Lecture 1: Overview
Virtual classrooms

• Virtual lectures on Zoom
  – Only host shares the screen
  – Keep video turned off in case of bandwidth issues
  – Microphones muted unless speaking to reduce noise
  – But please do speak up (remember to unmute!)
  – Slides to be uploaded on webpage just before class
  – Whiteboard through camera and iPad to write things

• Lectures recorded and upload on Kaltura
  – Available under “My Media” on Canvas
Virtual classrooms

• Virtual interactions on Zoom
  – Ask and answer plenty of questions
  – Try to have in-class discussions
  – “Raise hand” feature on Zoom when you wish to speak
  – Instructor or TAs can unmute
  – Post questions on chat window
  – TA will help keep track of raised hands and chat window
  – If instructor cannot notice, TA will interrupt at logical pause

• Happy to try other suggestions
Enrollment logistics

• Waitlist
  – You are welcome to attend lectures even if on waitlist
  – To limit TA workload, we can grade only enrolled students

• Canvas
  – All enrolled and waitlisted students should have access

• All announcements will be posted on Piazza
  – Send email to TA (CC instructor) if did not get Piazza invite

• 2 units or S-U
  – Send instructor an email and CC TAs
  – Assignments optional, graded on final exams and presentation
Course details

- Class webpage:
  - [http://cseweb.ucsd.edu/~mkchandraker/classes/CSE152B/Spring2020/](http://cseweb.ucsd.edu/~mkchandraker/classes/CSE152B/Spring2020/)

- Instructor email:
  - mkchandraker@eng.ucsd.edu

- TA: Rui Zhu
  - Emails: rzhu@eng.ucsd.edu

- Grading
  - 35% assignments
  - 25% midterm (open notes)
  - 40% final exam (open notes)

- Aim is to learn together, discuss and have fun!

CSE 152B, SP20: Manmohan Chandraker
Feedback from Quiz 0

• **Linear algebra**
  – Most were fine with eigenvalues and eigenvectors
  – Need reminders: rotation matrices
  – Discussion section: rigid-body transformations

• **Calculus**
  – Most remember gradients and stationary points
  – Need reminders: Hessians, conditions for saddle point
  – Discussion section: gradient, Hessian, chain rule differentiation

• **Optimization**
  – Most remember polynomial roots, limitations of gradient descent
  – Many forgot solving linear systems, mathematics of gradient descent
  – Discussion section: linear systems, SVD, background for optimization
Feedback from Quiz 0

• **Features**
  – Most remember interest points and descriptors
  – Many have forgotten SIFT
  – Briefly recap SIFT in lecture, our goal is deep learning for matching

• **Geometry**
  – Many remember epipolar lines, fundamental matrix
  – Need reminders: estimation, RANSAC
  – Briefly recap two-view geometry in lecture, our goal is SFM

• **Neural networks**
  – Most remember fully-connected layers and convolutions
  – Many forgot SGD or not familiar with architectures
  – Refresher for deep networks in lecture, also more in discussion section
Overall goals for the course

• Introduce the state-of-the-art in computer vision
• Study principles that make them possible
• Get understanding of tools that drive computer vision
• Enable one or all of several such outcomes
  – Pursue higher studies in computer vision
  – Join industry to do cutting-edge work in AI
  – Gain an appreciation of modern AI technologies
Using Deep Neural Networks
Deep learning is revolutionizing AI

- Tic-tac-toe (1952)
- Checkers (1994)
- Chess (1997)
- Atari (2015)
- Go (2016)

CSE 152B, SP20: Manmohan Chandraker
Computer vision is also riding the wave

- Autonomous driving (Waymo, Tesla, Mobileye, ....)
- Augmented reality (HoloLens, Oculus, MagicLeap)
- Social networks (Google, Facebook, ....)
- Mobile applications
- Surveillance
Traditional Image Categorization:
Training phase

Training

Training Images

Training

Image Features

Training Labels

Classifier Training

Trained Classifier

Slide credit: Jia-Bin Huang
Traditional Image Categorization:
Testing phase

Training Images

Training

Image Features → Classifier Training → Trained Classifier

Test Image

Testing

Image Features → Trained Classifier → Prediction

CSE 152B, SP20: Manmohan Chandraker

Slide credit: Jia-Bin Huang
Features have been key

SIFT [Lowe IJCV 04]

HOG [Dalal and Triggs CVPR 05]

SPM [Lazebnik et al. CVPR 06]

Textons

and many others:

SURF, MSER, LBP, GLOH, .....
Learning a Hierarchy of Feature Extractors

- Hierarchical and expressive feature representations
- Trained end-to-end, rather than hand-crafted for each task
- Remarkable in transferring knowledge across tasks

CSE 152B, SP20: Manmohan Chandraker
Significant recent impact on the field

Big labeled datasets → Deep learning

GPU technology

Error rates on ImageNet Visual Recognition Challenge, %

Sources: ImageNet; Stanford Vision Lab

CSE 252C, SP20: Manmohan Chandraker
Deep learning has opened new areas

- Advantages available for many areas of computer vision
  - Recognize objects across shape and appearance variations
  - Data-driven priors for 3D reconstruction
  - Predict long-term future behaviors in complex scenes
  - End-to-end training rather than expensive feature design

- Deep networks in this course
  - Will do a refresher of basic concepts
  - Will illustrate use for many problems
  - Only need functional understanding
  - Think of them as feature extractors
Limits of deep learning

• Deep networks are powerful
• But we must be aware of their limitations
Data: hardware and models scale more than labels

- More data helps
- 4 TB of data per day from a car
- Training effort
- Rare events matter more
- Purely supervised methods not scalable
Miles to go before ....

US vehicle miles traveled and proportionate fatality rate

Vehicle miles (tens of billions)

Traffic lights
Speed limits
Seat belts
Air bags
ABS
Electronic Stability Control

Annual deaths per billion miles

CSE 152B, SP20: Manmohan Chandraker
Miles to go before ....

Google's Total Autonomous Miles

TWO MILLION MILES!

20 million miles in Jan 2020
Miles to go before ....

CSE 152B, SP20: Manmohan Chandraker
Miles to go before ....

<table>
<thead>
<tr>
<th>SAE class*</th>
<th>Year</th>
<th>Availability of self-driving cars</th>
<th>Total autonomous test miles driven so far**</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 2</td>
<td>2016</td>
<td>Available today e.g. Tesla 'Autopilot'</td>
<td>Tesla (SAE class 2) 130m</td>
</tr>
<tr>
<td></td>
<td>2018</td>
<td>TESLA</td>
<td>Google (SAE class 3) 1.9m</td>
</tr>
<tr>
<td></td>
<td>2020</td>
<td>NISSAN, HONDA, HYUNDAI, TOYOTA, PSA GROUPE, Baidu, Volkswagen, BMW, Ford, FCA</td>
<td></td>
</tr>
</tbody>
</table>

275m miles are required to prove a self-driving vehicle is at least as safe as a human

---

* Levels 1 und 2 are assistance systems only. From level 3, the vehicle constantly monitors traffic. From level 4, driver intervention is not required even in an emergency

** To June (Tesla)/August 2016 (Google)

Sources: LSP Digital research, manufacturer information, SAE, RAND

CSE 152B, SP20: Manmohan Chandraker
Interpretability of outputs

Automobile industry wants models built by combining validated components

Trade-offs for various learning approaches

Generative and discriminative methods

Accuracy

CSE 152B, SP20: Manmohan Chandraker
Domain gaps: object detection for auto rickshaw
Getting self-driving to where it is needed the most

Number of people who will die on Indian roads:
- 20: During this lecture
- 400: By this time tomorrow
- 145000: By next year

Safety systems have reduced fatalities in Europe and China

Consumers Desire More Automated Automobiles
Consumers Trust Driverless Cars

<table>
<thead>
<tr>
<th>Country</th>
<th>Trust Driverless Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>95%</td>
</tr>
<tr>
<td>India</td>
<td>86%</td>
</tr>
<tr>
<td>France</td>
<td>45%</td>
</tr>
<tr>
<td>China</td>
<td>70%</td>
</tr>
<tr>
<td>UK</td>
<td>45%</td>
</tr>
<tr>
<td>USA</td>
<td>60%</td>
</tr>
<tr>
<td>Germany</td>
<td>37%</td>
</tr>
<tr>
<td>Japan</td>
<td>28%</td>
</tr>
<tr>
<td>Canada</td>
<td>52%</td>
</tr>
</tbody>
</table>


CSE 152B, SP20: Manmohan Chandraker
Security: all features are potential targets
Security: algorithms may face adversarial attacks
Limits of deep learning

- Large scale labeled data is not always available
- Lack of generalization to unseen domains
- Good at narrow “classification”, not at broad “reasoning”
- Lack of interpretability
- Lack of reliability, security or privacy guarantees
New approaches to overcome limits

• Weak supervision
• Semi-supervision
• Self-supervision
• Domain adaptation
• Physical modeling
• Privacy-preservation
Perception, understanding, prediction, adaptation

**Real-time perception and 3D scene understanding**

- CVPR 2017
- CVPR 2015
- CVPR 2018
- ECCV 2018

**Unsupervised domain adaptation**

- No adaptation
- Our adaptation
  - CVPR 2018, ICCV 2019

**Weakly supervised 3D reconstruction**

- CVPR 2017

**Long-term multimodal future behavior prediction**

- CVPR 2017
Geometric and semantic 3D reconstruction

- **Metric learning for correspondence**
- **Weakly supervised semantic reconstruction**
- **Input**
- **Occlusion reasoning and large transformations**

**Deep supervision for occlusion-reasoned parts**

**CVPR 2017, PAMI 2018**

**NeurIPS 2016, ECCV 2018**

**CVPR 2016**

**ECCV 2018, CVPR 2019**

CSE 152B, SP20: Manmohan Chandraker
Physically-based learning for shape and material

Materials and global illumination

Refractive interfaces

Shape and motion
Unsupervised adaptation for bias and privacy

**Face recognition for profile inputs**

- FF-GAN
- 3DMM Coefficients
- Pose-Variant Input
- Discriminator
- Recognition Engine
- Frontalized Output

**Privacy-aware visual recognition**

- Representations by Conventional Learning Methods
- Private Images
- Public Data
- Adversary
- w/o privacy protection
- ours (privacy-preserving)

**Car recognition across camera and lighting conditions**

- Elev: 10°
- Elev: 20°
- Elev: 30°

**From rainy to good weather**

- CVPR 2019
- CVPR 2018
- CVPR 2017
- AAAI 2020
- ICCV 2017

CSE 152B, SP20: Manmohan Chandraker
Take-home message

- Powerful new tools allow rapid progress
- Must gain insights and understand limitations
- Understand the old to inform the new
Correspondence

Relate projections of the same point in two or more images of the scene.
Find features that are invariant to transformations

- geometric invariance: translation, rotation, scale
- photometric invariance: brightness, exposure, ...
Correspondence

\[ G(x, y, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \frac{x^2 + y^2}{\sigma^2}} \]

\[ G(x, y, k\sigma) - G(x, y, \sigma) \approx (k - 1)\sigma^2 \nabla^2 G. \]
Correspondence using CNNs

Similar?

CNN

FC Layers

Similarity

CSE 152B, SP20: Manmohan Chandraker
Semantic correspondence

Matching across sub-categories

Input image

Large-scale dataset

Single-view 3D reconstruction

CSE 152B, SP20: Manmohan Chandraker
Stereo

Two images captured by a purely horizontal translating camera (rectified stereo pair)

\[ x_2 - x_1 = \text{the disparity of pixel } (x_1, y_1) \]

Estimating the disparity is equivalent to estimating depth.
Stereo: Graph cuts

Compute winning disparity at each pixel

\[ d(x, y) = \arg \min_d E(x, y; d) \]

\[ y = 141 \]

\[ E(x, y, d) \text{ the disparity space image (DSI)} \]
Stereo: Graph cuts

Compute winning disparity at each pixel
\[ d(x, y) = \arg \min_d E(x, y; d) \]

\[ y = 141 \]

\[ E(x, y, d) \text{ the disparity space image (DSI)} \]

[Boykov et al. “Graph cuts”]

CSE 152B, SP20: Manmohan Chandraker
Stereo: CNNs

\[ C_{\text{CNN}}(p, d) = -s(\langle \mathcal{P}_L(p), \mathcal{P}_R(p - d) \rangle) \]

[Zbontar and Le Cun, 2016]
Optical flow

*Brightness constancy constraint equation*

\[ I_x u + I_y v + I_t = 0 \]

- Number of equations and unknowns per pixel:
  - One equation, two unknowns \((u, v)\)
Optical flow: Lucas-Kanade

- Overconstrained linear system through patch coherence

\[
\begin{bmatrix}
I_x(p_1) & I_y(p_1) \\
I_x(p_2) & I_y(p_2) \\
\vdots & \vdots \\
I_x(p_{25}) & I_y(p_{25})
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix} = -
\begin{bmatrix}
I_t(p_1) \\
I_t(p_2) \\
\vdots \\
I_t(p_{25})
\end{bmatrix}
\]

\[
A d = b
\]

Least squares solution for \(d\) given by \((A^T A) d = A^T b\)

\[
\begin{bmatrix}
\sum I_x I_x & \sum I_x I_y \\
\sum I_x I_y & \sum I_y I_y
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix} = -
\begin{bmatrix}
\sum I_x I_t \\
\sum I_y I_t
\end{bmatrix}
\]

\[
A^T A \\
A^T b
\]

The summations are over all pixels in the K x K window
Optical flow: Coarse-to-Fine

- Gaussian pyramid of image 1
- Gaussian pyramid of image 2
- Run iterative L-K
- Warp and upsample

CSE 152B, SP20: Manmohan Chandraker
Optical flow: CNNs

CSE 152B, SP20: Manmohan Chandraker

[Dosovitskiy et al., “FlowNet”]
Structure from Motion (Basic)

- keypoints
- match
- fundamental matrix
- essential matrix
- $[R|t]$ triangulation

CSE 152B, SP20: Manmohan Chandraker
Structure from Motion \textsuperscript{(Real-Time)}

[CSE 152B, SP20: Manmohan Chandraker]

[Mur-Artal et al., "ORB-SLAM"]
Image retrieval use for SFM

When we see close points in feature space, we have similar descriptors, which indicates similar local content.

1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”
Structure from Motion

SFM robust to calibration, photometric variations, lens distortions and camera responses.

[Engel et al., “DSO”]
Face recognition

Verification

Identification

Intra-personal variation

Inter-personal variation

CSE 152B, SP20: Manmohan Chandraker

[Taigman et al., “DeepFace”]
Face alignment

[Taigman et al., “DeepFace”]

[Blanz et al., “3DMM”]

Normalization

[FF-GAN]

Augmentation

[Yin et al., “FF-GAN”]

[Zhu et al., “3DDFA”]
Metric learning for face recognition

\[ \|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \]

Metric learning: Features of different person further by at least a margin.

Triplet loss: \[ \sum_{i}^{N} \left[ \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_{+} \]

Challenge: too many triplets satisfy the margin easily (hard negative mining).

CSE 152B, SP20: Manmohan Chandraker

[Schroff et al., “FaceNet”]
Human pose estimation

- Localization task: needs more accurate spatial localization
- Cascade structure: crop initial prediction and refine

[Toshev and Szegedy, "DeepPose"]

CSE 152B, SP20: Manmohan Chandraker
Semantic segmentation

Full image

Extract patch

Classify center pixel with CNN

Cow
Cow
Grass

- Inefficient inference
- No shared computation between neighbors
- No global reasoning

CSE 152B, SP20: Manmohan Chandraker
Semantic segmentation

Locally accurate boundaries ↔ Global reasoning
Semantic segmentation

Combining what and where

- Fuse coarse semantic information with local appearance

→ Skip connections
Semantic segmentation

Insert dilated convolution module to preserve resolution and obtain multiscale information
Object detection

Training:
1. Obtain training data
2. Define features
3. Define classifier

Given new image:
1. Slide window
2. Score by classifier

Training examples

Car or non-car Classifier

Feature extraction

Object detection

CSE 152B, SP20: Manmohan Chandraker
Object detection: sliding windows

HoG + SVM

Orientation Voting

Overlapping Blocks

Local Normalization

DPM

Score is sum of filter scores minus deformation costs

Multiscale model captures features at two-resolutions
Parts are represented at twice the resolution of the root filter.

Object detection: sliding windows

Dog? YES
Cat? NO
Background? NO

Dog? NO
Cat? NO
Background? YES
Object detection: region proposals

R. Girschick, et. al, “R-CNN”

J. R. Uijlings, et. al, “Selective search for object recognition”

CSE 152B, SP20: Manmohan Chandraker
Object detection

Huang et al. Speed/Accuracy Tradeoffs for Modern Convolutional Object Detectors. CVPR 2017

CSE 152B, SP20: Manmohan Chandraker
Multi-target tracking

[Images of people walking in a sequence, with boxes highlighting individuals, and arrows indicating movement and tracking]

Observation $E_{\text{det}}$  
Appearance $E_{\text{app}}$  
Dynamic Model $E_{\text{dyn}}$  
Mutual Exclusion $E_{\text{exc}}$  
Target Persistence $E_{\text{per}}$

high energy
low energy

Tracking Area

[Images of connections and tracking areas]

[CSE 152B, SP20: Manmohan Chandraker]
Multi-target tracking

CRF or energy minimization

(a) Inputs at $t$
(b) Hypotheses Generation
(c) CRF Inference
(d) Outputs at $t$

RNN for prediction and association


[Choi, 2017] [Milan et al., 2017]
Action recognition

- Need a representation for videos
- SIFT in video volume

[Laptev et al., “STIP”]

• Different video representation: multiscale, track features, aggregate along trajectories

[Wang et al., “Dense trajectories”]
Action recognition

- 3D convolutions
- Appearance and flow
- Large-scale dataset

Kay et al., “Kinetics Dataset”, 2017
Domain adaptation

Abundant synthetic labeled data, little or no real data with labels

Training

Feature layers

Segmentation classifier

Labels

Classify as simulation or real

Domain discriminator

Testing

Simulation image

Real image

Good

Good

CSE 152B, SP20: Manmohan Chandraker
Bias and fairness

NY Times, Feb 09, 2018

Color Matters in Computer Vision
Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

Gender was misidentified in up to 1 percent of lighter-skinned males in a set of 385 photos.

Gender was misidentified in 35 percent of darker-skinned females in a set of 271 photos.

Wired, Jan 11, 2018

Is the iPhone racist? Chinese users claim iPhoneX face recognition can’t tell them apart

Hmmm... not seeing this clearly yet. What can Google Lens do?

WHEN IT COMES TO GORILLAS, GOOGLE PHOTOS REMAINS BLIND

Google Lens, which tries to interpret photos on a smartphone, also appears unable to see gorillas. Screenshot: Wired

Hacker News, Dec 21, 2017
Bias and fairness

- MS-Celeb-1M: 82% Caucasian, 9.7% African-American, 6.4% East-Asian, less than 2% Latino or South-Asian
Take-home message

- Powerful new tools allow rapid progress
- Must gain insights and understand limitations
- Understand the old to inform the new