Lecture 0: Introduction

CSE 252C: Advanced Computer Vision
Manmohan Chandraker
Defining computer vision

Wall-E: Fact and Fiction (Minh Do, Princeton University)
Defining computer vision

• Old: Computer programs that can
  • Process image information
  • Recognize instances of objects
  • Find distances of objects

• Modern: Understanding the world based on visual cues
  • Determining factors that govern image formation
  • Recognition across variations
  • Estimate semantic properties of a scene
  • Recognize complex actions
  • Predict long-term behaviors
The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".
Studying computer vision

- Images are everywhere around us

Source: Domo
Studying computer vision

- Images are everywhere around us
- Rapidly emerging technologies

自治驾驶

智能家庭

工厂自动化

游戏
Studying computer vision

• Images are everywhere around us
• Rapidly emerging technologies
• Deep and attractive scientific problems
  • How do we recognize objects?
  • Why do newborn babies respond to face-like shapes?
  • Beautiful marriage of math, physics, biology, CS, engineering

[Farroni et al., 2005]
Augmented Reality
Vision in augmented reality devices

Gaze tracking

Head pose estimation

Object detection

Material and lighting estimation

Depth estimation

Semantic segmentation
Autonomous driving

Source: Wired
Autonomous driving

Where is our car?
Structure from Motion
Visual SLAM

Where are other agents?
Object detection
3D localization

What is a safe path?
Behavior prediction
Path planning

Where are scene elements?
Semantic segmentation
My Interests
Scene understanding for self-driving

Distillation networks for fast and accurate object detection

Distillation for compressed CNN (student) to mimic uncompressed CNN (teacher), to achieve greater accuracy at the same speed.

Learning to predict uncertain future behavior

DESIRE: Deep Stochastic IOC RNN Encoder-Decoder
- Deep CVAE (autoencoder) to generate diverse hypotheses.
- RNN to rank predictions based on motion, scene and interactions.
- Deep inverse reinforcement learning for long-term future rewards.

Monocular SFM

Learning to simulate

Reinforcement learning for simulations. Ensuring diversity and coverage.

Large-scale, real-time, monocular SFM. Accuracy comparable to stereo systems.
Geometric and semantic 3D reconstruction

**Metric learning for correspondence**

- NIPS 2016, ECCV 2018

**Deep supervision for occlusion-reasoned parts**

- CVPR 2017, PAMI 2018

**Weakly supervised semantic reconstruction**

- CVPR 2016

**Occlusion reasoning and large transformations**

- ECCV 2018, CVPR 2019
Physically-based learning for shape and material

Materials and global illumination

Refractive interfaces

Shape and motion

SIGGRAPH Asia 2018

WACV 2018, 2019

ICCV 2017

WACV 2018, 2019

ICCV 2017
Unsupervised adaptation to new domains

Face recognition for profile inputs
- FF-GAN
- Recognition Engine
- Frontalized Output
- ICCV 2017

Reconstruction with unaligned data
- Mask Supervision
- Unlabeled 3D Shapes
- WGAN
- 3DV 2017

Car recognition across camera and lighting conditions
- Elev: 10°
- Elev: 20°
- Elev: 30°
- CVPR 2019

From rainy to good weather
- CVPR 2018
Deep Neural Networks
Deep learning is revolutionizing AI
Computer vision is also riding the wave

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

Training Images

Training

Image Features

Classifier Training

Trained Classifier

Training Labels

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

Training Images

Training
- Image Features
- Classifier Training
- Trained Classifier

Testing
- Image Features
- Trained Classifier
- Prediction Outdoor

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]

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
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
Deep learning has opened new areas

• Availability of large-scale image and video data
• Availability of computational power
  – Better and cheaper GPUs
  – Cloud computing resources
• Better understanding of how to train deep neural networks
• 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.
Limits of deep learning
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

Miles to go before ....

Google's Total Autonomous Miles

TWO MILLION MILES!
Miles to go before ....

![Graph showing the relationship between failure rate and miles needed to be driven to reach different failure rates. The graph indicates that 275 million miles need to be driven to achieve a failure rate of C = 95%.](image-url)
Miles to go before ....

<table>
<thead>
<tr>
<th>SAE class*</th>
<th>Year</th>
<th>Availability of self-driving cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 2</td>
<td>2016</td>
<td>Available today e.g. Tesla 'Autopilot'</td>
</tr>
<tr>
<td>3 to 5</td>
<td>2018</td>
<td><strong>TESLA</strong></td>
</tr>
<tr>
<td></td>
<td>2020</td>
<td><strong>NISSAN</strong>&lt;br&gt;<strong>HONDA</strong>&lt;br&gt;<strong>HYUNDAI</strong>&lt;br&gt;<strong>TOYOTA</strong>&lt;br&gt;<strong>PSA GROUPE</strong>&lt;br&gt;<strong>BAIDU</strong>&lt;br&gt;<strong>FORD</strong>&lt;br&gt;<strong>BMW</strong>&lt;br&gt;<strong>FCA</strong> (FIAT CHRYSLER AUTOMOBILES)</td>
</tr>
</tbody>
</table>

Total autonomous test miles driven so far**
- **Tesla** (SAE class 2): 130m
- **Google** (SAE class 3): 1.9m

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
Interpretability of outputs

Automobile industry wants models built by combining validated components

Trade-offs for various learning approaches

Generative and discriminative methods
Object detection for an auto rickshaw
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
New devices

- Time-of-flight sensors
- Structured light systems
- Light field cameras
- Coded apertures
Large-scale optimization

- Internet images pose challenges of scale and outliers
- Reconstructions with millions of images
- Choices to handle data
- Specific optimization approaches

Figure from Agarwal et al.
Real-time computation

- Mobile platforms, embedded systems (IoT devices)
- Stringent demands on computational resources
- Low power platforms (wattage) for automobile ECUs
- Carefully designed and multithreaded architectures

Song et al., CVPR 2014

Newcombe et al., CVPR 2015
Take-home message

• Computer vision is a key branch of AI
• Enables several modern applications around us
• A lot of highly visible and high-impact activity
• Huge industry interest
• This is a great time to study computer vision!
Course Details
Course details

• Each class will cover topics in computer vision

• Examples of topics
  • Correspondence
  • Stereo
  • Optical flow
  • Structure from motion
  • Face recognition
  • Human pose estimation
  • Material and lighting
  • Semantic segmentation
  • Object detection
  • Tracking
  • Action recognition
  • Domain adaptation
  • Privacy and fairness
Course details

• “Lightning” presentations
  – Four students to present in one class
  – Time limit: 7 minutes (5 for presentation, 2 for questions)
  – Papers to be assigned by instructor
  – Sign-up sheet will be posted
  – Send slides by 9pm two days before the class
  – Well-practiced and fluent presentations
  – Ask questions and encourage participation

• Each student does at least 1 presentation
Course details

• Presentation format (1 slide for each):

  1. Motivation and problem description
  2. Prior work
  3. Method overview
  4. Method details and analysis
  5. Experiments
  6. Future work and discussion
Course details

• Class webpage:
  – http://cseweb.ucsd.edu/~mkchandraker/classes/CSE252C/Spring2019/

• Instructor email:
  – mkchandraker@eng.ucsd.edu

• TA: Zhengqin Li
  – Email: zhl378@eng.ucsd.edu

• Grading
  – 15% presentation
  – 50% assignments (mini-projects)
  – 30% final exam
  – 05% participation

• Aim is to learn together, discuss and have fun!
Overview of Topics
Overview of topics

• Examples of topics
  • Correspondence
  • Stereo
  • Optical flow
  • Structure from motion
  • Face recognition
  • Human pose estimation
  • Material and lighting
  • Semantic segmentation
  • Object detection
  • Tracking
  • Action recognition
  • Domain adaptation
  • Privacy and fairness
Correspondence

Relate projections of the same point in two or more images of the scene.
Correspondence

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 as similarity
Correspondence using CNNs

Similar?

CNN

FC Layers

Similarity
Correspondence beyond similarity CNN

- Detection of interest points
- Normalization of patches
- Multiscale information
- Obtaining training data
- Efficient training and testing
Semantic correspondence

Matching across sub-categories

Single-view 3D reconstruction

Input image

Reconstruction

Large-scale dataset

Input Supervised Ours ILSVRC-CNN DSP
Stereo

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

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

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

Compute winning disparity at each pixel

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

\( y = 141 \)

\( E(x, y, d) \) the disparity space image (DSI)
Stereo: Graph cuts

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

\( y = 141 \)

\( E(x, y, d) \) the disparity space image (DSI)

(a) initial labeling  (b) standard move  (c) \( \alpha-\beta \)-swap  (d) \( \alpha \)-expansion

[Boykov et al. “Graph cuts”]
Stereo: CNNs

Left input image

Right input image

Output disparity map

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

[Zbontar and Le Cun, 2016]
Stereo: Weak Supervision (Monocular)

1. Left Image $I_l(x)$
2. Warp Image $I_w(x)$
3. Reconstruction Error

$\|I_w(x) - I_l(x)\|$
Optical flow

Brightness constancy constraint equation

\[ I_x u + I_y v + I_t = 0 \]
Optical flow: Ambiguity

**Brightness constancy constraint equation**

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

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

The component of the motion perpendicular to the gradient (that is, parallel to the edge) cannot be measured.

If \((u, v)\) satisfies the flow equation, so does \((u+ku', v+kv')\) for any \(k\) if

\[ I_x u' + I_y v' = 0 \]
Optical flow: Lucas-Kanade

- Overconstrained linear system through patch coherence

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

Least squares solution for \(d\) given by

\[
\left( A^T A \right) d = A^T b
\]

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

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

- Coarse-to-Fine image I
- Coarse-to-Fine image J
- Gaussian pyramid of image 1
- Gaussian pyramid of image 2
- Run iterative L-K
- Warp and upsample
- Run iterative L-K

Gaussian pyramid of image 1
Gaussian pyramid of image 2
Optical flow: Coarse-to-Fine

Lucas-Kanade without pyramids
Fails in areas of large motion

Lucas-Kanade with Pyramids

run iterative L-K
warp and upsample

Gaussian pyramid of image 1
Gaussian pyramid of image 2
Optical flow: CNNs

[Dosovitskiy et al., “FlowNet”]