1970s

Early Days of Computer Vision
Introduction

- Datasets with 5 or 10 images
- Large-Scale Experiment: 800 photos (Takeo Kanade Thesis, 1973)
Today

Visual Data is Exploding!
Introduction

- Billions of cell phones equipped with cameras
- 3 Billion Facebook photo uploads per month (100M photos per day)
- 500 Billion consumer photos are taken each year world-wide, 633 million photos taken per year in NYC alone
- The United Kingdom has installed more than 4 million security cameras over the past decade
- Tremendous amount of visual Data available (Era of Big Data)
Introduction

Personal photo albums

Movies, news, sports

Google Image Search
Picasa
flickr
webshots
picsearch
YouTube

Surveillance and security

Medical and scientific images

Slide Credit: Lana Lazebnik

Visual Recognition And Search
Columbia University, Spring 2013
Introduction

- How to annotate large amounts of visual data?
- Crowdsourcing: Amazon Mechanical Turk
- ImageNet: **Millions of Annotated Images**, Thousands of Categories

http://www.image-net.org/
## ImageNet 2012 Challenge

<table>
<thead>
<tr>
<th>Team name</th>
<th>Filename</th>
<th>Error (5 guesses)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperVision</td>
<td>test-preds:141-146.2009-131-137-146.2011-145f.txt</td>
<td>0.15315</td>
<td>Using extra training data from ImageNet Fall 2011 release</td>
</tr>
<tr>
<td>SuperVision</td>
<td>test-preds:131-137-145-135-145f.txt</td>
<td>0.18422</td>
<td>Using only supplied training data</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_wLACs_weighted.txt</td>
<td>0.25172</td>
<td>Weighted sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_weighted.txt</td>
<td>0.26602</td>
<td>Weighted sum of scores from classifiers using each FV.</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_summed.txt</td>
<td>0.26646</td>
<td>Naive sum of scores from classifiers using each FV.</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_wLACs_summed.txt</td>
<td>0.26952</td>
<td>Naive sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.</td>
</tr>
</tbody>
</table>

Impressive results obtained by deep convolutional networks.
Exciting Time for Computer Vision

- + DATA
- + Computational Processing
- + Advances in Computer Vision and Machine Learning
Plan for Today

- Visual Recognition and Search: Applications & Challenges
- Student Introductions
- Course Overview and Requirements
- Syllabus Tour
- Project Ideas
Applications

Visual Recognition and Search
Self-Driving Cars

- Object Recognition and Tracking – mostly 3D point clouds (laser scanner)
- Passed Driver’s license test in Nevada!

Check Sebastian Thrun short TED talk:
http://www.ted.com/talks/sebastian_thrun_google_s_driverless_car.html

CVPR 2012 talk with more details:
http://techtalks.tv/talks/self-driving-cars/56391/
Applications

Mars Exploration Rover

- Autonomous Navigation
- Highly textured environment – good for stereo!
- Terrain Classification
- Automatic Photo capture of interesting scenes
- Horizontal velocity estimation in landing – feature tracking
Applications

Gaming

- Kinect Camera: RGB + Depth sensor
- Human pose Detection and Gesture Recognition
- See J. Shotton et al, Real-Time Human Pose Recognition in Parts from a Single Depth Image, CVPR 2011 (Best paper award)
Applications

Surveillance

IBM Smart Vision Suite (SVS)

User driven queries
- Find red cars
- Find tailgating incidents involving this person

Analytics & Framework
Watches the video for alerts & events
- Analytics modules:
  - Object tracking and classification
  - Face capture and recognition
  - License Plate Recognition
  - Many others
- Gathers event meta-data & makes it searchable
- Provides plug and play framework for analytics

Real-time alerts
- Perimeter violation
- Tailgating attempt
- Red car on service road

DVR - records & streams video
Video Capture/Encoding & Management

Sensors & Transactions
Watch for Abnormal Events!
What did you notice in the video?

- People walking?
- Vehicles driving?
- Abandoned bag???
People Search in Surveillance Videos
Video Demo:
http://rogerioferis.com/demos/PeopleSearch_BlurredFaces.wmv
Cashier Fraud Detection: 
Faked Scans (aka 
Sweethearting fraud)

Giving-away of merchandise 
without charge to a "sweetheart" 
customer (e.g., friend, family, fellow 
employee) by faked product scan 

Q. Fan et al, Recognition of 
Repetitive Sequential Human 
Activity, CVPR 2009
Like a normal field guide…

- that you can search and sort
- and with visual recognition

See N. Kumar et al, "Leafsnap: A Computer Vision System for Automatic Plant Species Identification, ECCV 2012"
- Nearly 1 million downloads
  - 40k new users per month
  - 100k active users
- 1.7 million images taken
  - 100k new images/month
  - 100k users with > 5 images
- Users from all over the world
- Botanists, educators, kids, hobbyists, photographers, …

Slide Credit: Neeraj Kumar
Applications

Google Goggles

[Image of a smartphone with Google Goggles app open, showing a picture of Golden Gate Bridge and the app's results page with the location of Golden Gate Bridge, San Francisco, United States.]

[Image of a smartphone showing an app called SnapTell, with a picture of the movie "Happy Feet" on the screen and the app's results page with options to watch the movie online or get movie-related information.]
Applications

Auto-focus by Face Detection

Video Demo:
http://www.rogerioferis.com/demos/FaceTrackerRealTime.avi

TAAZ Virtual Makeover  
http://www.taaaz.com/

Creating your own new look is easy!

1. Upload your photo
2. Apply some makeup
3. Choose a hairstyle
Finding Visually Similar Objects
Exploring Community Photo Collections
http://phototour.cs.washington.edu/
Applications

Organizing Photo Collections – IBM IMARS

http://researcher.watson.ibm.com/researcher/view_project.php?id=877
Applications

Many More Applications...

Medical Imaging

Biometrics

Handwritten Digit Recognition

And so on ....
Does it really work?

Many problems still remain to be solved...
Challenges

- Illumination
- Object pose
- Clutter
- Occlusions
- Intra-class appearance
- Viewpoint

Slide Credit: Kristen Grauman
Introductions

Student Info Form (see website: Homework #0)
Send by Monday, Jan 28
Course Goals:

- Understand and analyze the state-of-the-art in visual recognition and search.

- Hands-on experience with cutting-edge methods for visual recognition and search.

Note: 7:00pm – 8:50pm (3 credits)
Pre-requisites:

1) Courses/Background in Computer Vision or Image Processing is mandatory.

2) Programming skills in C/C++ or Matlab are also required.
Course Overview

Course Format:

☐ Each class will consist of:
  ✓ Lecture by instructor
  ✓ Student presentations and discussion

☐ Project
  ✓ Final project report: Paper (4-8 pages)

☐ Grading
  ✓ Participation (10%) Presentations (30%) Project (60%)
  ✓ See late policy in the course webpage
Course Overview

Resources:

- Check course webpage for:
  - Publicly Available Source Code
  - Datasets
  - Related Courses
Presentations:

- Students will be divided in groups to present a specific topic (covering 3 papers – see schedule in the course webpage).

- Groups and topic assignment will be decided next week. More information about the format and length/time of the presentation will be available in the course webpage.

- All students will have to read the papers before each class and participate in the discussion.
Presentations:

- Presentation slides will have to be sent to instructors at least one day before the presentation.

- Grading based on clarity, organization, knowledge, and whether the presentation meets the time constraints.

- Check [http://www.slideshare.net/cameraculture/how-to-give-a-good-talk](http://www.slideshare.net/cameraculture/how-to-give-a-good-talk)
Course Overview

Projects:

- Groups of two or three depending on the total number of students attending the class.

- Keep checking deadlines in the project webpage.

- Project Proposal, Mid-Project Report (Status Update), and Final Project Report (4-8 pages)
Projects:

- Projects will be graded according to the following criteria:

  - Mid-Project Report (15%)
  - Final Project Report (40%)
    - Write-up (15%) - Clarity, language, organization, literature survey, references, discussion
    - Technical (15%) - Originality, correctness, depth
    - Evaluation and Results (10%) - Thoroughness in analysis and experimentation, results and performance
  - Project Presentation (5%)
Syllabus Tour

Broad set of topics; focus on the state-of-the-art
Low-Level Representation: Feature Detection and Description

Classical Descriptors: SIFT and SURF

Modern Descriptors for Real-Time Applications: FAST, BRISK, ...

Check ECCV 2012 Tutorial on Modern Features:
https://sites.google.com/site/eccv12features/slides
Mid-level Representation: feature coding and pooling

Classical Bag-of-Word Models

Modern higher-level representations:
Fisher vector and super-vector encoding, LLC, ...

[K. Chatfield et al, The devil is in the details: an evaluation of recent feature encoding methods, BMVC 2011]
Part-based and Hierarchical Models

Structureless

Rigid

Deformable Part Models

ConvNets / Hierarchy

Objects

Parts

Features
Syllabus Tour

High-level Representation: Attributes and Semantic Features

Output of Semantic Classifiers as Features, Zero-shot Learning, ...
Fast Indexing and Image Retrieval

Focus on Hashing, Compact codes, ...
Large-Scale Classification

Insights on how to handle huge amounts of data and categories
Fast Object Detection and Localization

Focus on efficient approaches based on cascades, branch-and-bound, ...
Extensions to 3D and Video

Visual Recognition with Depth Cameras (e.g., Kinect)

Action Recognition
Case Studies

IBM Smart Vision Suite (SVS)

IBM Multimedia Analysis and Retrieval System (IMARS)
NOT Covered:

- Basic Machine learning algorithms. We assume you are familiar with e.g., SVMs, ...

- Computer vision fundamentals. We assume you are familiar with e.g., edge detection, ...
Project Ideas
**Brief Description:** TV shows exhibit human faces in a wide range of poses, expressions, and illumination conditions. The idea of this project is to improve the accuracy of a naïve frontal face detector by placing a webcam in front of a TV, and using face tracking around high-confidence frontal face detections to discover novel face poses and expressions, which in turn can be used to augment the initial training data to obtain a more robust detector.

**Task #1:** Train a face detector using frontal face images only (e.g., using off-the-shelf methods such as convolutional nets or Adaboost learning)

**Task #2:** Implement a face tracker using clothing color as context. Automatically select high-confidence tracklets around frontal-face detections to discover non-frontal poses and novel expressions.

**Task #3:** Augment the training data of the original detector with the new data (novel poses/expressions) and train a new detector

**Task #4:** Compare the performance of the Bootstrapped detector with the original detector on the FDDB Dataset (ROC Curves)
Mapping the Universe: Vision for Astroinformatics

**Brief Description:** Astroinformatics is an emerging discipline at the intersection of astronomy and computer science. Millions of galaxy images are now publicly available (see SDSS and Galaxy Zoo website), with tremendous opportunities for large-scale visual recognition and search. This is a rather unexplored application field in computer vision. The goal of this project is to create an efficient system for large-scale Galaxy search based on images and textual descriptions.

**Task #1:** Dataset preparation – Contact instructors to obtain a large-scale galaxy dataset (.sql file) with fine-grained annotated attributes.

**Task #2:** Train semantic attribute classifiers (e.g., cigar-shaped, smooth, round, boxy, no buldge, etc.) for enabling attribute-based galaxy search based on textual descriptions.

**Task #3:** Train additional discriminative classifiers to enable efficient content-based retrieval.

**Task #4:** Quantitative Performance Evaluation.
Local Features for Low-resolution Images

**Brief Description:** SIFT, HOG and other local edge histogram-based features are the workhorse for a lot of modern computer vision applications. However, these feature descriptors are sensitive to low image resolution. This project aims to find the intrinsic resolution for local features, and explores whether there are better local representations for low resolution images.

**Task #1:** Prepare an image dataset for evaluation of local feature matching at multiple resolutions. You can generate multi-resolution images from standard evaluation packages (e.g., see VLFeat Benchmark) by subsampling, and also consider your own multi-resolution dataset containing several categories of objects, including human faces, cars, pedestrians, buildings, and beach scenes.

**Task #2:** Evaluate the performance of state-of-the-art feature descriptors (e.g., using VLFeat or OpenCV source code) at multiple resolutions. Find the resolution where these methods work and where they don’t.

**Task #3:** Design your own local feature descriptor for low resolution image matching.

**Task #4:** Perform quantitative analysis of your method and discuss whether it performs better than other existing methods in low-resolution images.
Historical Photo Finder

**Brief Description:** The goal of this project is to design a robust method for matching present-day photos with historical photos. As an example, the project could be used to match current views of the Columbia campus with historical Columbia photos.

**Task #1:** Create a dataset of present-day photos and corresponding historical photos. You may take advantage of Google/Bing image search using the option “black & white” to locate historical photos. You may also include in the dataset your own photos of the Columbia campus and corresponding historical images.

**Task #2:** Evaluate state-of-the-art local feature matching algorithms (e.g., using VLFeat or OpenCV source code) on your dataset.

**Task #3:** Design your own method for robust matching of recent photos with historical images.

**Task #4:** Perform quantitative analysis of your method and discuss whether it performs better than other existing methods in this particular scenario.
State-of-the-art Object Detection: DPM versus CNN

**Brief Description:** Deformable part models (DPM) have been the de facto approach for object detection in the vision community (top performance in Pascal VOC, ...). Recently, Convolutional Neural Networks (CNN) achieved breakthrough results in the ImageNet challenge, arising as a top contender approach. The goal of this project is to critically study these two state-of-the-art techniques, assessing their weaknesses and strengths, as well as similarities and differences.

- **Task #1:** Select a dataset to evaluate the two detectors
- **Task #2:** Train a DPM detector using available source code.
- **Task #3:** Train a CNN Detector using available source code
- **Task #4:** Plot ROC Curves with respect to different parameters (e.g., amount of training data, ...) and different conditions for error diagnosis. Critically draw conclusions about both approaches.

You may also extend the analysis to other datasets.
Is there Overfitting? Revisiting High-Dimensional Features

**Brief Description:** Non-sparse high-dimensional features may cause overfitting due to the large number of parameters involved in the learning process. However, recent studies on PASCAL VOC and ImageNet suggest that high-dimensional features always get superior results for both small-scale and large-scale image classification problems. This project aims to study the issue of overfitting in this context.

**Task #1:** Play with available source code for image classification using state-of-the-art feature coding techniques (e.g., supervector/fisher vector and sparse coding)

**Task #2:** Try the classification task in PASCAL VOC (for one category), with both low-dimensional features and high-dimensional features.

**Task #3:** Compare the training/testing accuracy subject to #dimensions and #training samples. Discuss whether overfitting occurs.

**Task #4:** Repeat your experiments on a subset of the ImageNet dataset and discuss the overfitting issue.
Brief Description: The goal of this project is to study object recognition for Kinect-style RGB-D cameras, using the “RGB-D dataset” which contains color and depth images for 300 household objects. The idea is to exploit the vast amount of 3D models available in Google 3D Warehouse to learn geometric attributes or 3D classemes [Torresani et al, 2010]. These geometric classifiers could be combined with RGB attribute classifiers to serve as high-level features for RGB-D object recognition from few examples.

Task #1: Train 3D attribute classifiers or 3D classemes using the 3D object models available in Google 3D Warehouse

Task #2: Train RGB attribute classifiers or RGB classemes using standard RGB datasets

Task #3: Learn RGB-D object classifiers using the output of the models above as high-level features

Task #4: Compare performance with the state-of-the-art, analyzing recognition accuracy with respect to the number of training examples.
**TrafficCam: Robust and Efficient Vehicle Detection**

**Brief Description:** Real-time video feeds of thousands of public traffic cameras are now available on the web. This rich source of data, however, has received little attention by the vision community. The goal of this project is to develop an online learning system for video-based vehicle detection in traffic cameras, using multiple features (e.g., appearance, blob shape, motion) and scene structure information. Note that the size, appearance, pose, motion of a car are directly correlated with its position in the video frame.

**Task #1:** Choose a specific traffic camera from the web and collect training and test data, considering various conditions (shadows, crowds, different time periods, etc.)

**Task #2:** Train a vehicle detector using online learning (e.g., neural net) with multiple features (e.g., motion blob obtained by background subtraction, appearance, motion)

**Task #3:** Incorporate other scene-specific features, such as vehicle size and position

**Task #4:** Perform quantitative performance evaluation (accuracy and efficiency), assessing the contribution of each feature
Fusing Feature Descriptors for Action Recognition

**Brief Description:** Action recognition is still in its early stage. There exists no consensus on which features work best across datasets. The goal of this project is to evaluate multiple state-of-the-art feature descriptors for action classification, and investigate whether the fusion of multiple features improve results. You may include your own actions in the evaluation dataset.

**Task #1:** Prepare your own dataset. You may record your own action by referring to e.g, HMDB51 datasets or MSR action datasets.

**Task #2:** Play with available source code for spatio-temporal feature detectors and descriptors (STIP, HOF, ...) using, for example, SVM as classifier.

**Task #3:** Develop a method for fusing multiple feature descriptors.

**Task #4:** Compare the performance of the fusion algorithm with individual descriptors in your dataset and at least another standard dataset.
How to Invent:

See http://www.slideshare.net/cameraculture/raskar-ideaheaxagonapr2010
Next Steps

- Don’t forget to submit Homework #0 (Student Info Form) if you are taking this class

- Check the presentation topics in the course website

- Reading for next class