인공지능

19차시 : Vision

22년 삼성 AI 전문가과정
6월 9일 목요일 4교시
장병탁

서울대학교 컴퓨터공학부
담당 교수: 장병탁

Seoul National University
Byoung-Tak Zhang




Lecture Overview

인공지능

19차시 : Vision

서울대학교 컴퓨터공학부
담당 교수: 장병탁

Seoul National University
Byoung-Tak Zhang




Introduction: Vision

- **Vision:** Around 80 percent of our sensory impressions are registered through our eyes.
- Most animals have eyes, often at significant cost. However, this cost is justified by the immense value that eyes provide.
- An agent that can see can predict the future:
  - It can tell what it might bump into
  - It can tell whether to attack or to flee or to court
  - It can guess whether the ground ahead is swampy or firm
  - It can tell how far away the fruit is
- In this chapter, we describe how to recover information from the flood of data that comes from eyes or cameras.

Detecting Objects

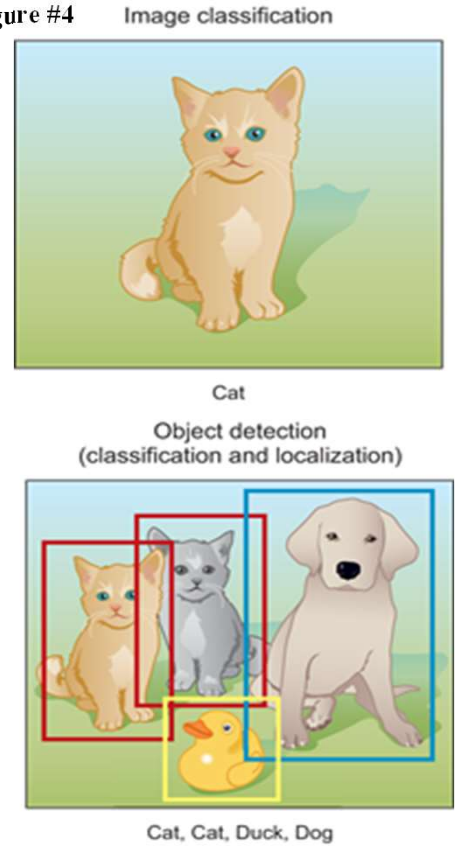
Image classifiers

- Predict what is in the image

Object detectors

- Find multiple objects in an image, report what class each object is, and report where each object is by giving a bounding box around the object
  - Looking at a small sliding window onto the larger image—a rectangle
  - Classify what we see in the window

출처: Figure #4



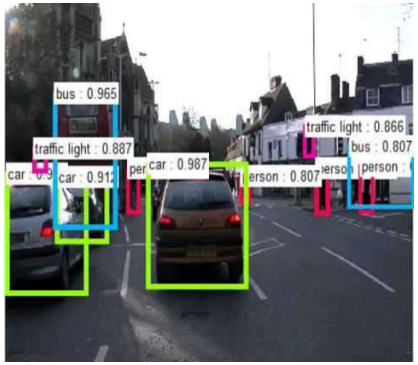
출처: Figure #5

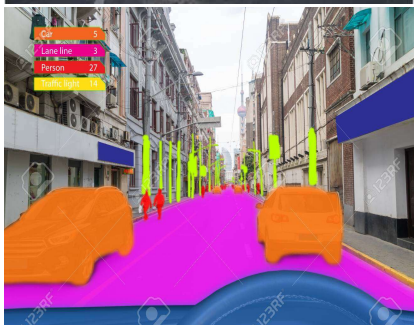
Visual Perception

- **Perception:** How we connect the agent to the raw, unwashed world?
  - **Perception** provides agents with information about the world
  - **Sensors** measure some aspects of the environment that are input to the agent.
  - **Sensory modalities:** vision, hearing, and touch etc.
- **Vision:** How the agent sees and understands the visual world?
  - **Sensor models** for vision $\mathbf{P}(\mathbf{E}_t \mid \mathbf{X}_t)$
  - **Object model:** 3D geometric model of objects
  - **Rendering model:** Physical, geometric, statistical processes
- **Approaches** to Visual Perception
  - **Feature extraction** approach: Simple computations to the sensor observations
  - **Recognition** approach: Labels each image with a yes or no
  - **Reconstruction** approach: Builds a geometric model of the world from images

Applications

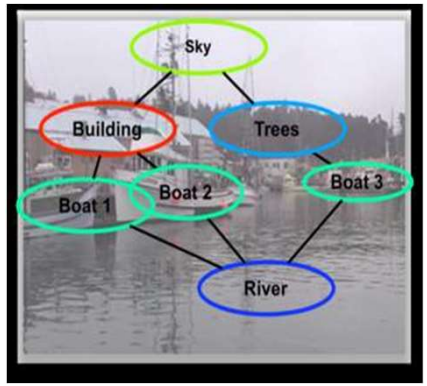
- Self-driving cars
- Machine inspection
- Optical character recognition (OCR)
- 3-D model building (photogrammetry)
- Retail (e.g. automated checkouts)
- Medical imaging
- Automotive safety
- Match move
- Motion capture (mocap)
- Surveillance
- Fingerprint recognition and biometrics





출처: Figure #2

출처: Figure #1



출처: Figure #3

Outline (Lecture 19)

19.1 Introduction

19.1 Introduction (1/1)

Vision

- A perceptual channel that accepts a stimulus and reports some representation of the world
- Feature
  - A number obtained by applying simple computations to an image
- The model-based approach to vision
  - **Object model**
  - **Rendering model**: describes the physical, geometric, and statistical processes
- Two core problems of computer vision
  - **Reconstruction**: an agent builds a model of the world from an image
  - **Recognition**: an agent draws distinctions among the objects

19.2 Image Formation

19.2 Image Formation (1/3)

Images without lenses: The pinhole camera

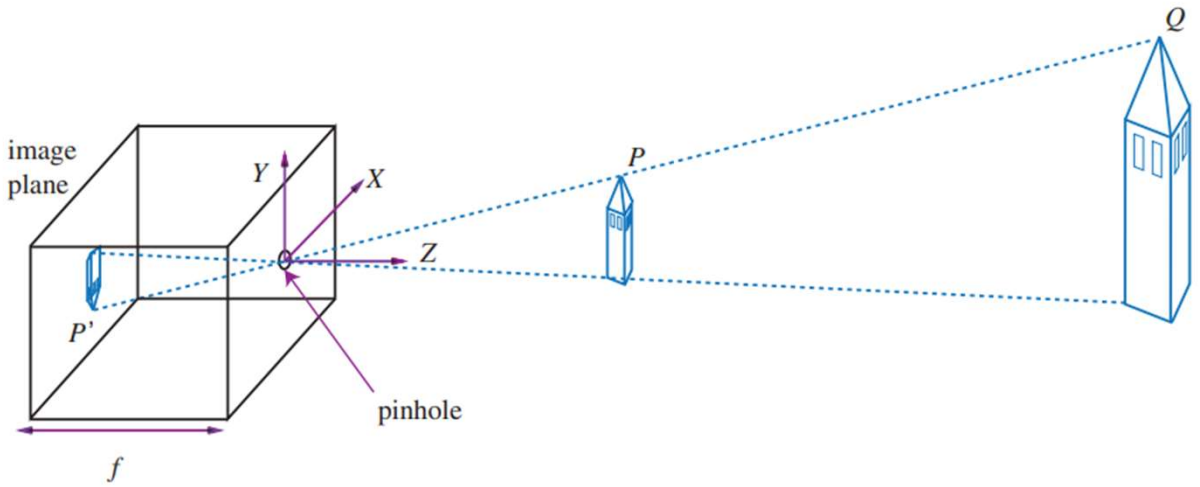


Figure 27.2 Each light sensitive element at the back of a pinhole camera receives light that passes through the pinhole from a small range of directions. If the pinhole is small enough, the result is a focused image behind the pinhole. The process of projection means that large, distant objects look the same as smaller, nearby objects—the point P' in the image plane could have come from a nearby toy tower at point P or from a distant real tower at point Q.

출처: Figure #6

19.2 Image Formation (2/3)

Lens systems

- A single piece of transparent tissue in the eye and a system of multiple glass lens elements in a camera

- A camera (or an eye) with a lens captures all the light that hits anywhere on the lens

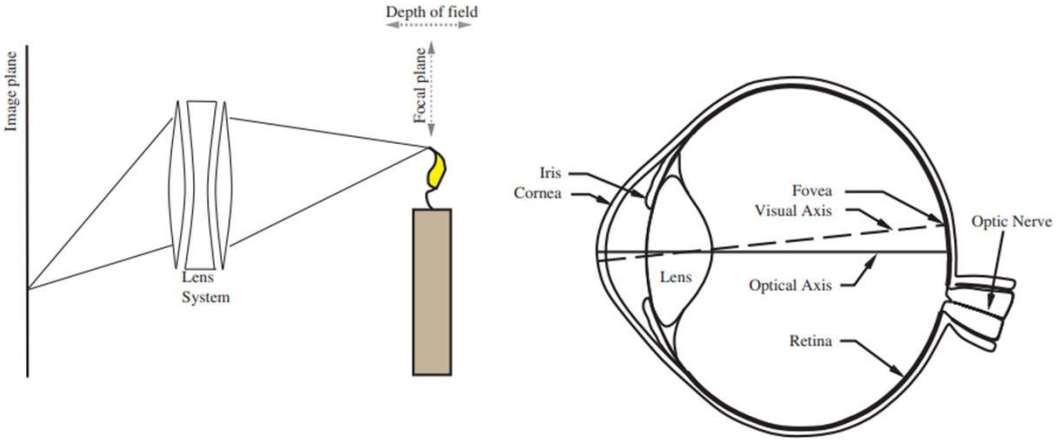


Figure 27.3 Lenses collect the light leaving a point in the scene (here, the tip of the candle flame) in a range of directions, and steer all the light to arrive at a single point on the image plane. Points in the scene near the focal plane—within the depth of field—will be focused properly. In cameras, elements of the lens system move to change the focal plane, whereas in the eye, the shape of the lens is changed by specialized muscles.

출처: Figure #7

19.2 Image Formation (3/3)

Light and shading

Figure 27.4 This photograph illustrates a variety of illumination effects. There are specularities on the stainless steel cruet. The onions and carrots are bright diffuse surfaces because they face the light direction. The shadows appear at surface points that cannot see the light source at all. Inside the pot are some dark diffuse surfaces where the light strikes at a tangential angle. (There are also some shadows inside the pot.) Photo by Ryman Cabannes/Image Professionals GmbH/Alamy Stock Photo.

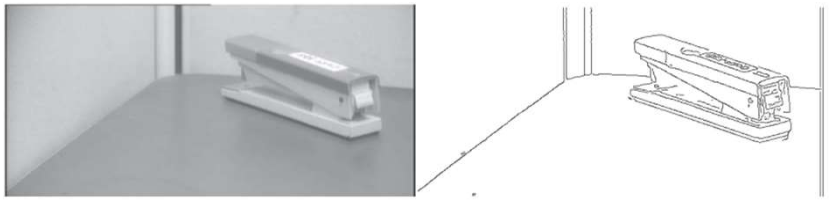
출처: Figure #8

19.3 Simple Image Features

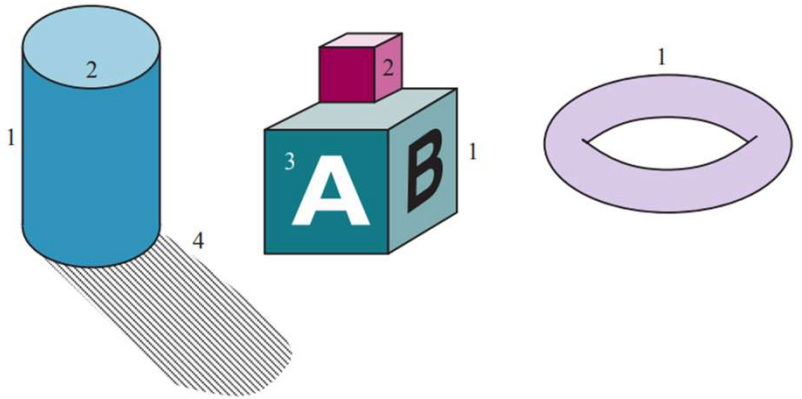
19.3 Simple Image Features (1/8)

Edge detection

- Edges are straight lines or curves in the image plane across which there is a "significant" change in image brightness.
- Kinds of edges
  1. Depth discontinuities
  2. Surface orientation discontinuities
  3. Reflectance discontinuities
  4. Illumination discontinuities (shadows)



출처: Figure #9



출처: Figure #10

19.3 Simple Image Features (2/8)

- **Gaussian filter**: a weighted average that weights the nearest pixels the most, then gradually decreases the weight for more distant pixels.
  - $G_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$ in 1D, or
  - $G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$ in 2D

- The application of the Gaussian filter replaces the intensity $I(x_0, y_0)$ with the sum, over all (x, y) pixels, of $I(x,y)G_\sigma(d)$, where d is the distance from (x_0, y_0) to (x, y).

19.3 Simple Image Features (3/8)

- We say that the function h is the convolution of two functions f and g (denoted $f * g$) if we have

$h(x) = (f * g)(x) = \sum_{u=-\infty}^{+\infty} f(u)g(x-u)$ in 1D, or

$h(x,y) = (f * g)(x,y) = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} f(u,v)g(x-u,y-u)$ in 2D

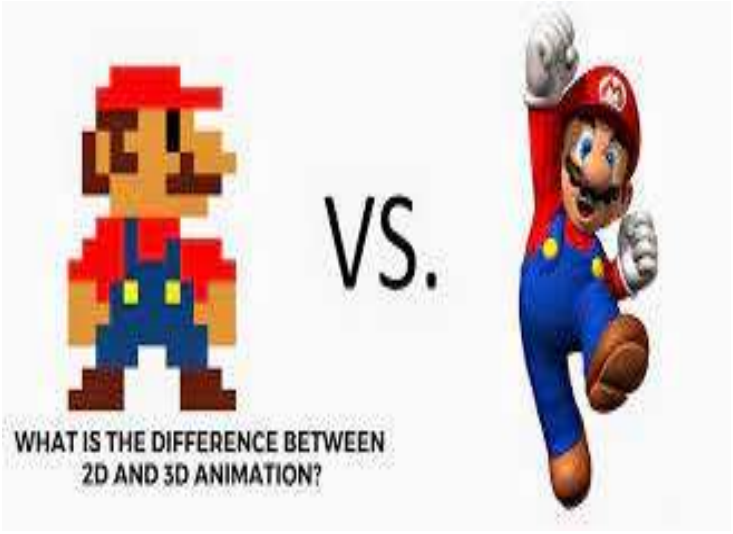
- So the smoothing function is achieved by convolving the image with the Gaussian, $I * G_\sigma$.

19.3 Simple Image Features (4/8)

Generalization to 2D images

- In two dimensions, edges may be at any angle θ.
- Considering the image brightness as a scalar function of the variables x, y, its

gradient is a vector $\nabla I = \begin{pmatrix} \frac{\partial I}{\partial x} \\ \frac{\partial I}{\partial y} \end{pmatrix} = \begin{pmatrix} I_x \\ I_y \end{pmatrix}$



19.3 Simple Image Features (5/8)

Generalization to 2D images

- Edges correspond to locations in images where the brightness undergoes a sharp change, and so the magnitude of the gradient, $\| \nabla I \|$, should be large at an edge point.

- Of independent interest is the direction of the gradient $\frac{\nabla I}{\|\nabla I\|} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$

- This gives us a $\theta = \theta(x, y)$ at every pixel, which defines the edge orientation at that pixel.

19.3 Simple Image Features (6/8)

Texture

- In everyday language, **texture** is the visual feel of a surface—what you see evokes what the surface might feel like if you touched it.

Optical flow

- When an object in the video is moving, or when the camera is moving relative to an object, the resulting apparent motion in the image is called optical flow.

출처: Figure #11

19.3 Simple Image Features (7/8)

Optical flow

- The optical flow vector field can be represented at any point (x, y) by its components $v_x(x, y)$ in the x direction and $v_y(x, y)$ in the y direction.
- Consider a block of pixels centered at pixel p, (x_0, y_0), at time t_0 . One possible measure of similarity is the sum of squared differences (SSD):

$$\mathrm{SSD}(D_x, D_y) = \sum_{(x,y)} \big(I(x, y, t) - I(x + D_x, y + D_y, t + D_t)\big)^2$$

- We find the (D_x, D_y) that minimizes the SSD. The optical flow at (x_0, y_0) is then $(v_x, v_y) = (D_x/D_t, D_y/D_t)$.

19.3 Simple Image Features (8/8)

Segmentation of images

- Breaking an image into **regions** of similar pixels.

(a) (b) (c) (d)

Figure 27.10 (a) Original image. (b) Boundary contours, where the higher the P_b value, the darker the contour. (c) Segmentation into regions, corresponding to a fine partition of the image. Regions are rendered in their mean colors. (d) Segmentation into regions, corresponding to a coarser partition of the image, resulting in fewer regions. (Images courtesy of Pablo Arbelaez, Michael Maire, Charless Fowlkes and Jitendra Malik.)

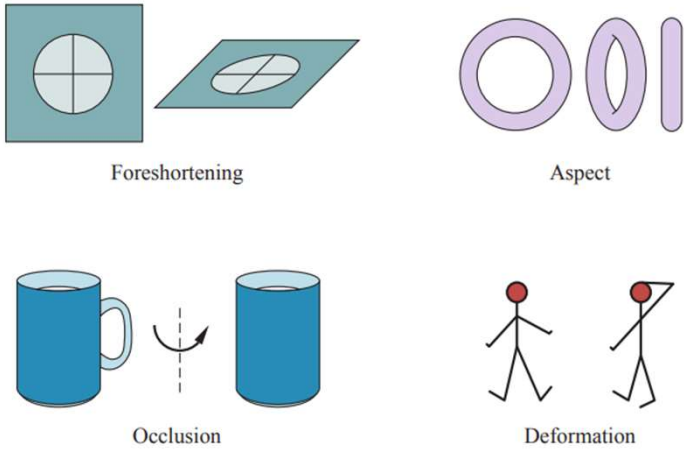
출처: Figure #12

19.4 Classifying Images

19.4 Classifying Images (1/4)

Complex appearance and pattern elements

- Foreshortening, which causes a pattern viewed at a slant to be significantly distorted.
- Aspect, which causes objects to look different when seen from different directions.
- Occlusion, where some parts are hidden from some viewing directions.
- Deformation, where internal degrees of freedom of the object change its appearance.



출처: Figure #13.1

19.4 Classifying Images (2/4)

Image classification with convolutional neural networks (CNNs)

➢ Why do CNN classify well?

- Features used by CNN classifiers are learned from data not hand-crafted by a researcher
- Availability of large, challenging data sets e.g. ImageNet
- Detect spatial relations between local patterns
- Deeper layers of convolutional neural networks represent “patterns of patterns”

➢ Dataset augmentation

- Training samples are copied and modified slightly
- E.g. Shifting, Rotation, Random cropping, Padding, Resizing

출처: Figure #13.2

19.4 Classifying Images (3/4)

ImageNet Large-Scale Visual Recognition Challenge (LSVRC)

- Image Classification/Localization
- 1.2M labeled images, 1000 classes
- Deep learning methods won since 2012:
  - 2012 non-CNN: 26.2% (top-5 error)
  - 2012: (Hinton, AlexNet)15.3%
  - 2013: (Clarifai) 11.2%
  - 2014: (Google, GoogLeNet) 6.7%
  - (pre-2015): (Google) 4.9%

    → Beyond human-level performance

출처: Figure #14

19.4 Classifying Images (4/4)

Deep convolutional networks

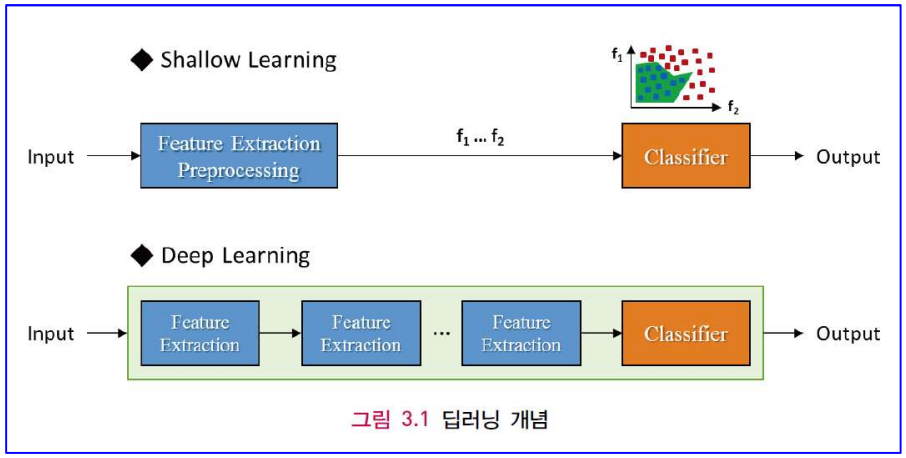
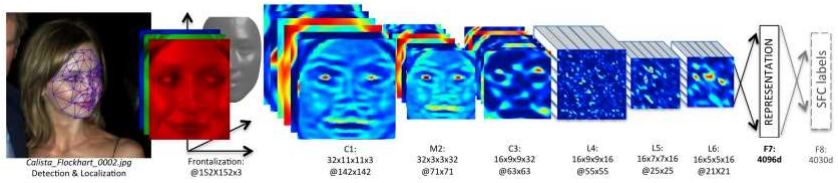


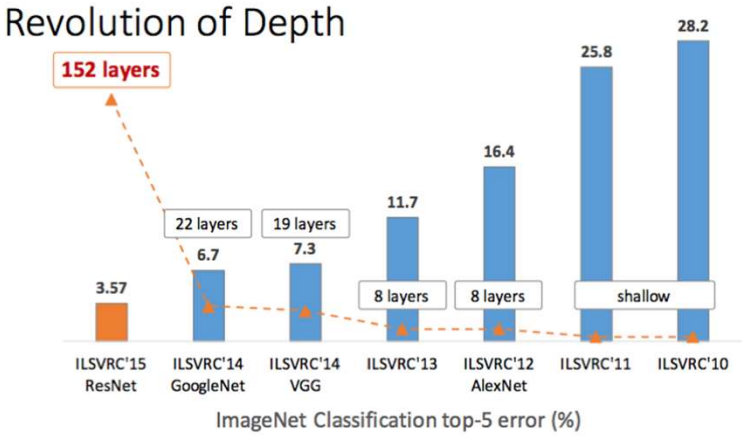
그림 3.1 딥러닝 개념

출처: Figure #15



출처: Figure #16

Depth vs. performance



출처: Figure #17

19.5 Detecting Objects

19.5 Detecting Objects (1/4)

Image classifiers

- Predict what is in the image

Object detectors

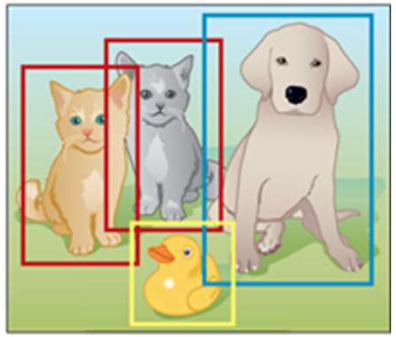
- Find multiple objects in an image, report what class each object is, and report where each object is by giving a bounding box around the object
  - Looking at a small sliding window onto the larger image—a rectangle
  - Classify what we see in the window

Image classification

Cat

출처: Figure #4

Object detection
(classification and localization)



Cat, Cat, Duck, Dog

출처: Figure #5

19.5 Detecting Objects (2/4)

Details to workout

- **Decide on a window shape**
  - Faster RCNN uses nine boxes: small, medium, and large sizes; and tall, wide, square aspect ratios
- **Build a classifier for windows**
  - Use CNN classifier
- **Decide which windows to look at**
  - Select ones that are likely to have interesting objects in them
  - Regional proposal network (RPN): A network that finds regions with objects
  - Region of interest (ROI): box with a good enough objectness score
  - ROI pooling: sampling the pixels to extract features

19.5 Detecting Objects (3/4)

Details to workout (cont.)

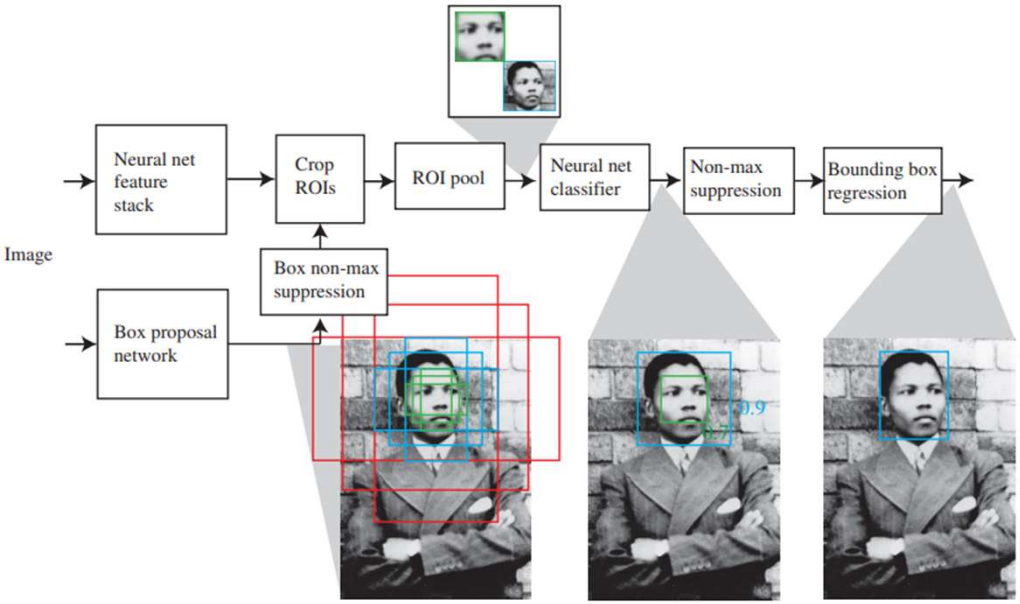
- **Choose which windows to report**
  - Don’t want to report the same object multiple times in slightly different windows
  - Perhaps only the objects that appear large in the image
  - Non-Maximum Suppression (NMS)
    - First, build a sorted list Non-maximum suppression of all windows with scores over a threshold.
    - Then, choose the window with the highest score
- **Report precise locations of objects using these windows**
  - Bounding box regression

19.5 Detecting Objects (4/4)

Faster RCNN

- First network computes "objectness" scores of candidate image boxes, called "anchor boxes," centered at a grid point

- Second network is a feature stack that computes a representation of the image suitable for classification



출처: Figure #21

19.6 The 3D World

19.6 The 3D World (1/2)

3D cues from multiple views

- Two images of the same scene taken from different viewpoints and know enough about the two cameras
  - Can construct a 3D model by figuring out which point in the first view corresponds to which point in the second view + geometry
- Two views of enough points, and know which point in the first view corresponds to
  - Do not need to know much about the cameras to construct a 3D model
  - Two views of two points gives you four *x, y* coordinates, and you only need three coordinates to specify a point in 3D space

19.6 The 3D World (2/2)

Binocular stereopsis

- **Disparity**

  Try!

  - 1) both index fingers up in front of your face, with one eye closed
  - 2) adjust them so the front finger occludes the other finger in the open eye's view
  - 3) Now swap eyes
  - 4) you should notice that the fingers have shifted position with respect to one another

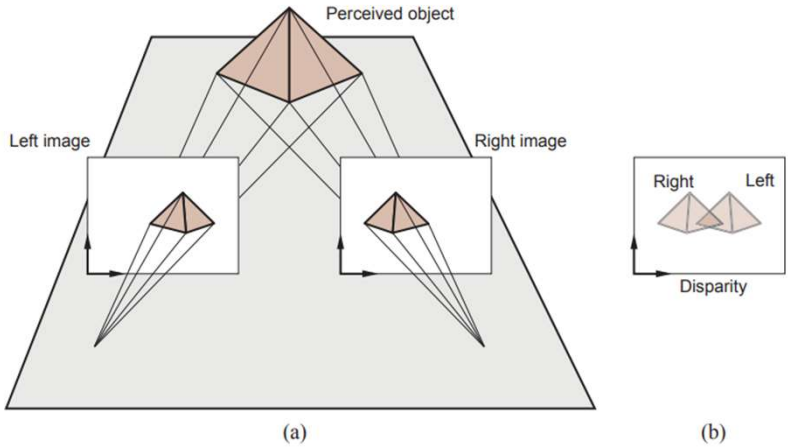


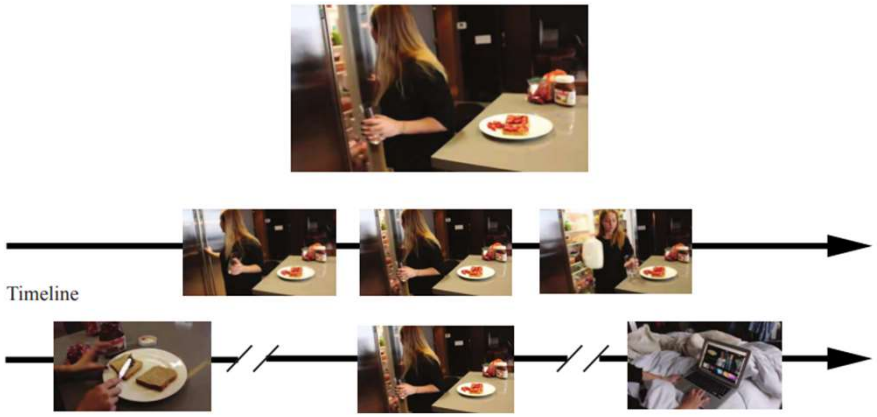
Figure 27.14 Translating a camera parallel to the image plane causes image features to move in the camera plane. The disparity in positions that results is a cue to depth. If we superimpose left and right images, as in (b), we see the disparity.

출처: Figure #22

19.7 Using Computer Vision

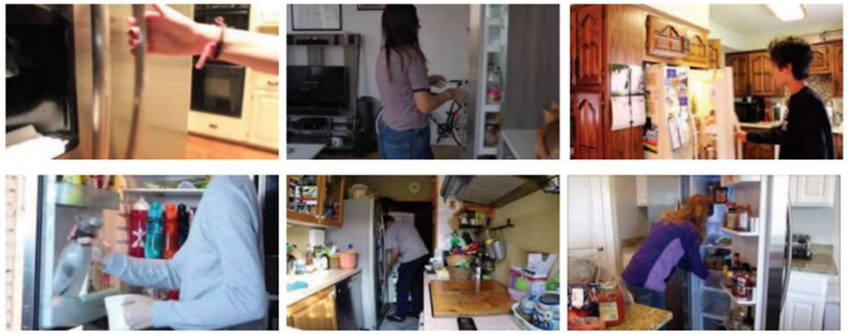
19.7 Using Computer Vision (1/7)

Understanding what people are doing



출처: Figure #23

출처: Figure #24

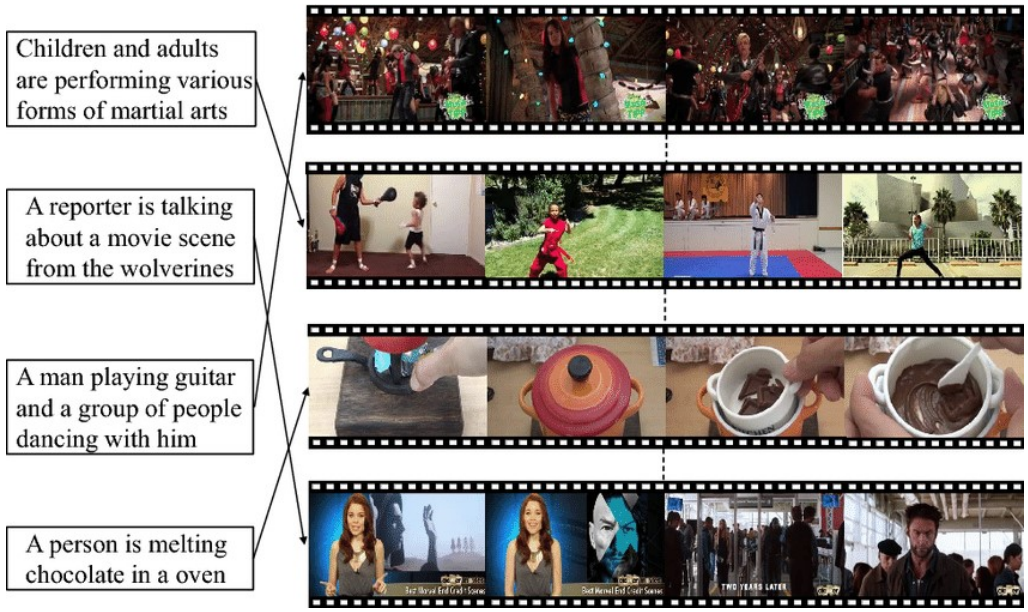


출처: Figure #25

19.7 Using Computer Vision (2/7)

Linking pictures and words

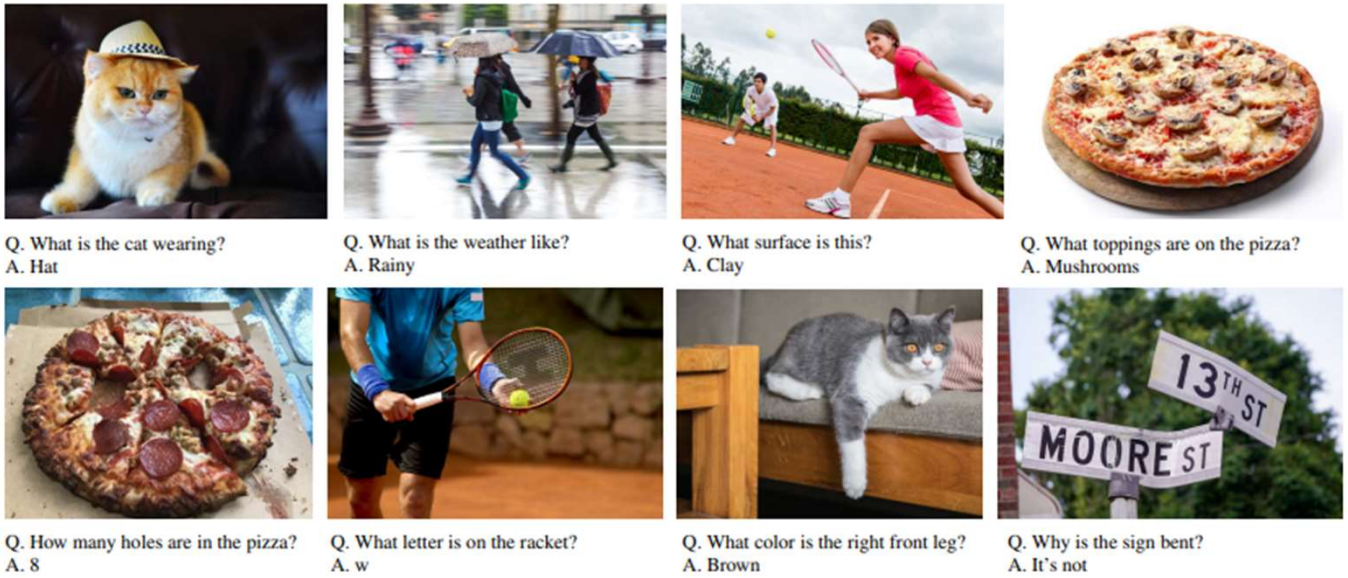
- How can we find the images we want?
  - Some of the images will have keywords or captions attached
  - For these, image retrieval can be like text retrieval.



19.7 Using Computer Vision (3/7)

Linking pictures and words

- Visual question-answering systems

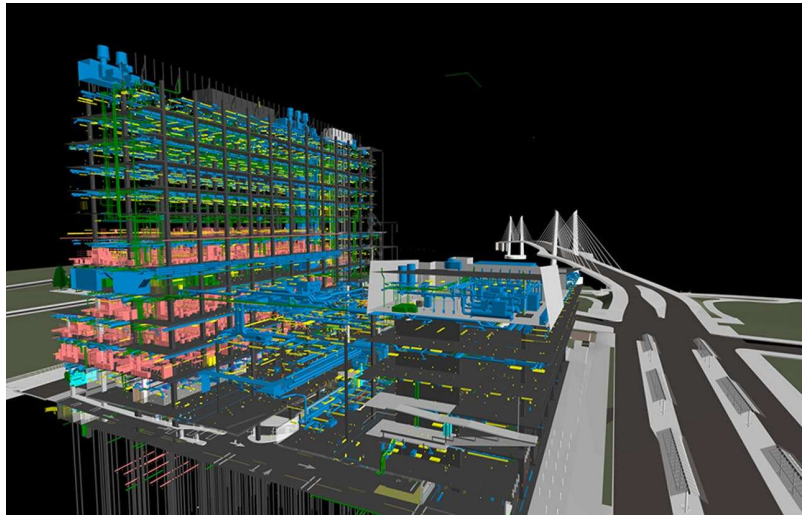


출처: Figure #26

19.7 Using Computer Vision (4/7)

Reconstruction from many views

- **Construction management**
  - Keeping track of what is happening during construction is difficult and expensive
  - Fly drones through the construction filming the current state
  - Build a 3D model of the current state and explore



출처: Figure #27

19.7 Using Computer Vision (5/7)

Making pictures

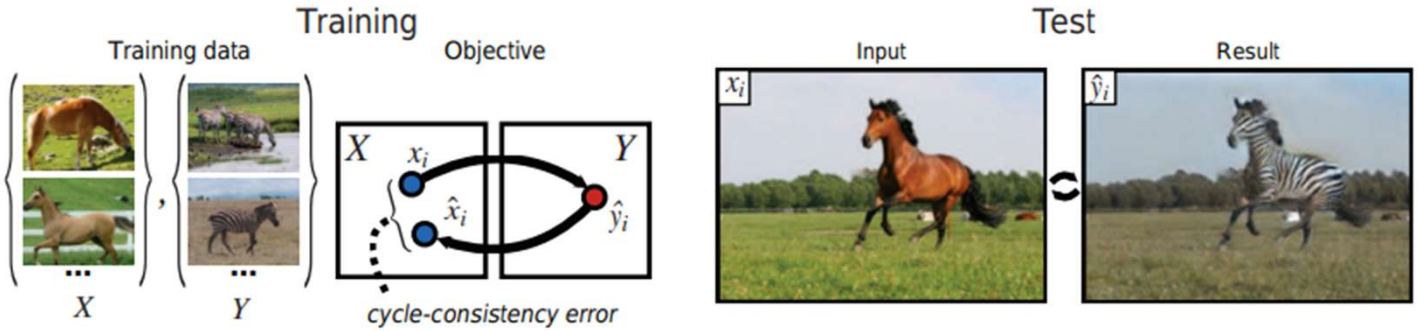


Figure 27.25 Unpaired image translation: given two populations of images (here type X is horses and type Y is zebras), but no corresponding pairs, learn to translate a horse into a zebra. The method trains two predictors: one that maps type X to type Y, and another that maps type Y to type X. If the first network maps a horse x_i to a zebra $\hat{y}_i$, the second network should map $\hat{y}_i$ back to the original x_i. The difference between x_i and $\hat{x}_i$ is what trains the two networks. The cycle from Y to X and back must be closed. Such networks can successfully impose rich transformations on images. Figure courtesy of Alexei A. Efros; see Zhu *et al.* (2017). Running horse photo by Justyna Furmanczyk Gibaszek/Shutterstock.

출처: Figure #28

19.7 Using Computer Vision (6/7)

Using vision for controlling movements

- A vision system for an automated vehicle driving
  - Lateral control: Ensure that the vehicle remains securely within its lane
  - Longitudinal control: Ensure that there is a safe distance to the vehicle in front.
  - Obstacle avoidance: Monitor vehicles in neighboring lanes and be prepared for evasive maneuvers if one of them decides to change lanes.
- Generate appropriate steering, acceleration, and braking actions to best accomplish these tasks.
- The more general case of mobile robots navigating in various indoor and outdoor environments has been studied, too.

19.7 Using Computer Vision (7/7)

Video Turing Test (VTT) Project at SNU

Learn

Understand

TV Sitcom Video "Friends" © Warner Brothers

Human Q	Who entered the café wearing a bridal dress?
Robot A	Rachel.
Human Q	Why did Rachel try to leave her fiancé?
Robot A	Because he betrayed her.
•••	

Summary

Although perception appears to be an effortless activity for humans, it requires a significant amount of sophisticated computation. The goal of vision is to extract information needed for tasks such as manipulation, navigation, and object recognition.

1. Representations of images capture edges, texture, optical flow, and regions. These yield cues to the boundaries of objects and to correspondence between images.
2. Convolutional neural networks produce accurate image classifiers that use learned features. Rather roughly, the features are patterns of patterns of patterns. . . . It is hard to predict when these classifiers will work well, because the test data may be unlike the training data in some important way.
3. Image classifiers can be turned into object detectors. One classifier scores boxes in an image for objectness; another then decides whether an object is in the box, and what object it is.
4. With more than one view of a scene, it is possible to recover the 3D structure of the scene and the relationship between views. In many cases, it is possible to recover 3D geometry from a single view.
5. The methods of computer vision are being very widely applied.

References

Figures

#6, 7, 8, 9, 10, 11, 12, 13.1, 22, 23, 24, 25, 26, 27, 28 Stuart J. Russell and Peter Norvig (2021). Artificial Intelligence: A Modern Approach (4rd Edition). Pearson

#13.2 https://www.kdnuggets.com/2018/05/data-augmentation-deep-learning-limited-data.html

#14 Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

#15 장병탁. 장교수의 딥러닝. 홍릉과학출판사, 2017.

#16 Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

#17 Delivering competitive advantage to companies creating and using vision and machine learning technology https://www.bdti.com/InsideDSP/2017/06/29/Microsoft

References

Figures

#1, 19 https://int.search.myway.com/search/AJimage.jhtml?&n=78674121&p2=%5EBSB%5Echr999%5ETTAB02%5E&pg=AJimage&pn=2&ptb=37D82E3B-48C7-47BB-A51A-347A9F03E0D2&qs=&searchfor=optical+character+recognition&si=&ss=sub&st=tab&tpr=sbt&trs=wtt&ots=1590294714543&imgs=1p&filter=on&imgDetail=true

#2, 20 https://int.search.myway.com/search/AJimage.jhtml?&enc=0&n=78674121&p2=%5EBSB%5Echr999%5ETTAB02%5E&pg=AJimage&pn=1&ptb=37D82E3B-48C7-47BB-A51A-347A9F03E0D2&qs=&searchfor=scene+understanding&si=&ss=sub&st=tab&tpr=sbt&trs=wtt&imgs=1p&filter=on&imgDetail=true

#3 https://int.search.myway.com/search/AJimage.jhtml?&enc=0&n=78674121&p2=%5EBSB%5Echr999%5ETTAB02%5E&pg=AJimage&pn=1&ptb=37D82E3B-48C7-47BB-A51A-347A9F03E0D2&qs=&searchfor=scene+understanding&si=&ss=sub&st=tab&tpr=sbt&trs=wtt&imgs=1p&filter=on&imgDetail=true

#4 https://int.search.myway.com/search/AJimage.jhtml?&n=78674121&p2=%5EBSB%5Echr999%5ETTAB02%5E&ptb=37D82E3B-48C7-47BB-A51A-347A9F03E0D2&qs=&si=&ss=sub&st=tab&trs=wtt&tpr=sbt&enc=2&searchfor=fSOrccErETqNBTaUBd9A8w99viweJWdZ2j_cF0xVQ8mf9yQNM4-5WtDD42-PRPE8ZN7z0zWqv_pYJBiS19egEEIEHKdBZrZWN_R6m9RflvbETdL6GWC7j0OnNZUiA-GoQyZCb_M2_JC4d2UEoZMiPaCqo1zHjdldaeMtLjygu7R0-nrV7i0-qOPGyWq3Kn0kMph69BAXLj3glCfbKkcHZl0lzE6t-IXvOLMBuyGCvpBNtYrWumcrDOrMCs9IJ8JrpNOlvd57zWUnzGQOerjrnb3G54QkKVT-fwQoS4GEGTHgc90CCTsUUa1okqy1got-rPVbeYrXYjHqzIJRNqxC9Q&ts=1590286927954&imgs=1p&filter=on&imgDetail=true

#5 https://int.search.myway.com/search/AJimage.jhtml?&n=78674121&p2=%5EBSB%5Echr999%5ETTAB02%5E&pg=AJimage&pn=6&ptb=37D82E3B-48C7-47BB-A51A-347A9F03E0D2&qs=&searchfor=car+detection&si=&ss=sub&st=tab&tpr=sbt&trs=wtt&ots=1590287197619&imgs=1p&filter=on&imgDetail=true