What do we want from Computational Scene Understanding?

Alexei “Alyosha” Efros
Carnegie Mellon University
Many ways to Understand a Scene…

- **Scene Categories** (e.g. street scene, city, Boston, etc.)
- **Innumerate objects** (people, cars, lampposts, etc)
- **Label/Segment scene elements** (road, buildings, sky, etc)
- **Scene Geometry** (qualitative or quantitative) and **Illumination**
- **Physical Affordances** (where can I walk?)
- **Prediction** (What will happen next?)
Scene Categorization

- Animal vs. no animal (Thorpe, Poggio & Oliva)
- Basic Scene Categories (Oliva, Renenger, Fei-Fei, etc.)
- Image classification, e.g. “Boat” (Caltech, etc.)
- Spatial envelope (Oliva & Torralba)

Spatial categories:
- Beach
- Mountain
- Forest
- City
- Street
- Farm

Spatial envelope categories:
- Scene-centered: Large space, Man-made scene, Semiclosed space
- Scene-centered: Small space, Man-made scene Enclosed

im2gps
**Poster Spotlight:** SUN Attributes: A Large-Scale Database of Scene Attributes

Genevieve Patterson and James Hays, Brown University

Global, binary attributes describing:
- Affordances / Functions (e.g. farming, eating)
- Materials (e.g. carpet, running water)
- Surface Properties (e.g. aged, sterile)
- Spatial Envelope (e.g. enclosed, symmetrical)

Statistics of database:
- 14340 images from 717 scene categories
- 106 attributes
- 2 million+ labels collected so far
- Outlier workers manually graded, good workers ~90% accurate.
Enumerating Objects
good for scene retrieval

Standard detection task, e.g. PASCAL
But picture is worth… 4 words?
Where can I sit?

Couch

Lamp

Couch

Table
Labeling Pixels

See Alan and Lana talk…
3D Scene Understanding

See Martial and Silvio’s talk...
Pushing Back Evaluation Horizon…

• So far, we have acted as cognitive psychologists:
  – proposing and evaluating intermediate “mental representations”
  – E.g. object(scene) categories, pixel labels, geometry

• But we can also be more behaviorist:
  – Focus on tasks instead of “mental states”
  – Evaluate action plans and behavior predictions
Task-Specific Questions

Pushable, Reachable, Sittable ……
Human Centric Scene Understanding

Reasoning in terms of set of allowable actions/body poses.
Human Workspace

3D Scene Geometry

Joint Space of Human-Scene Interactions

Gupta et al, CVPR’11
Subjective Interpretation
Holy Grail: Predicting the Future
Event prediction

Event prediction

Input image

Video database

Liu, Yuen, Torralba. CVPR 2009; Yuen, Torralba. ECCV 2010
What do we need to get to there?

The Op-Ed Part 😊
Organizing Our Data

“It irritated him that the “dog” of 3:14 in the afternoon, seen in profile, should be indicated by the same noun as the dog of 3:15, seen frontally…”

“My memory, sir, is like a garbage heap.”

Fumes the Memorious
Jorge Luis Borges
Trouble with Classic Platonic Categorization

• Step 1: cut up the world into categories

• Step 2: train an SVM on positives vs. negatives
It gets more complicated...

• Number of objects * number of interactions * number of outcomes… = too many categories

• Don’t want to categorize too early
  – “Dealing with the world as it comes to us” [Derek]

• Let’s categorize at run-time, once we know the task!
The Dictatorship of Librarians
categories are losing money…

vs.

Yahoo! vs. Google
Association instead of categorization

Ask not “what is this?”, ask “what is this like”

– Moshe Bar

• Exemplar Theory (Medin & Schaffer 1978, Nosofsky 1986, Krushke 1992)
  – categories represented in terms of remembered objects (exemplars)
  – Similarity is measured between input and all exemplars
  – think non-parametric density estimation

• Vanevar Bush (1945), Memex (MEMory EXtender)
  – Inspired hypertext, WWW, Google…
Bush’s Memex (1945)

- Store publications, correspondence, personal work, on microfilm
- Items retrieved rapidly using index codes
  - Builds on “rapid selector”
- Can annotate text with margin notes, comments
- Can construct a *trail* through the material and save it
  - Roots of hypertext
- Acts as an external memory
Visual Memex, a proposal

Types of edges:
- Visual similarity
- Spatial, temporal co-occurrence
- Geometric structure
- Language
- Geography

Milosewicz, Efros, NIPS’08
Poster Spotlight: Relative Attributes

- Previous work restricts attributes to binary categories, but many attributes are more fluid and should be expressed relatively.

[Parikh & Grauman, ICCV 2011]
Relative attributes

[Parikh & Grauman, ICCV 2011]

- Learn a ranking function per attribute, given ordering constraints among exemplars or categories

  Youth: ![Images](image1.png), ![Images](image2.png), ![Images](image3.png) ...

- Allows two novel tasks:
  1) Zero-shot learning from comparisons
     Train: “Unseen person C is younger than S, older than H”, ...
  2) Description relative to examples/classes
     is more dense than ![Images](image4.png), and less dense than ![Images](image5.png)

Precise descriptions are more recognizable to human subjects
Poster Spotlight: Ensemble of Exemplar-SVMs for Object Detection and Beyond

Milosewicz, Gupta, Efros, ICCV 2011
Exploring Photobios

Ira Kemelmacher-Shlizerman\textsuperscript{1}  Eli Shechtman\textsuperscript{2}  Rahul Garg\textsuperscript{1,3}  Steven M. Seitz\textsuperscript{1,3}

\textsuperscript{1}University of Washington\textsuperscript{*}  \textsuperscript{2}Adobe Systems\textsuperscript{†}  \textsuperscript{3}Google Inc.

Source  Automatically generated transition  Target