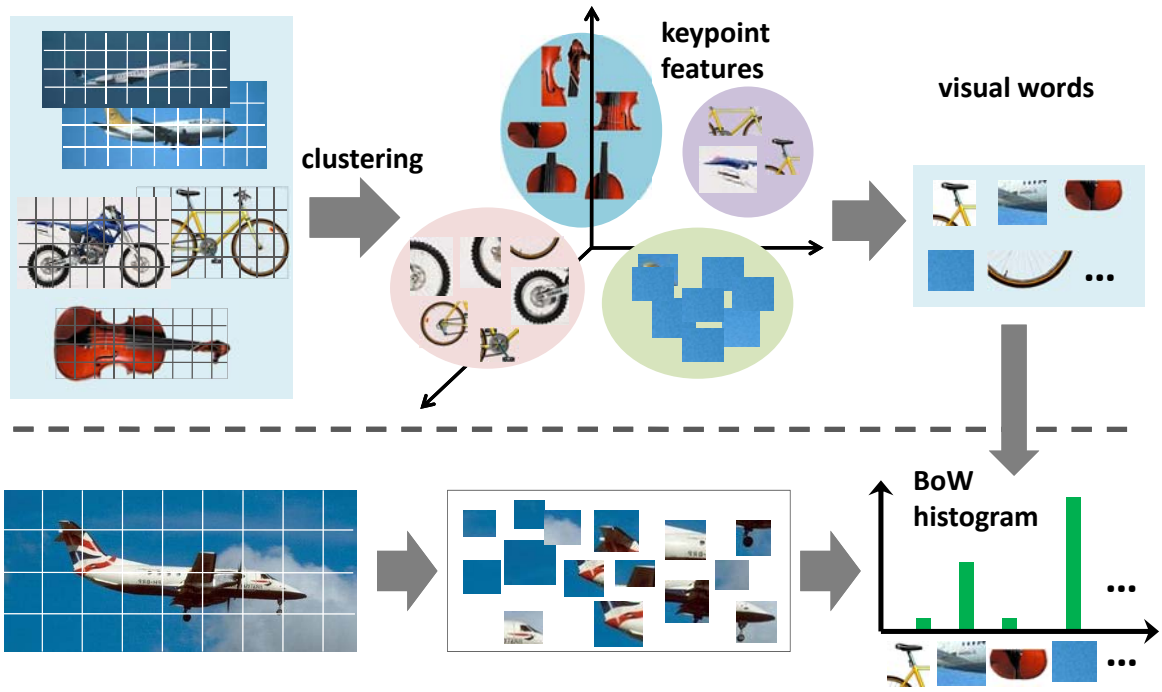
letters

Sivic and Zisserman, "Video Google", 2006

From local features to *Visual Words*



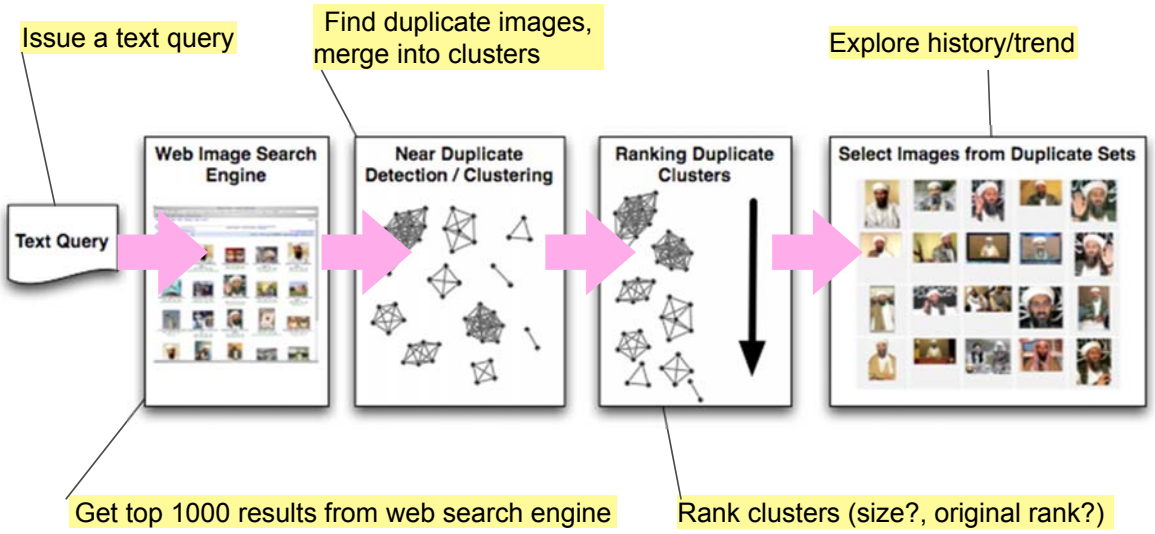
Represent Image as Bag of Words



Content Based Image Search

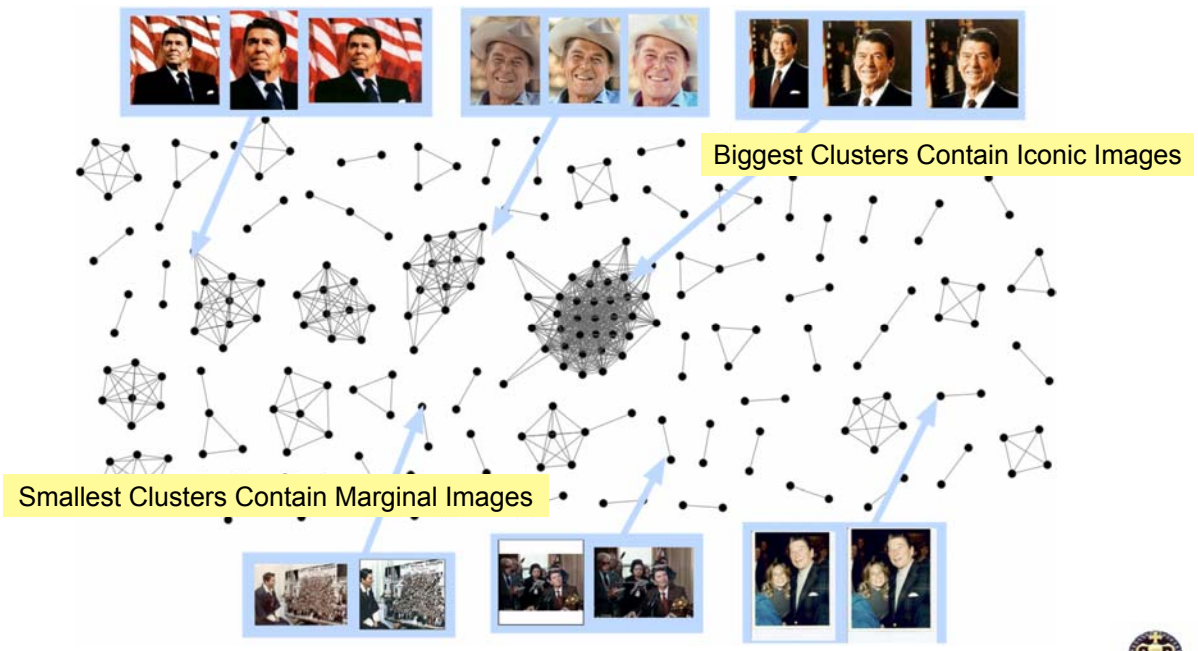
- Demo: Object Retrieval
- Demo 2: Flickr Image Search

Application of Image matching: search result summary



Slide of Lyndon Kennedy

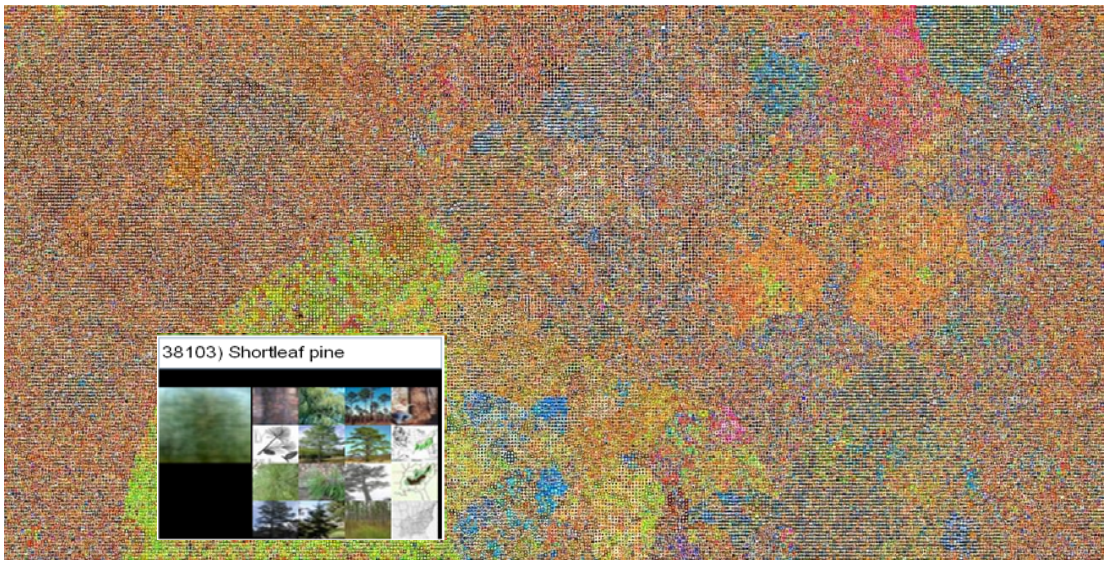
Matching Reveals Image Provenance



Scale Up: Find similar images over Internet

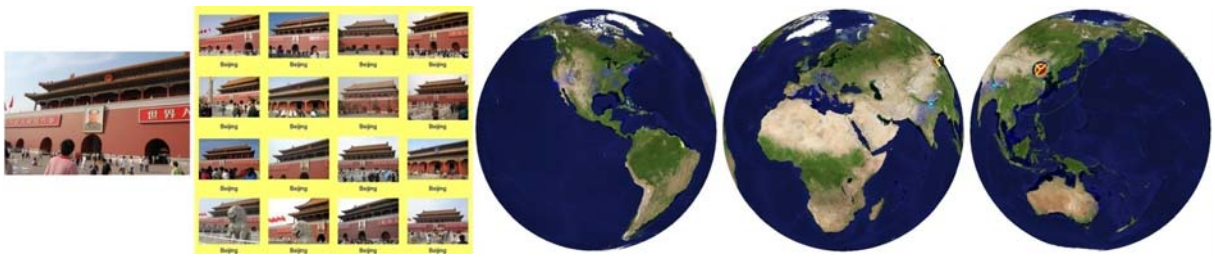
- Billions of images online as dense sampling of the world
- For every image taken, likely to find images that look alike

80 Million Tiny Images, Torralba, Fergus & Freeman, PAMI 2008



IM2GPS: where is this photo taken? (Hays & Efros, 2008)

Similar images Most likely locations



IM2GPS: where is this photo taken? (Hays & Efros, 2008)

Similar images Most likely locations

IM2GPS: where is this photo taken? (Hays & Efros, 2008)

Similar images Most likely locations

»digital video | multimedia lab»

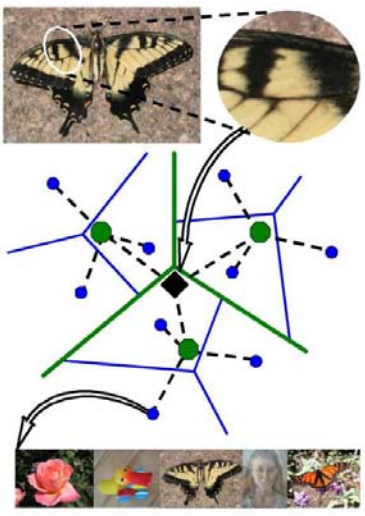
Columbia University in the City of New York

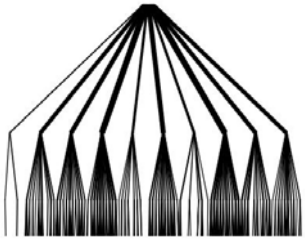
Images on Social Networks

- Understanding social behaviors by media mining
  - Crandall et al, WWW 2009, 35 million Flickr photos, 300,000 users, photographer movement paths

Indexing Gigantic Dataset

- Exhaustive matching of every image is infeasible
- Use hierarchical clustering to speedup
  - Reduce clustering complexity from $O(dk^2)$ to $O(d*\log(k))$
    d: feature dimension, k: clusters
- Each local feature mapped to a path in the tree
- Each image represented as a sub-tree plus occurrence frequency of nodes
- Each node linked with an inverted file of images
- Similarity between query and database images = similarity between two sub-trees

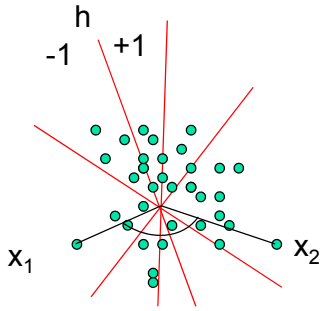




Nister and Stewenius '06

Search over Billions: Scalability is a Big Issue

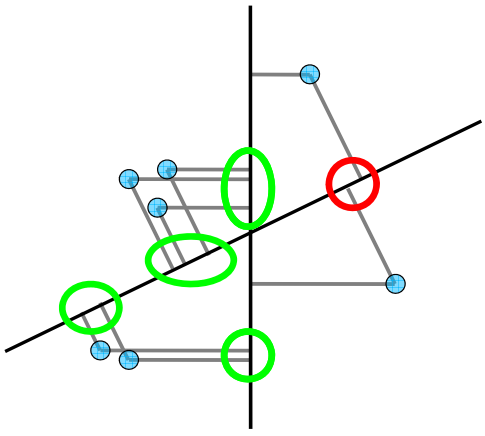
- Similarity Search: traditional tree-based methods (e.g., kd-tree) not suitable in high dimension, because of back tracing
- Need accurate, sublinear solutions (o(N), O(log(N)), O(1))
- Recent trends:
  Hashing based index
  - Random projection:
    Locality Sensitive Hash (LSH)
    [Indyk & Motwani 98, Charikar 02]
  - Principal projection:
    Spectral Hashing [Weiss et al 08]
  - Restricted boltzman machines
    [Hinton et al. 06, Torralba et al. 08]
  - Kernel LSH
    [Kulis et al. 09 & Mu et al. 10]



$P(h(x_1) = h(x_2)) = 1 - \cos^{-1}(x_1 \cdot x_2)/\pi$
$= Sim(x_1, x_2)$
random projection h with *N(0,1)*

Beyond Tree Indexing: Locality Sensitive Hashing (LSH)

- Choose a random projection
- Project points
- Points close in the original space remain close under the projection
- Unfortunately, converse not true

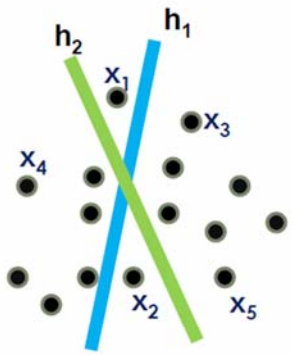


- Answer: use multiple quantized projections which define a high-dimensional "grid"

Slide credit: J. Sivic

Binary Codes

Linear projection (hyperplane) based partitioning



X	x_1	x_2	x_3	x_4	x_5
y_1	0	1	1	0	1
y_2	1	0	1	0	1
…	…	…	…	…	…
y_k	…	…	…	…	…
	010…	100…	111…	001…	110…

Linear Projection based hashing

$$h_k(\mathbf{x}) = \mathrm{sgn}(f(\mathbf{w}_k^\top \mathbf{x} + b_k))$$
$$y_k(\mathbf{x}) = (1 + h_k(\mathbf{x}))/2$$

Very efficient training and testing

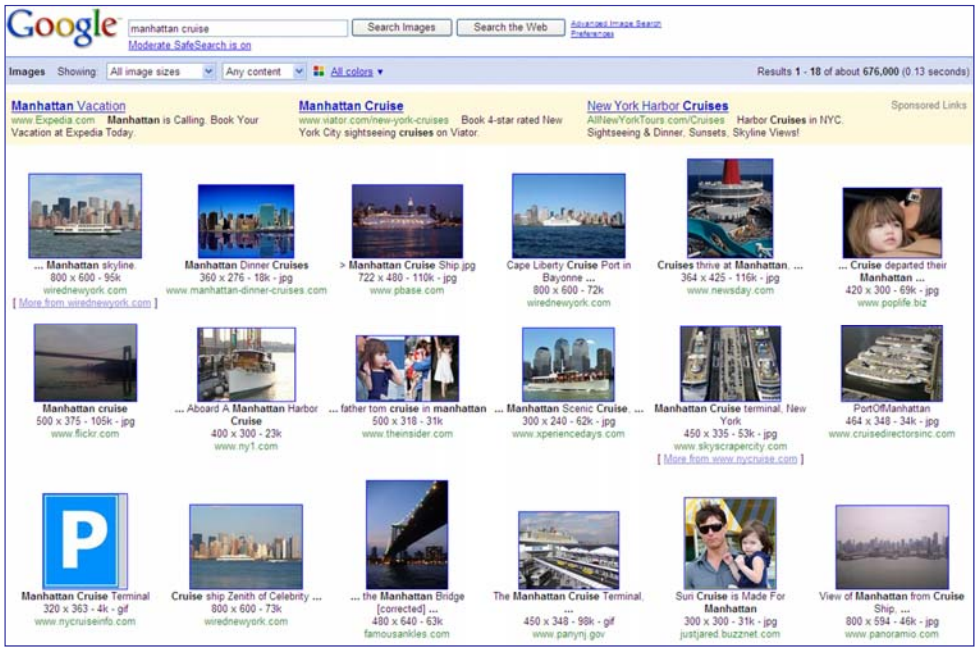
Probabilistic guarantee of finding true targets within ε distance range

[Indyk & Motwani 98]

Slide of Sanjiv Kumar

Going to Higher Level: Text-based Search

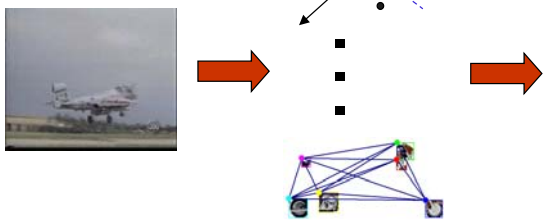
Current system still flawed, e.g., keyword: Manhattan Cruise



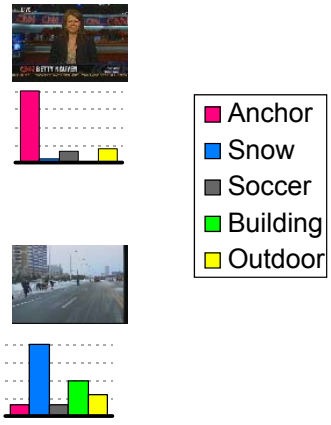
Auto Image Tagging May Help Fill the Gap

- Audio-visual features
- User social features
- Camera/location info

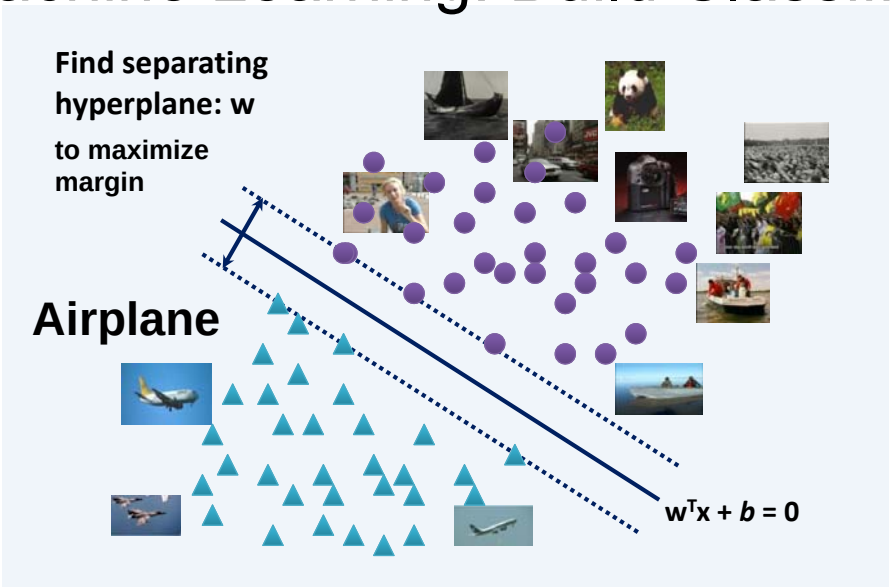
- Rich semantic description based on content recognition



Statistical models



Machine Learning: Build Classifier



Decision function: $f(x) = \text{sign}(w^T x + b)$

$w^T x_i + b > 0$ **if label** $y_i = +1$

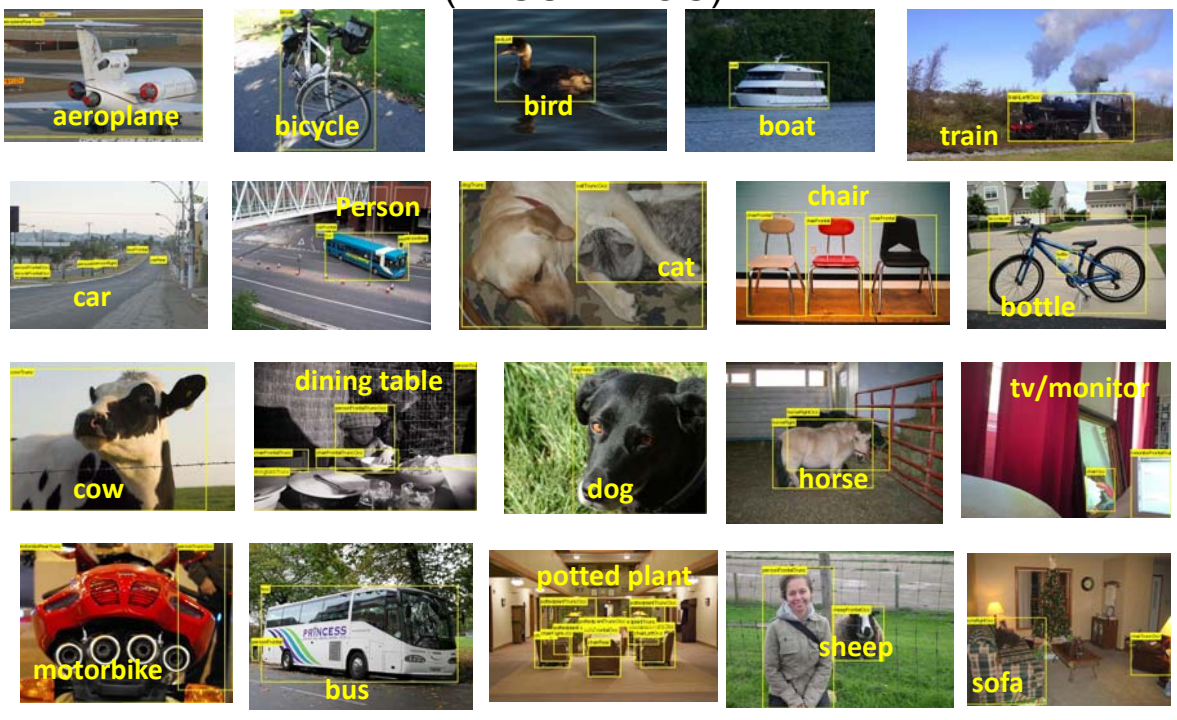
$w^T x_i + b < 0$ **if label** $y_i = -1$

TRECVID: Detection Examples

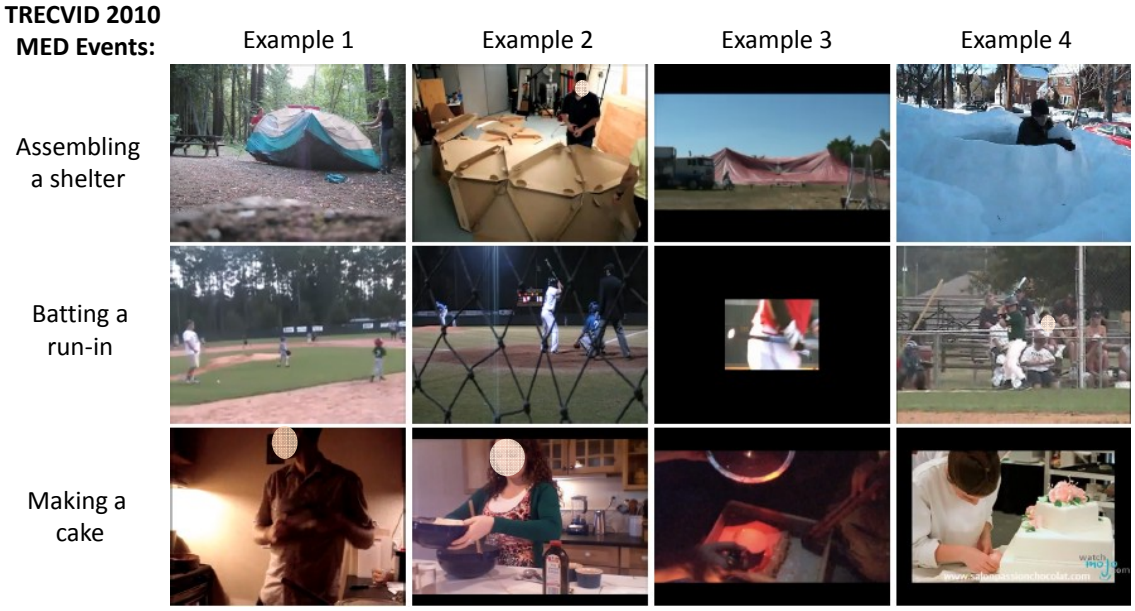
- Top five classification results

Object Localization

(PASCAL VOC)



High-Level Multimedia Event Detection



Need fusion of multimodal analysis: visual, audio, text, temporal

Model Event Context

Understanding contexts is critical for event modeling.

Classifiers Enable Concept-Level Search

- Offline concept detection

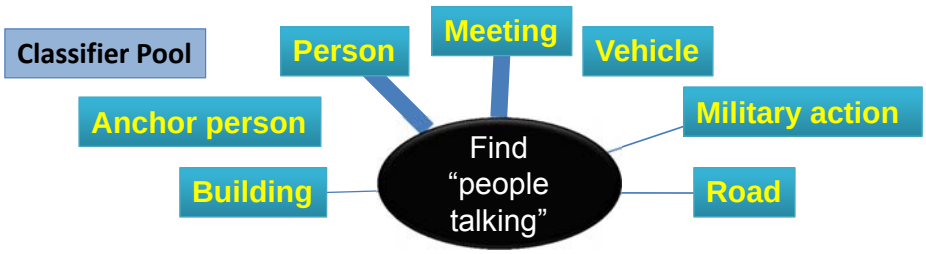
Visual Concept Classifiers

Anchor person

Person, Meeting, …

Military action, Vehicle, Road, Building…

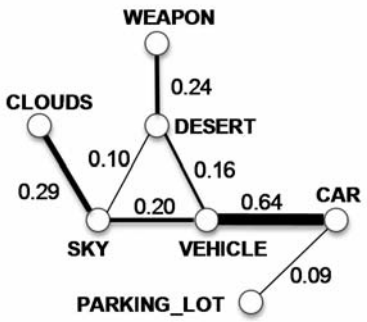
- Online search



Explore Concept Correlation: Semantic Diffusion via Graph

Individual Classifiers:
Desert: 0.68; Sky: 0.60;
Weapon: 0.38; Car: 0.43;
Vehicle: 0.35 …

Concept correlation graph



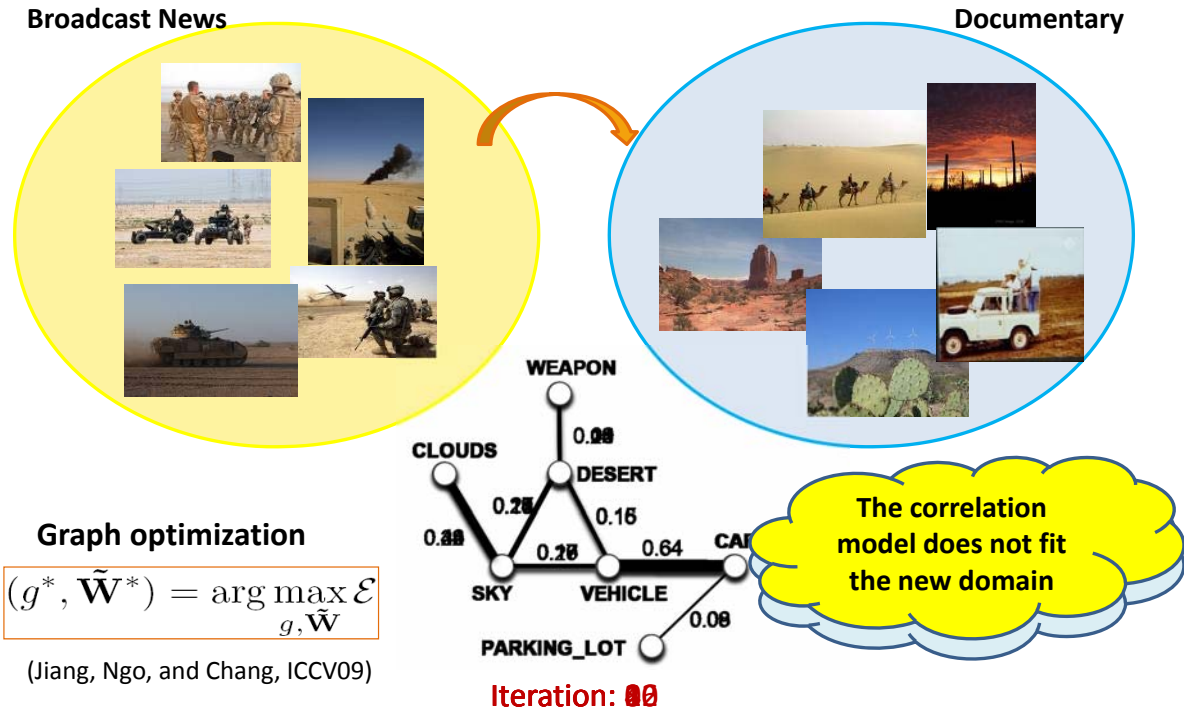
correlation matrix

Classifier c_j score

$$\mathcal{E}(g) = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} W_{ij}\left\|\frac{g(c_i)}{\sqrt{d(c_i)}} - \frac{g(c_j)}{\sqrt{d(c_j)}}\right\|^2$$

$$(g^*, \tilde{\mathbf{W}}^*) = \arg\max_{g,\tilde{\mathbf{W}}} \mathcal{E}$$

Adapting Graph Weights to New Domain



Graph optimization

$$(g^*, \tilde{\mathbf{W}}^*) = \arg\max_{g,\tilde{\mathbf{W}}} \mathcal{E}$$

(Jiang, Ngo, and Chang, ICCV09)

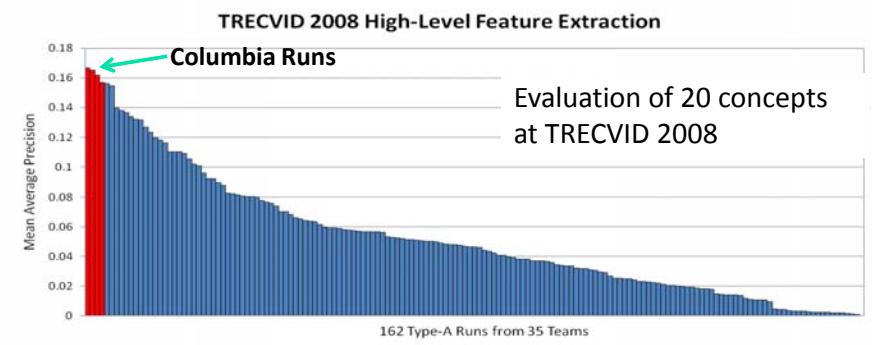
Iteration: 00

Columbia CuZero: 400+ classifiers

Columbia University CuZero digital video | multimedia laboratory

concept detection models: objects, people, location, scenes, events, etc

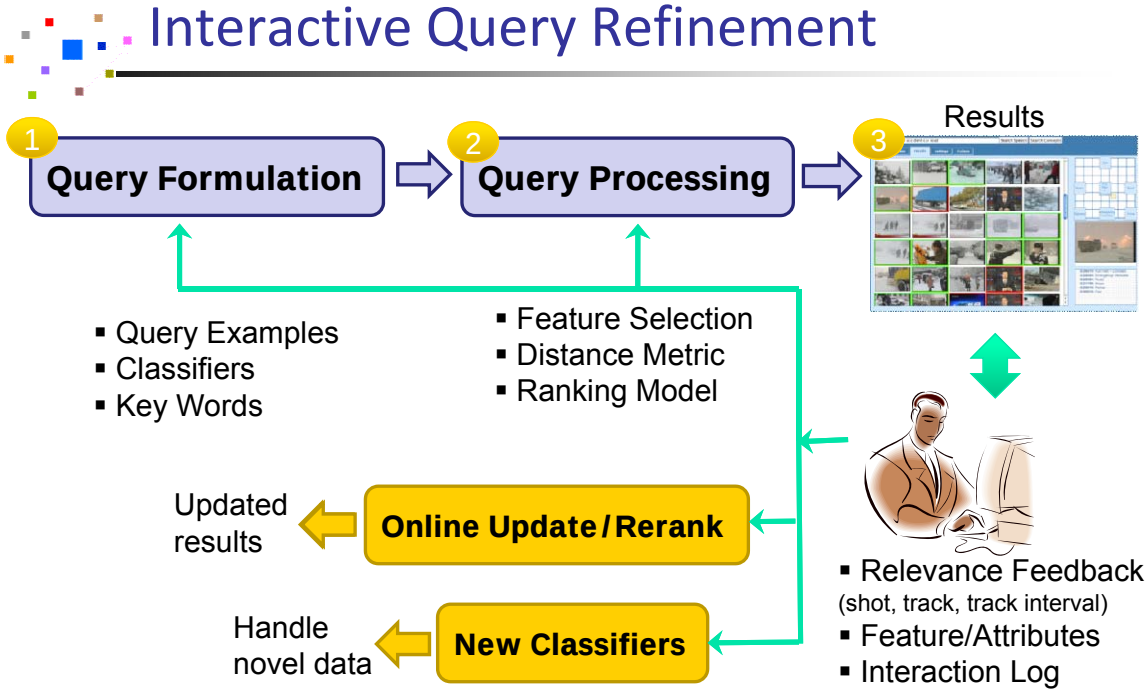
airplane airplane_takeoff airport_or_airfield armed_person building car cityscape crowd desert dirt_gravel_road entertainment explosion_fire forest highway hospital insurgents landscape maps military military_base military_personnel mountain nighttime people-marching person powerplants riot river road rpg shooting smoke tanks urban vegetation vehicle waterscape_waterfront weapons weather



Demos: classifier-based search

- Find lake front buildings in the park
- Find person walking around building
- Find a car on a road in a snowy condition

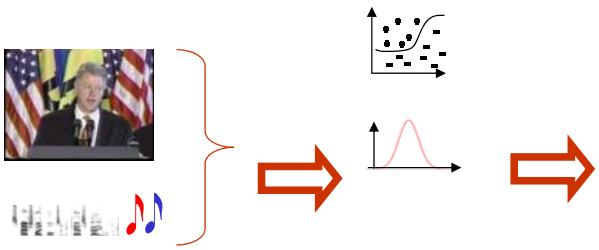
When User in the Loop: Interactive Query Refinement



Columbia TAG Interactive Image Search System

- Demo:
  Rapid Image Annotation with User Interaction

Instead of Automatic Tagging 100% of Concepts



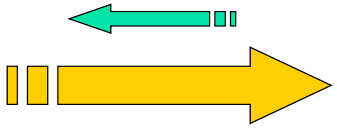
Tags: Person: 79%
Clinton: 75%
US Flag: 80%
Podium: 70%
Give speech: 75%
Press conference: 65%

What if we can ask users to help 1-2 labels?

Partial Active Tagging

(Jiang, Chang, Loui, ICIP 06)

User labels: park, picnic



Auto. generated labels:
people, tree, mountain, etc

Examples of Best Questions

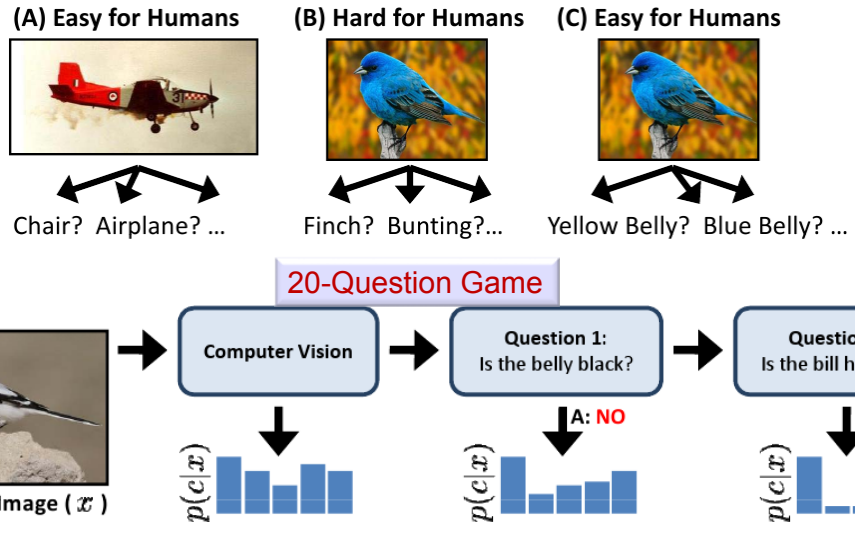
Active Tagging

Best Questions to Ask User?

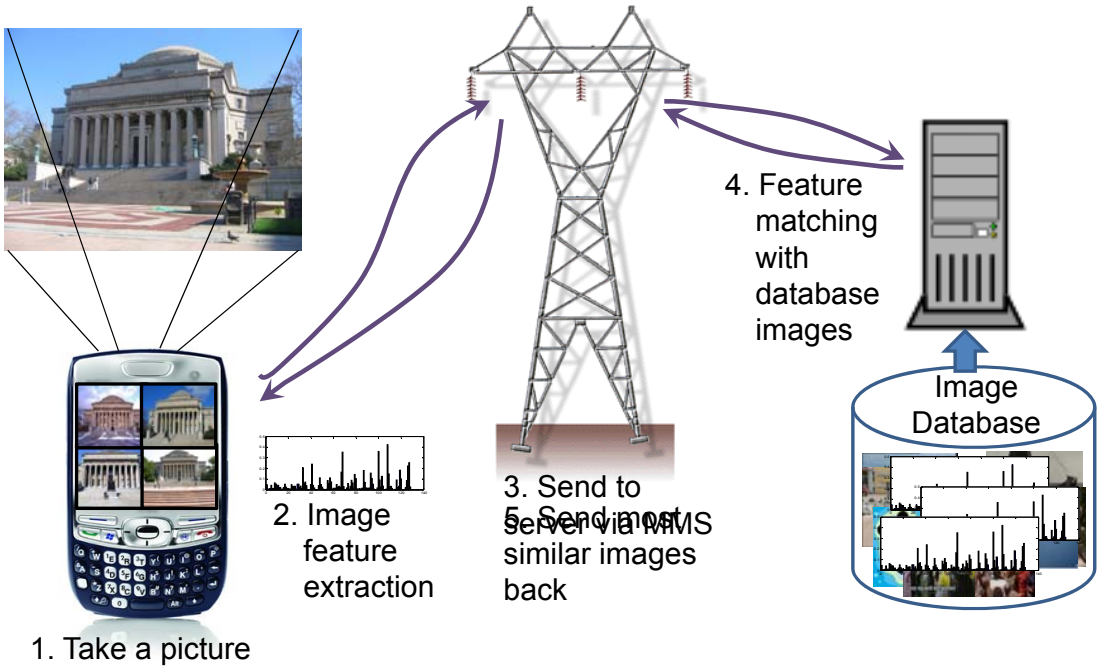
Airplane, animal, boat, building, bus, car, chart, court , crowd, desert, entertainment, explosion_fire, face, flag_us, government_leader, map, meeting, military, mountain, natural_disaster, office, outdoor, people_marching, person, road, sky, snow, sports, urban, waterscape, etc

User in the Loop - Relevance Feedback

- Human-machine collaboration
  - Humans/machines do what they are best at [Branson et al, ECCV, 2010]



Mobile Visual Search

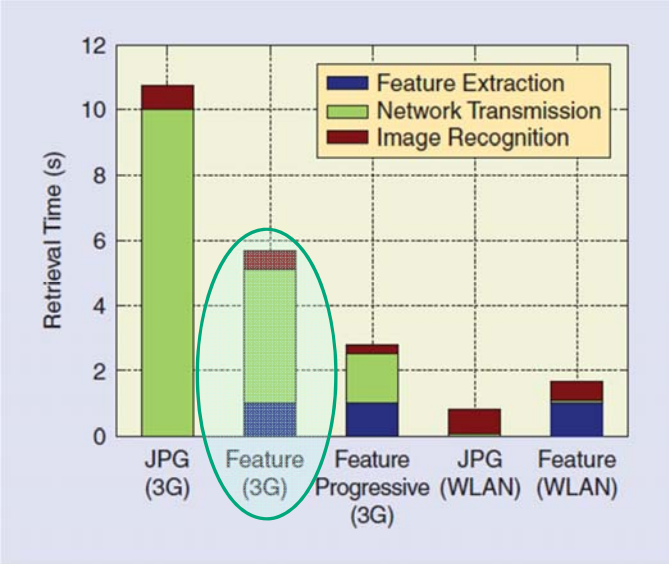


System level issues

- Speed
  - Feature extraction
  - Transmitting features or images (up and down)
  - Searching large databases
- Storage
  - Features and codebooks
- User interface
  - Quality of captured images
  - Visualization of search results

Mobile Challenge: Speed and Bandwidth

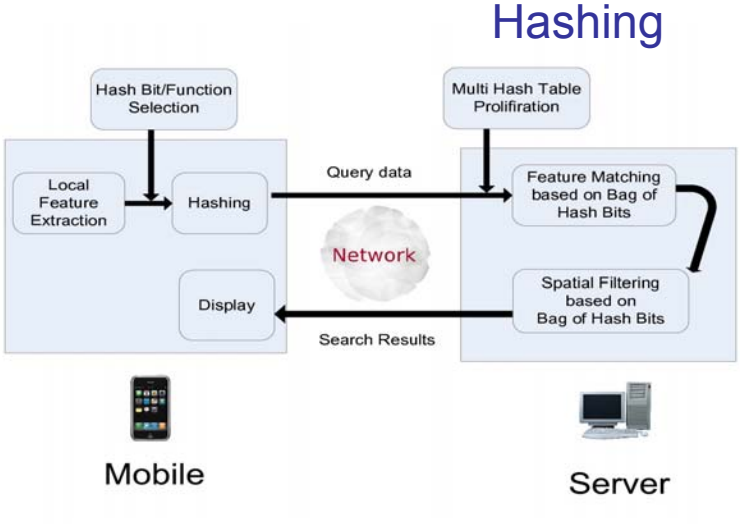
- Speed still limited by bandwidth and power



Mobile Visual Search, Girod, et al, SPM, 2011

Columbia Mobile Product Search System based on Hashing

He, Lin, Feng, and Chang, ACM MM 2011



Server:

- **400,000 product images crawled from Amazon, eBay and Zappos**
- **Hundreds of categories; shoes, clothes, electrical devices, groceries, kitchen supplies, movies, etc.**

Speed

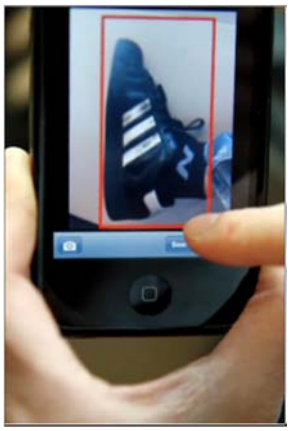
- **Feature extraction: ~1s**
- **Transmission: 80 bits/feature**
- **Serer Search: ~0.4s**
- **Download/display: 1-2s**

video demo

Mobile App Demo

Add Interactive Tools on Mobile Devices

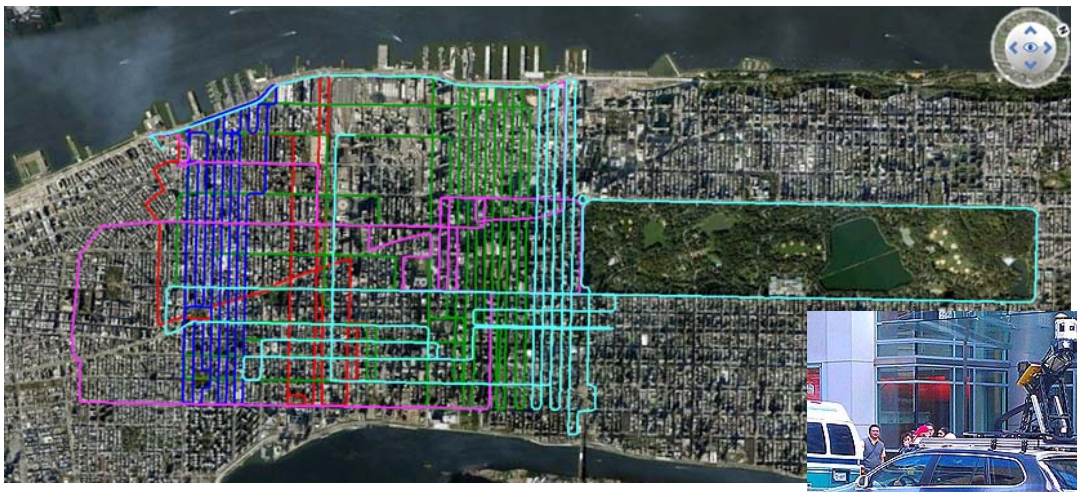
- Interactive Segmentation
  - User helps machine identify point of interest



1:01

Mobile Location Search

- 300,000 images of 50,000 locations in Manhattan
- Collected by the NAVTEQ street view imaging system



Geographical distribution

Challenge

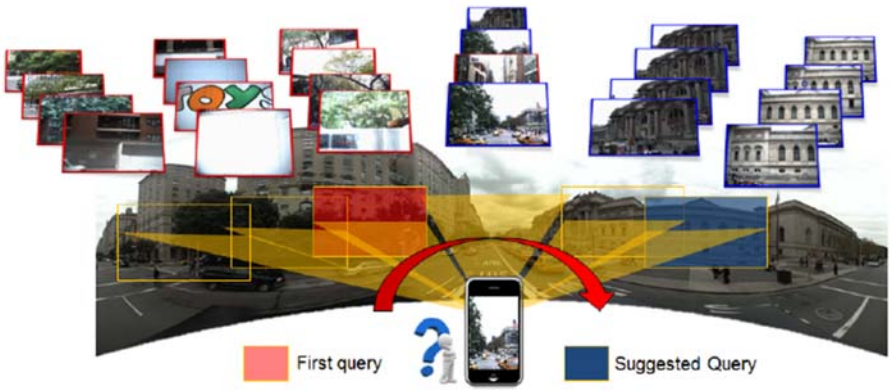
How to guide the user to take a successful mobile query?

- Which view will be the best query?
  - For example, in mobile location search:

  - Or in mobile product search:

Solution: Active Query Sensing

- Guide User to a More Successful Search Angle
  - Active Query Sensing [Yu, Ji, Zhang, and Chang, ACMMM '01]

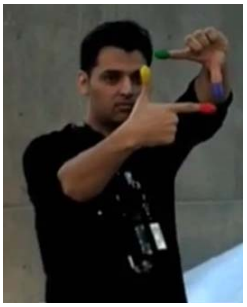


Video demo

Mobile App Demo

Mobile Augmented Reality

- MIT Sixth Sense Project (Pranav Mistry and Pattie Maes, MIT)
  - Mobile wearable computer
  - Camera and projector
  - Gesture interaction
  - Visual recognition



EE 6882, Spring 2011

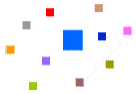
- Course web site:
  - http://www.ee.columbia.edu/~sfchang/course/vse
- Instructor: Prof. Shih-Fu Chang
  - Office hour: Monday 11-12, CEPSR 709
- Asst. Instructor: Dr. Rong-Rong Ji
  - Office Hour: Friday 2-4pm, CEPSR 707
  - Staff Assistants: Tongtao Zhang, and Jinyuan Feng
- Prerequisites:
  - Image processing or computer vision, pattern recognition, probability (a 15 mins quiz)

Course Format

- Required background: familiarity with image processing, pattern recognition. There will be a quiz.
- Lectures + two hands-on homeworks (due 2/13, 2/27)
- Mid-term project
  - Review and experiment topics of interest, 2 students each team
  - Proposal due 3/5, narrated slides due 3/26
  - Selected projects presented and discussed in class (3/26-4/9)
- Final project
  - Extension of mid-term projects encouraged, 2 students each team
  - Proposal due 4/2, narrated slides due 4/30
  - Selected projects presented and discussed in class (4/30-5/7)
- Grading:
  - class participation (20%), homework (20%), mid-term (20%), final (40%)
  - Everyone has a total "budget" of 4 days for late submissions. No other delayed submission accepted.

Examples of Final Projects

- Mobile visual search: feature extraction, quality enhancement, real-time systems
- Mobile augmented reality
- Image search for specific domains: product, patent trademark, roadside objects, landmarks, 3D objects
- Hashing for search over million scale datasets
- Gesture recognition with depth sensors
- Fast video copy detection
- Search by sketch drawings
- Multimedia summarization



Reading List

Many papers available at http://www.ee.columbia.edu/ln/dvmm/newPublication.htm/

- Rui, Y., T.S. Huang, and S.-F. Chang, *Image retrieval: current techniques, promising directions and open issues. Journal of Visual Communication and Image Representation, 1999. 10(4): p. 39-62.*
- Smeulders, A.W.M., et al., *Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell., 2000. 22(12): p. 1349-1380.*
- Sivic, J. and A. Zisserman, *Video Google: A text retrieval approach to object matching in videos, in ICCV. 2003.*
- Mikolajczyk, K. and C. Schmid, *A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005: p. 1615-1630.*
- Nister, D. and H. Stewenius. *Scalable recognition with a vocabulary tree. in CVPR. 2006.*
- Jiang, Y.-G., et al. Consumer Video Understanding: A Benchmark Database and An Evaluation of Human and Machine Performance. ACM International Conference on Multimedia Retrieval (ICMR), 2011.
- Zavesky, E. and S. Chang. *CuZero: embracing the frontier of interactive visual search for informed users. in ACM Multimedia Information Retrieval (MIR). 2008.*
- Kennedy, L. and M. Naaman. *Generating diverse and representative image search results for landmarks. in ACM WWW. 2008.*
- Yu, F., R. Ji, S.-F. Chang. Active Query Sensing for mobile location search. In Proceeding of ACM International Conference on Multimedia (ACM MM), 2011.