Visual Search

Shih-Fu Chang
Department of Electrical Engineering
Columbia University

Boston University ECE Lecture, September 2010
Many pictures are taken everyday/everywhere...

Taking pictures everyday/everywhere...

http://cache1.asset-cache.net/xc/200545511-001.jpg?v=1&c=NewsMaker&k=2&d=CED48661B87C5DBFF5EA7A9E4218F9EE1F6F6178A68B340C
http://cache4.asset-cache.net/xc/200483959-001.jpg?v=1&c=NewsMaker&k=2&d=E76F2F2F4F969CEBD9A551141E9DF12C0E47AD5C832857DF0450484851C07200123AA3B5A18ED0
Taking pictures everyday/everywhere...
But we are not good at organizing them ...

- Most photos and videos remain in shoebox or e-shoebox
- But we love to share them ...
Example:

- **My family’s video channel**
- Sharing ~ 100 videos on Youtube
- This single video has been viewed >130,000 times
- More popular than all of my published papers!!!
Challenge after capture and upload

It will be nice to tag ...

Performance, classical music, ensemble, concert; Pachelbel Canon in D Major, romantic; young musicians, girls, boys; outdoor, Bryant park in NYC, statue, columns, stage, flowers; Instruments, audio equipments, violins; . . .

- Tagging is boring and hard
- Each uploaded photo has only 0.97 tag on average

(Naaman, Yahoo!/Rutgers)
Scarce Tags \(\rightarrow\) Faulty Search Engines

“Manhattan Cruise”
Scarce Tags $\rightarrow$ Faulty Search Engines

"Cruise ship in Manhattan"
The other side of coin

Visual Search: use pixels to find information

Recent commercial applications:

- product search
- landmark search
- document search

iPhone App SnapTel

Google Goggles
How Does It Work?

1. Take a picture

2. Image feature extraction

3. Send to server via MMS

4. Feature matching with database images

5. Send most similar images back
Basic Image Features

Grid Color Moment
- Mean
- Variance
- Energy in different filter banks

Wavelet Texture
- Edge direction counts

Canny Edge
- Edge direction counts
Recent Popular Features: Sample keypoints in images

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

- Computation in Gaussian scale pyramid
- Local maxima in scale space

\[ \sigma = 2^\frac{1}{4} \]

Original image

Sampling with step \( \sigma^4 = 2 \)

(Slide of K. Grauman)
Extract Appearance Descriptor from Keypoints

Compute gradient in a local patch

SIFT: Histogram of oriented gradients over local grids
- e.g., 4x4 grids and 8 directions
  \[ \rightarrow 4 \times 4 \times 8 = 128 \text{ dimensions} \]
- Scale invariant

[Lowe, ICCV 1999]
Clustering of Image Patch Patterns

Corners

Blobs

eyes

letters

Sivic and Zisserman, “Video Google”, 2006
A New Representation: Visual Words

clustering

keypoint features

visual words

BoW histogram
Image matching with local interest points

- Measure useful information
  - Image similarity
  - Copy/source identification
  - Discover possible transformation between images
- Can be done efficiently

Matching SIFT points [Lowe, 1999]
Content Based Image Search

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

S.-F. Chang, Columbia U.
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)
Search over Billions: Scalability is a Big Issue

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

Linear projection (hyperplane) based partitioning

<table>
<thead>
<tr>
<th>X</th>
<th>x_1</th>
<th>x_2</th>
<th>x_3</th>
<th>x_4</th>
<th>x_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y_1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>y_2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>y_k</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Linear Projection based hashing

\[ h_k(x) = \text{sgn}(f(w_k^T x + b_k)) \]

\[ y_k(x) = (1 + h_k(x))/2 \]

Very efficient training and testing

Probabilistic guarantee of finding true targets within \( \varepsilon \) distance range

[Indyk & Motwani 98]
Beyond Randomness: Semi-Supervised Learned Projection

- Given Pair-Wise Relations
  \((x_i, x_j) \in \mathcal{M}: \text{neighbor pair}\)
  \((x_i, x_j) \in \mathcal{C}: \text{nonneighbor-pair}\)

- Measure empirical fit of hash bits

- Are the partitions balanced?
  Measure hash bit variance

- Elegant eigen-decomposition solution

\[
J(W) = \frac{1}{2} \text{tr} [W^T X_l S X_l^T W] + \frac{\eta}{2} \sum_k E[||w_k^T x||^2]
\]

\[
S_{ij} = \begin{cases} 
1 & : (x_i, x_j) \in \mathcal{M} \\
-1 & : (x_i, x_j) \in \mathcal{C} \\
0 & : \text{otherwise.}
\end{cases}
\]

- Incremental learning via AdaBoosting

References:
[Wang, Kumar, Chang CVPR10, ICML10]
• Learned projection hash increases accuracy > 2X
• Query time: a few seconds
• Compact code - 48 bits vs. 128 bytes per sample

• Challenge: scale to billions?

Tiny Image – 80M

SIFT-1M

Binary Reconstructive Embedding

Learned Projection Hash
Beyond pair-wise similarity: Image Manifold

- Recover manifold distributions in high-dimensional space
- Each point represents an image, construct sparse graph
- Graph for information propagation, like PageRank
Manifold vs. Supervised Learning

- Poor classifier when ignoring manifold distribution
Graph-based Semi-Supervised Learning

• Construct Sparse for Label Propagation

\[ f^* = \min_f Q(f, y, G(V, W)) \]

- \( G \) -- graph
- \( V \) -- graph node
- \( W \) -- weight matrix
- \( y \) -- label matrix
- \( f \) -- classification
- \( Q \) -- risk function
An active area in Machine Learning

- Given initial labels, $Y$, find classification $F$ over graph

$$Q(F) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \left[ \frac{F_i.}{\sqrt{D_{ii}}} - \frac{F_j.}{\sqrt{D_{jj}}} \right]^2 + \mu \sum_{i=1}^{l} \|F_i. - Y_i.\|^2$$

$$= \text{tr} \{ F^\top L F + \mu (F - Y)^\top (F - Y) \}$$

(Zhou, et al NIPS04)

- Gaussian fields & Harmonic functions (Zhu et al ICML03)

$$Q(F) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \|F_i. - F_j.\|^2$$

1) $\Delta F = 0$ on unlabeled data, where $\Delta = D - W$ is the graph Laplacian;
2) $F_i. = Y_i.$ on labeled data.
Non-Trivial Issues

LGC Method

GFHF Method

Unbalanced Labels

Bad Label Locations

Noisy Labels

S.-F. Chang, Columbia U.
A Key Problem: How to Construct a Good Sparse Graph from Massive Data?

- **B-Match preserves sparse structures**
- **Efficient method exists much faster than** $O(N^3)$

\[
\min_{P \in \mathbb{B}} \sum_{i,j} P_{ij} D_{ij} \\
\text{s.t. } \sum_{j} P_{ij} = b, P_{ii} = 0, P_{ij} = P_{ji},
\]

(Jebara, Wang, Chang ICML 2009)
Another Key Issue: Noisy Labels

- Treat both labels & prediction as optimization variables

\[ Q(F, Y) = \frac{1}{2} \text{tr} \left\{ F^T L F + \mu (F - V Y)^T (F - V Y) \right\} \]

- Diagnose and add/remove labels
  - Select the most reliable and informative label

\[ Q(Y) = \frac{1}{2} \text{tr} \left( Y^T V^T \left[ P^T L P + \mu (P^T - I)(P - I) \right] V Y \right) \]

- Propagate trusted labels to the whole graph
  - given label (Y), propagate over graph, predict F

\[ \frac{\partial Q}{\partial F^*} = 0 \Rightarrow F^* = (L/\mu + I)^{-1} V Y = P V Y \]

(Wang, Jebara, Chang, ICML08) (Wang and Chang, CVPR09)
Iteratively Tune and Propagate Label Information

(Wang and Chang CVPR’09)

Decline of the cost function $Q$ over iterations (with vs. without label tuning)
Application: Internet Search Result Reranking

Google Search “Tiger”

Keyword Search

Search Engine

Treat top results as +
Bottom imgs as -

Label Diagnosis Diffusion
Application: Internet Search Result Reranking

Keyword Search

Search Engine

Treat top results as +

Bottom imgs as -

Label Diagnosis Diffusion

Reranked Results
Application: Internet Search Result Reranking

Keyword Search

Search Engine

Treat top results as +
Bottom imgs as -

Label Diagnosis Diffusion

Google Search “Statue of Liberty”
Application: Internet Search Result Reranking

Reranked Results

Keyword Search

Search Engine

Treat top results as +
Bottom imgs as -

Label Diagnosis Diffusion
Interactive Image Annotation: Columbia TAG System

Image/Video data

Processing (denoising, cropping …)

Feature Extraction

Compute Similarity

Graph Construction

Interactive browse / label

Graphic User Interface

Graph Diffusion

Interactive Mode

Applications

Concept detection, Object retrieval, Image search

S.-F. Chang, Columbia U.
Columbia TAG Image Search System

- **Demo:**
  Rapid Image Annotation with User Interaction
Beyond Matching: Automatic Visual Recognition

- Audio-visual features
- Geo, time, camera metadata
- User context

- Rich semantic labels

Analysis models

[Image of analysis models with points and graphs]

Legend:
- Anchor
- Snow
- Soccer
- Building
- Outdoor
Hot topic … community fast growing!

(As of Nov. 2009)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data domain</th>
<th>amount</th>
<th>types</th>
<th>Lexicon size</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSCOM</td>
<td>Broadcast news video</td>
<td>170 hours Television News 61901 subshots</td>
<td>video subshots</td>
<td>1000+ concepts</td>
</tr>
<tr>
<td>CalTech256</td>
<td>Internet Images</td>
<td>30,607 images</td>
<td>images</td>
<td>256 classes</td>
</tr>
<tr>
<td>PASCAL</td>
<td>Internet Images</td>
<td>9,963 images 24,640 annotated objects</td>
<td>images, objects</td>
<td>20 classes</td>
</tr>
<tr>
<td>Tiny Image</td>
<td>Internet Images</td>
<td>80,000,000 tiny images (32x32)</td>
<td>images</td>
<td>75,378 WordNet nouns</td>
</tr>
<tr>
<td>LabelMe</td>
<td>Internet and user uploaded images and videos</td>
<td>30,369 images from 183 folders</td>
<td>images, keyframes</td>
<td>111,490 object labels</td>
</tr>
<tr>
<td>ImageNet</td>
<td>Internet images</td>
<td>9,386,073 images</td>
<td>images</td>
<td>14,847 WordNet synsets</td>
</tr>
<tr>
<td>Lotus Hill Dataset</td>
<td>Internet Images</td>
<td>500,000+ images and keyframes</td>
<td>images, keyframes</td>
<td>280 object classes</td>
</tr>
</tbody>
</table>
Image/Video Classification

TRECVID

Boat
Street
Protest

CalTech 101

Camera
Airplane
Panda
Object Localization
(PASCAL VOC)
Create Large Corpora: Crowd Sourcing

- Amazon Mechanic Turk
- Web Open Market for Human Computing

What can I annotate for you?

Task

Is this a DOG?
- Yes
- No

earn $0.03 per image tag

Internet-scale workforce
Crowd sourcing for image annotation

Such task is nicely called Human Intelligence Task (HIT)!
Rapid Advances in Image Annotation

- “Moore’s Law”: accuracy doubles in about 2 years
CuZero: tag videos with 400+ classifiers

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

TRECVID 2008 High-Level Feature Extraction

Columbia Runs
Evaluation of 20 concepts at TRECVID 2008
TRECVID: Detection Examples

- Top five classification results

Classroom

Demonstration
Or Protest

Cityscape

Airplane flying

Singing
What can a small recognition engine do?

• How to leverage a noisy and small visual vocabulary?
• State of the Art video search engines have ~1000 visual concepts
  – IBM IMAR
  – U. Amsterdam MediaMill
  – CMU Informedia
  – Columbia CuZero
Obvious Problems with Small Vocabulary

• Given a search topic, users often have difficulty in choosing matched concept classifiers

Find shots of something **burning** with **flames** visible

Cannot find matched classifiers! Which classifiers work?
Ideas from Information Retrieval: Query Expansion

Classifier Suggestion

query time concept mining
Use visual examples to compensate the deficiency

Query: “find person running around a building”
Semantic Search Target (query topic)

1. Instant Query Concept Mapping
   - Concept Mapping
   - Concept Visualization

2. Real-Time Navigation & Feedback
   - Multi-Concept Navigation
   - Real-time Query Manipulation

Video Content (with metadata)

content analytics

thousands of semantic concepts
- car
- urban
- fire
- building
- airplane
- road
- sky
- outdoor
- person

CuZero Search Engine

user

Concept Pool

(Zavesky and Chang, Multimedia Info Retrieval MIR '08)
Demos

- Find lake front buildings in the park
- Find person walking around building
- Find a car on a road in a snowy condition
Visual Search via Brain State Decoding

• Human Vision is Superb by quick “gist” in the “Blink of an Eye”

Joint work with Paul Sajda’s group
Brain Machine Interface for Image Retrieval

Use EEG brain signals to detect target of interest
(video)

Use image graph to tune & propagate information

Rapid Serial Presentation of Caltech 101 Images

Graph-Based Visual Pattern Discovery

C3Vision System by Sajda et al

TAG System by Wang and Chang

(Wang, Pohlmeyer, Hanna, Jiang, Sajda, Chang, ACMMM09)
The Neural Signatures of “Visual Attention”

From D. Linden, 2005
Single-trial EEG Analysis

- Typically EEG is averaged over trials to increase the amplitude of the signal correlated with cortical processes relative to artifacts (very low SNR)

- High-density EEG systems were designed without a principled approach to handling the volume of information provided by simultaneously sampling from large electrode arrays.

- Our solution: identifying neural correlates with individual stimuli via single trial EEG analysis.

- We apply principled methods to find optimal ways for combining information over electrodes and moments in time contained in individual trials
Identifying Discriminative Components in the EEG Using Single-Trial Analysis

LDA or Logistic Regression is used to learn the contributions of EEG signal components at different spatial-temporal locations (Parra, Sajda et al. 2002, 2003)

Optimal spatial filtering across electrodes within each short window (e.g., 100ms)

Optimal temporal filtering over time windows after onset
The Visual Interest Readout Experiment

User thinks about what he/she wants to search

Database (any target that may interest users)
The Paradigm

Database

Neural (EEG) decoder

Interest-scores
The Paradigm

- Database
  - Neural (EEG) decoder
  - Exemplar labels (noisy)
  - Semi-supervised
    - Graph-based propagation
  - Features from the entire DB
  - prediction score
The Paradigm

Pre-trieage

Post-trieage
The Paradigm

Human inspects only a small sample set via BCI

Machine filters out noise and retrieves targets from very large DB

- General: no predefined semantic classes
- High Throughput: neuro-vision as bootstrap of fast computer vision

Pre-triage

Post-triage
Experiments

- CalTech101: 3798 images from 62 categories
  Satellite images
- Generic neural decoder trained per user using images (Soccer Ball or Baseball Gloves) from Caltech256
- A subset images randomly sampled to construct 6-Hz RSVP sequence
- Initial Trials: 4 subjects, 3 targets (Dalmatian, Chandelier/Menorah, & Starfish)
Example results

Top 20 results of Neural EEG detection

Top 20 results of Hybrid System (BCI-VPM)
DARPA NIA Program: Remote Sensing Target Search

Initial EEG neural signal detection:

After graph refinement and diffusion

Images from DigiGlobe)
Summary: Cross Fertilization of Several Fields

Human Vision

Internet/Data Vision

Image Recognition/Machine Learning

Image video search engines
Conclusions

- Great opportunity for video search research

Exciting topics

- **Semantic Search:**
  Large-scale visual ontology and intuitive search

- **Machine Learning and Computer Vision:**
  Robust classification and image understanding

- **Matching of Billions of Images or More**
  Robust features and fast matching

- **Internet Vision:**
  Explore new applications on Internet

- **Neuro-Computer Vision:**
  Synergistic integration with neural vision systems
Acknowledgments

- Columbia University
  - Eric Zavesky, Yu-Gang Jiang, Jun Wang, Junfeng He, Wei Liu, Wei Jiang, Akira Yanagawa
- Yahoo! Research
  - Lyndon Kennedy
- City University, Hong Kong
  - Chong-Wah Ngo
- IRIT, France
  - Elie El Khoury
References

(many papers can be found at http://www.ee.columbia.edu/dvmm )


- Wei Liu, Junfeng He, Shih-Fu Chang, “Large Graph Construction for Scalable Semi-Supervised Learning,” ICML 2010.
