Introduction

Introduction to Computer Vision
CSE 152
Lecture 1

• We’ll start with some introductory material

• And end with
  – Syllabus
  – Organizational materials
  – Prerequisites

What is Computer Vision?

• Szeliski (Text): “In computer vision, we are trying to describe the world that we see in one or more images and to reconstruct its properties.”

• Trucco and Verri (secondary text): “Computing properties of the 3-D world from one or more digital images”

• Ballard and Brown: “The construction of explicit, meaningful description of physical objects from images.”

• Forsyth and Ponce: “Extracting descriptions of the world from pictures or sequences of pictures.”

• Stockman and Shapiro: “To make useful decisions about real physical objects and scenes based on sensed images.”

Why is this hard?

What is in this image?
1. A hand holding a man?
2. A hand holding a mirrored sphere?
3. An Escher drawing?
4. A 1935 self portrait of Escher

• Interpretations are ambiguous
• The forward problem (graphics) is well-posed
• The “inverse problem” (vision) is not
The challenge has been underestimated:
We all make mistakes.

- “In 1966, Minsky hired a first-year undergraduate student and assigned him a problem to solve over the summer: connect a television camera to a computer and get the machine to describe what it sees.”
  – Crevier 1993, pg. 88

- “640K ought to be enough for anybody.”
  – Bill Gates, 1981

What do you see?
- Changing viewpoint
- Moving light source
- Deforming shape

What was happening
- Changing viewpoint
- Moving light source
- Deforming shape

Should computer vision follow from our understanding of human vision?
Yes & No

1. Who would ever be crazy enough to even try creating machine vision?
2. Human vision “works”, and copying is easier than creating.
3. Secondary benefit – in trying to mimic human vision, we learn about it.

1. Why limit oneself to human vision when there is even greater diversity in biological vision
2. Why limit oneself to biological vision when there may be greater diversity in sensing mechanisms?
3. Biological vision systems evolved to provide functions for “specific” tasks and “specific” environments. These may differ for machine systems
4. Implementation – hardware is different, and synthetic vision systems may use different techniques/methodologies that are more appropriate to computational mechanisms

Hermann Grid
Scan your eyes over the figure. Do you see the gray spots at the intersections? Stare at one of them and it will disappear.

Why does this happen? Is it useful? Eye does not seem be acting like a camera an faith fully capturing scene?
How many red X’s are there? 
Raise your hand when you know.

Ubiquitous Vision

• Digital video has become really cheap.
• It’s widely embedded in cars, phones, games, doorbells, etc.
• 99.9% of digitized video isn’t seen by a person.
• That doesn’t mean that only 0.1% is important!
• And there’s an enormous amount of image and video content on the internet...

How many red X’s are there? 
Raise your hand when you know.

Applications: touching your life

• Optical Character Recognition
• Football
• Movies
• Surveillance
• HCI – hand gestures
• Aids to the blind
• Face recognition & biometrics
• Road monitoring
• Industrial inspection
• Virtual Earth; street view
• Robotic control
• Autonomous driving
• Space: planetary exploration, docking
• Medicine – pathology, surgery, diagnosis
• Microscopy
• Military
• Remote Sensing
• Digital photography
• Google Goggles
• Video games
Vision to explore the world

Place Recognition

Image from Microsoft Virtual Earth (see also: Google Earth)

Vision to explore other worlds

- Panorama stitching
- 3D terrain modeling
- Obstacle detection
- Position tracking

For more, read "Computer Vision on Mars" by Matthies et al.

Vision to look at stuff:
Object recognition

- Point & Find, Nokia
- SnapTell.com (now Amazon)
- Google Photos
- Apple Photos
- Google Image Search

Visual Field Guides:
Leafsnap.com -> Dogsnap -> Birdsnap -> Merlin, iNaturalist, etc

Object recognition (in supermarkets)

"A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized, the cashier verifies the quantity of items that were found under the basket, and continues to close the transaction. The item can remain under the basket, and with LaneHawk, you are assured to get paid for it..."
Amazon Go
1. Turn-style entry. Consumer scans in with Amazon App on smartphone
2. Consumer goes around the store, picks up items, adds to bag, shops like normal
3. Consumer exits

Vision to look at people
- Digital cameras, phones, Facebook, Google Photos, Snapchat etc.

Face recognition
Who is she?

Vision-based biometrics
“How the Afghan Girl was Identified by Her Iris Patterns” Read the story

Login without a password…
Fingerprint scanners on laptops, mice, other devices
But all is not so bright

Pastors are getting in on Ring’s ever-growing surveillance state

Vision to create special effects: Shape Capture

Vision to create special effects: Motion Capture

The Matrix movies, ESC Entertainment, XYZRGB, NRC
Vision for Entertainment

Sportvision first down line

Nice explanation on www.howstuffworks.com

Vision for augmented reality

- AR Toolkit
- Blippar
- Magic Leap
- Microsoft HoloLens

Augmented reality

- Text detection, localization, and translation, then render with similar font

Augmented reality

- Pokémon Go

Vision in Immersive Augmented Reality

- Gaze tracking
- Head pose estimation

Material and lighting estimation

Depth estimation

Semantic segmentation

Vision for smart cars

- Mobileye
  - Vision systems currently in high-end BMWs

Slide content courtesy of Amnon Shashua
**Vision for medicine**

3D imaging
MRI, CT

Image-guided surgery
Grimson et al., MIT

**Vision for Science**

Molecular Reconstruction from Cryo-electron Microscope Images
Mallick, Zhu, Kriegman

**coralnet.ucsd.edu**

CoralNet
A Web Solution for Coral

**Vision to bridge communication between physical and digital worlds:**

Optical Character Recognition (OCR)

License plate readers
Digit recognition, AT&T labs

Handwriting recognition

**Scene Text: Text Recognition in the Wild**

COCO-Text
A large-Scale Scene Text Dataset
https://bgshih.github.io/cocotext/

**Automatic image captioning**

- Deep learning

- Human captions from the training set

- Automatically captioned
**Automatic image captioning**

- Deep learning

**Video understanding**

- Video classification
- Activity recognition
- Video segmentation
- Activity detection

**Related Fields**

- Graphics
- Computational Photography

**Related Fields**

- Deep Learning

**Four Rs of computer vision**

- Reconstruction
  - Multiple view geometry, structure from motion, shape from X (where X is texture, shading, contour, etc.), etc.
- Registration
  - Tracking, alignment, optical flow, correspondence, etc.
- Reprojection
  - Rendering a scene and features from a different view, under different illumination, under different surface properties, etc.
- Recognition
  - Recognizing objects, scenes, events, etc.

Others may have slightly different Rs.
Rudiments: The implied fifth R

- image filtering
- interest point detection
- edge detection
- probability
- statistics
- linear algebra
- projective geometry
- optics

- Fourier analysis
- sampling
- algorithms
- photometry
- physics of color
- human vision
- psychophysics
- performance evaluation

How do we solve computer vision problem such as face recognition?

1. Craft a solution using software libraries of established methods (e.g., OpenCV) and tailor them to the particulars of the problem.
2. Build mathematical/physical model of the problem and implement algorithm with provably correct properties.
3. Gather image data, label it, and use machine learning to provide solution.

Craft a solution

- Distance between facial features (eye center, tip of nose, lip corners, eyebrows) should be good for recognition.
- Feature descriptors like SIFT or ORB are pretty insensitive to lighting and available in OpenCV
- Let’s detect those features, build a Delaunay triangulation, and use an SVM for classification.

An example problem:

Facial appearance as a result of differences in lighting is

1. A source of information
2. An annoyance

Illumination variability can be a real annoyance

“The variations between the images of the same face due to illumination and viewing direction are almost always larger than image variations due to change in face identity.”

-- Moses, Adini, Ullman

Build mathematical/physical model and use it to create algorithms and then code

1. Construct a model of the domain (usually mathematical, based on physics).
2. Prove properties of that model to better understand the model and opportunities of using it.
3. Develop algorithms to solve a problem that is correct under the model.
4. Implement & evaluate it.
5. Question assumptions of the model & start all over again.
Step 1. Build a Mathematical/Physical Model: Image Formation

At image location \((x,y)\) the intensity of a pixel \(I(x,y)\) is

\[ I(x,y) = a(x,y) \cdot n(x,y) \cdot s \]

where

• \(a(x,y)\) is the albedo of the surface projecting to \((x,y)\).
• \(n(x,y)\) is the unit surface normal.
• \(s\) is the direction and strength of the light source.

Step 2. Prove properties under model

The set of images of a Lambertian surface with no shadowing is a subset of 3-D linear subspace.

\[ L = \{ x \mid x = Bs, \forall s \in \mathbb{R}^3 \} \]

where \(B\) is a \(n\) by 3 matrix whose rows are product of the surface normal and Lambertian albedo.

Steps 3,4 : Develop a provable algorithms and implement it: Relighting

Steps 3,4 : Develop a provable algorithms and implement it: Photometric Stereo

Basic idea: 3 or more images under slightly different lighting

Step 5. Question Assumptions

• What about cast shadows?
• Many objects are glossy or have more complex reflectance functions

Use lots of data and machine learning

• Google FaceNet trained on hundreds of millions of cropped face images
• Triplet loss function
• 140 Million Parameters
• 1.6B Flops

[Schroff, Kalenichenko, Philbin, 2015]
About the Course

Teaching Staff

• Professor: David Kriegman
• TA’s:
  – Stephen Guerin
  – Nikhil Bangalore Mohan

Primary Text 1
Computer Vision: Algorithms and Applications

Primary Text 2
Deep Learning (Adaptive Computation and Machine Learning series)

Secondary Text

The Syllabus

• Discussion Section: Tuesday, 7:00-7:50PM in Center Hall 212
• Lab session: Contrary to what’s online, there is no separate lab section. Center Hall, Tuesday 8-8:50PM

http://cseweb.ucsd.edu/classes/wi20/cse152-a/
Background

- Linear algebra
- Multivariable calculus
- Probability
- Programming (we’ll use Python)

Announcements

- HW0 to be posted on by Wed class
- Read Chapter 1, Szeliski

Academic Integrity Policy

Integrity of scholarship is essential for an academic community. The University expects that both faculty and students will honor this principle and in so doing protect the validity of University intellectual work. For students, this means that all academic work will be done by the individual to whom it is assigned, without unauthorized aid of any kind.

Collaboration Policy

It is expected that you complete your academic assignments on your own and in your own words and code. The assignments have been developed by the instructor to facilitate your learning and to provide a method for fairly evaluating your knowledge and abilities (not the knowledge and abilities of others). So, to facilitate learning, you are authorized to discuss assignments with others; however, to ensure fair evaluations, you are not authorized to use the answers developed by another, copy the work completed by others in the past or present, or write your academic assignments in collaboration with another person. If the work you submit is determined to be other than your own, you will be reported to the Academic Integrity Office for violating UCSD's Policy on Integrity of Scholarship.