Lecture 2: CS677

Aug 27, 2020

Review

- Previous class
  - Course requirements
  - Assignments, grading
  - Adding more students to the class
  - Summary of syllabus
- Today’s objective
  - Topics to be studied in class
  - Some example state-of-art apps
  - Human visual system (very briefly)
  - Image formation

Online Class

- How to maintain interactivity in lectures?
  - Please ask questions, participate in discussion
  - Chat seems effective, we will pause occasionally and answer chat questions
- Slack channel, Piazza other sharing tools can be added
  - Should not turn into sharing assignment solutions
- Office hours
  - Instructor, Tu, Th; 4:30-6:00 P.M., other times by appointment
  - Online Zoom Meeting https://usc.zoom.us/j/96312659047
  - Try setting appointments even during office hours
  - TA office hours posted separately

Pre-requisites

- Proficiency in Python
- Significant programming experience
- Basic Math: Calculus, Linear Algebra, Probability Theory
- ML, DL or AI courses are **not** pre-requisites

Exams and Grading

- There will be two in-class, closed book exams
  - Exam 1, 7th or 8th week of classes (will be announced >1 week in advance)
  - Exam 2, Nov 24, last class day (this is not a cumulative “final” exam)
  - Term paper due on December 3, 2:00 PM (university requirement to have a “summative” experience)
    - We have requested approval to drop this requirement
- Grading weights
  - Assignments 30%; Exam1: 25%, Exam2: 25%; Term paper: 10%, Class attendance 10% (DEN students will be assumed to have perfect attendance)
    - If term paper requirement is removed, Exam1 and Exam2 will count for 30% each

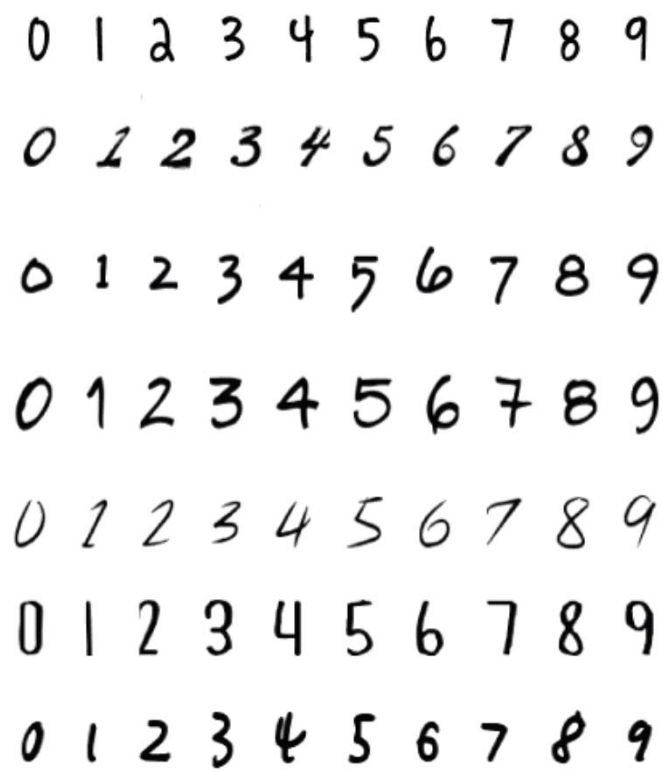
Course Objectives

- Understanding the key **problems** of vision
- **Alternative approaches** to solving the fundamental problems
- Specific **applications** will be covered only to illustrate the basic techniques
- Provide enough **background for further study** and for **implementation of some practical vision systems**
- We begin with no prior knowledge of computer vision but still will study several very recent techniques (hence "advanced" in the title)
- However, it is **not possible to cover "everything" about "everything"**
  - Not even all state-of-art methods can be covered; these change rapidly
    - >2000 major papers published each year

Why is Vision Hard?

- Seems easy to us, no conscious effort is needed by human viewers
- Small variations in human population’s ability to see/perceive
  - Does not require training/education for everyday tasks
- Can’t we just recognize objects based on “how they look”?
  - Isn’t a pen (a chair) a pen (chair) because it looks like a pen (chair)?
  - What does a pen (chair) look like?
  - Do we memorize images of pens or extract some more abstract representations (such as thin, mostly cylindrical objects with a conical section narrowing to a small circle at the end)?
  - We also need to detect/segment objects from others

From MNIST Database



From MNIST Database

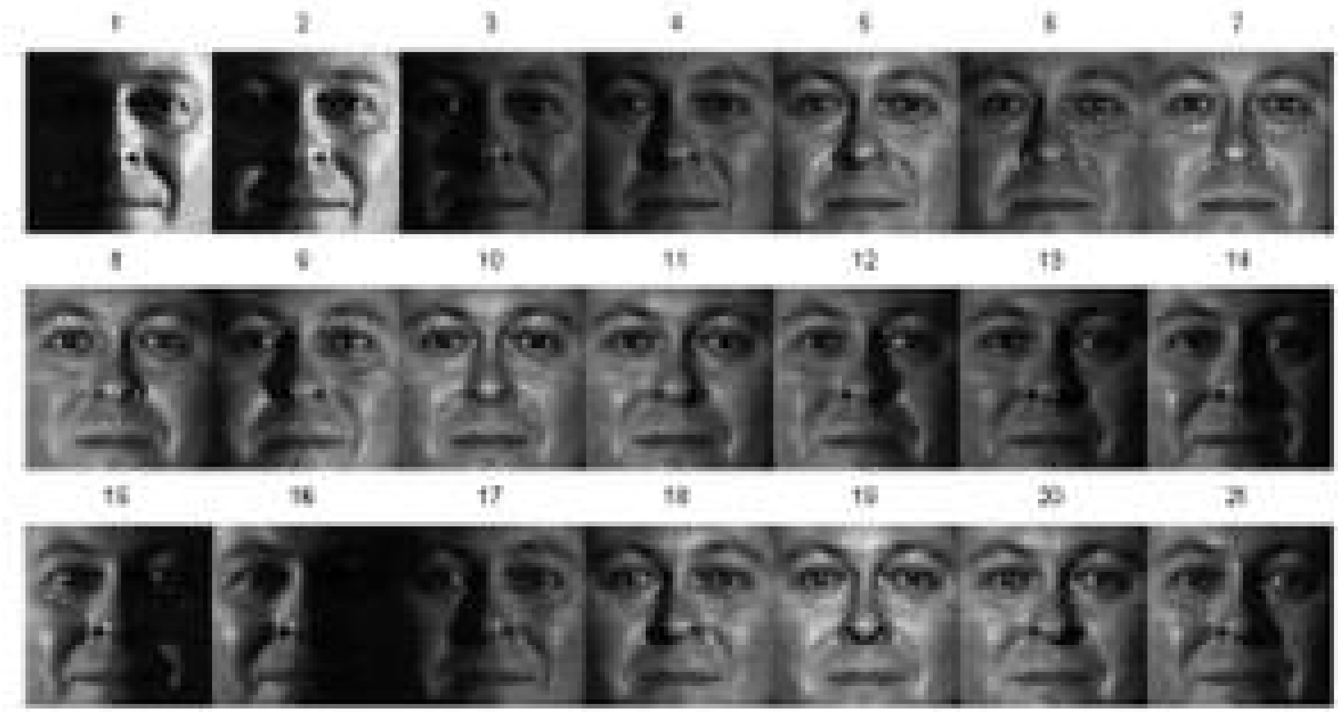
Same Object Class?

Some Issues: Representation

- What is representation of an object
- Objects of same class can have large variations in shape, size, color, material and other properties
  - Think about every day objects, such as chairs, coffee mugs, telephones…
- What is representation of an action (say throw an object)?
- Same action can be performed in different ways by different actors or even the same actor at different times or in different contexts

Viewpoint Change Examples

Illumination Change Examples



Find Objects in this Image

- Where is the object of interest? (Figure-ground problem)
- Do we need to know we are looking for a bicycle?
- How do we know if the object is a bicycle?
  - Do we need to know bikes have two wheels, handlebar etc
  - If so, how do we find the wheels and the other parts?

Find Objects

- What is figure, what is ground?
- Different shape of bicycle, with a rider
- What color is the backpack of the rider?
- How far is the fence from the biker?

Additional Complexities

- Harder to segment figure from ground
- If we draw a box around bicycle, image will also have a car in it. Do we need to separate the two before we can recognize or do we recognize first and then separate?
- How far is the car from the bicycle?

Depth Ambiguity and Occlusions

- World is 3-D, images are 2-D
  - There is an inherent loss of information; process is not truly invertible
    - Many 3-D environments could produce the same 2-D images
  - Our perception of 3-D from single 2-D images must take advantage of some regularities of the natural world
    - How do we isolate and exploit these regularities?
- Occlusion is (almost) ever-present
  - Objects occlude one another
  - Self-occlusion

Complexity

How many objects are in this image?

What can we say about each?

What can we say about this scene?

Two Major Components of the Objectives

- Infer 3-D scene geometry
  - Needed for navigation and manipulation
  - May be helpful for object/activity recognition
  - How can we infer 3-D info from a single 2-D image?
  - Can we use multiple images to simplify the problem?
  - Can we measure 3-D directly (and bypass some basic vision problems)?
  - Above problems relatively well understood, many working systems
- Semantic understanding
  - Recognition of objects, relations, activities…
  - Difficult to formulate mathematically
  - Very active area of research: methods have changed from "intuitive" to "statistical" to "deep learning"

Model-Based or Data-Driven Analysis

- All vision problems can be stated as learning a function between input and output, say $\mathbf{y} = f(\mathbf{x})$
- If *f* can be described (or well approximated) by an analytical function, say a polynomial in case of scalar values, the task reduces to find the parameters of the function
- An alternative is to fit *f* by a composition of simpler functions:
  - $f(\mathbf{x}) = f_3(f_2(f_1(x)))$; each f_i may be simple but non-linear function
  - This is the approach taken by deep learning
- Which is better?
  - If *f* is indeed a simple, derivable function, we can be confident of the solution; otherwise, it may "underfit" the data
  - Deep learning is susceptible to "overfitting" and requires huge amounts of training data
  - Transparency, ease of human interaction

Evolution of Computer Vision Approaches

- Early methods used representations based on intuition
  - “Hand-designed” descriptors and classification rules
- Later methods incorporated sophisticated mathematical models
  - These turned out to be very effective for recovering 3-D geometry from multiple images as problem can be posed as one of solving a system of algebraic equations
  - Less effective for semantic analysis such as object segmentation and recognition
    - Trend was to use hand-designed features but machine learned classifiers
- Recent and Current trend
  - Let machine learn the complete pipeline though structure of the pipe is still defined by designers
  - Achieves much higher accuracies when sufficient training data is available but methods are not transparent; hard to find source of errors

What kind of methods are we going to study?

- A combination of “model-driven” and “data-driven” methods
- More emphasis on mathematical methods in first part of the course as the geometry problems are relatively well-defined
- More emphasis on machine learning (deep learning) in second part as problems are not easy to describe in precise math terms
- Anticipation that future systems will use a combination of techniques so best to learn basic principles of both
- Traditional methods continue to be used for applications; job interviewers seem to test for these skills as well
- For students interested in DL only, it may be better to just take CS566

Topics to be studied in this class (*updated*)

- **Introduction (1 week)**
  Background, requirements and issues, human vision.
- **Image formation: geometry and photometry (1.5 weeks)**
  Geometry, brightness, quantization, camera calibration, photometry
- **Image segmentation (1.5 weeks)**
  Region segmentation, Edge and line finding
- **Multi-view Geometry (1.5 weeks)**
  Shape from stereo and motion, feature matching, surface fitting, Active ranging
- **Object Recognition: Traditional Methods (1.5 weeks)**
  HoG/SIFT features, Bayes classifiers, SVM classifiers
- **Neural Network Basics (1 week)**
  - Neural nets, CNNs, Backprop, SGD, Batch Normalization
- **Object Recognition: (2.5 weeks)**
  Image classification, object detection, semantic segmentation, Human pose estimation
- **Adversarial Attacks and Defense (.5 week)**
- **Motion Analysis and activity Recognition (1 week)**
  - Optical flow, motion features, classification network
- **Selected Topics (1 Week)**
  Face Identification, Vision and language …

Next Class

- Read ch. 1 of Forsyth/Ponce book
  - Sections 1.1, 1.2.