AI基础

Lecture 14: Computer Vision

Bin Yang

School of Data Science and Engineering

byang@dase.ecnu.edu.cn

[Some slides adapted from Philipp Koehn, JHU]

考试安排

- 时间：8:15到12:15，6月26日
- 地点：数学馆西113（平时周五的教室）
- 每人9分钟：4分钟报告，5分钟问答

- 8:15~9:45 1~10
- 9:55~11:25 11~20
- 11:30~12:15 21~25

序号	学号	姓名
1	10215300402	朱维清
2	10222140408	谷杰
3	10222140454	陈予瞳
4	10223903406	曹可心
5	10224507041	姚凯文
6	10224602413	朴祉燕
7	10225101419	贺云航
8	10225101440	韩晨旭
9	10225101447	唐硕
10	10225101469	朱陈媛
11	10225101483	谢瑞阳
12	10225101529	田亦海
13	10225101535	徐翔宇
14	10225101546	陈胤遒
15	10225102444	李文奇
16	10225102459	杨鸣谦
17	10225102463	李畅
18	10225102480	额尔琪
19	10225102491	张宇昂
20	10225102494	陈稷豪
21	10225102509	童言
22	10225501422	林童奕凡
23	10225501435	王雪飞
24	10225501447	姜嘉祺
25	10225501448	李度

Lecture 13 ILOs

- Feedforward neural networks
    - Universal approximation theorem
    - Nonlinear activation functions
    - Computation graphs
- Convolutional neural networks
    - Kernel, Receptive field
    - Pooling, down-sampling
- Recurrent neural networks
    - Back-propagation through time
    - Long-short term memory
- Learning algorithms
    - Stochastic gradient descent
    - Batch normalization
- Generalization
    - Network architecture, neural architecture search
    - Weight decay
    - Drop out
- Beyond supervised learning
    - Unsupervised learning, transfer learning, semi-supervised learning

Lecture 14 ILOs

- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Outline

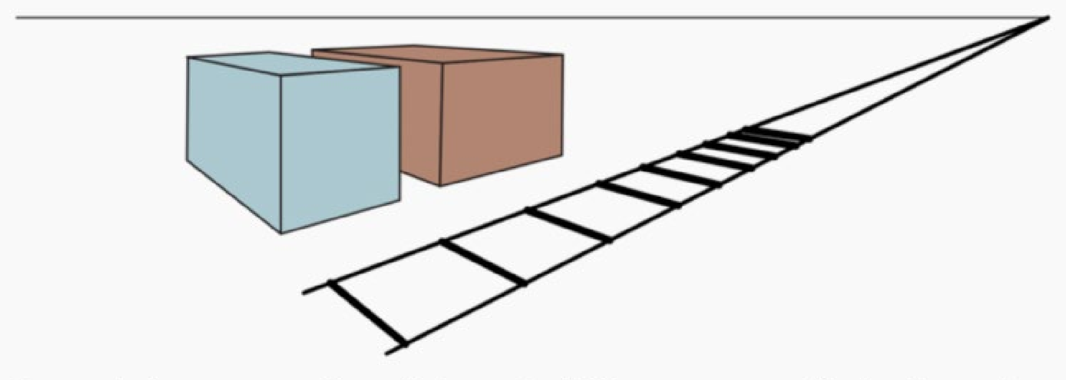
- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Vision

- Vision is a perceptual channel that accepts a stimulus and reports some representation of the world.
- Passive sensing: does not need to send out light to see.
- Active sensing: involves sending out a signal such as radar or ultrasound, and sensing a reflection.
  - Bats (ultrasound), dolphins (sound), abyssal fishes (light), and some robots (light, sound, radar).
- Computer vision
  - How to recover information from the data that comes from eyes or cameras
- The two core problems of computer vision are
  - Reconstruction: builds a model of the world from an image or a set of images
  - Recognition: draws distinctions among the objects it encounters based on visual and other information

Image Formation

- Imaging distorts the appearance of objects
- **Foreshortening：**
  - A picture taken looking down a long straight set of railway tracks will suggest that the rails converge and meet.
  - If you hold a book flat in front of your face and tilt it backward and forward, it will seem to shrink and grow in the image.

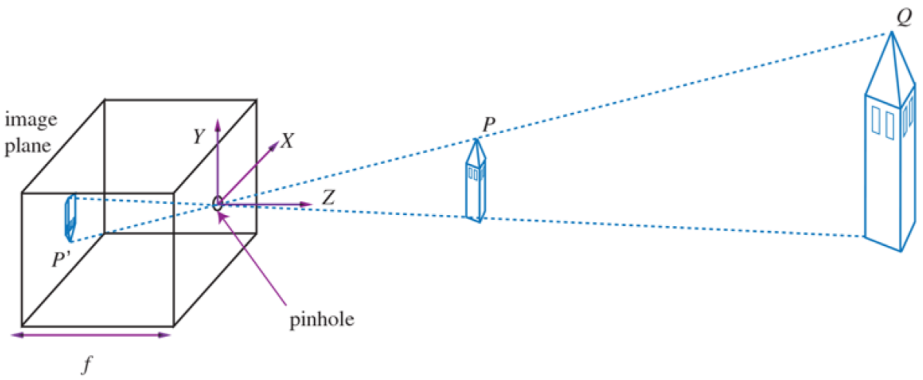


Geometry in the scene appears distorted in images. Parallel lines appear to meet, like the railway tracks in a desolate town. Buildings that have right angles in the real world scene have distorted angles in the image.

Images without Lenses: the pinhole camera

- The simplest way to form a focused image is to view stationary objects with a pinhole camera
- If the pinhole is small enough, the result is a focused image behind the pinhole.
- Pinhole cameras make it easy to understand the geometric model of camera behavior (which is more complicated—but similar—with most other imaging devices).

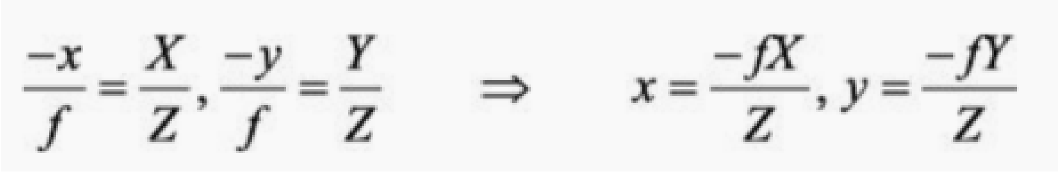
Figure 25.2



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.

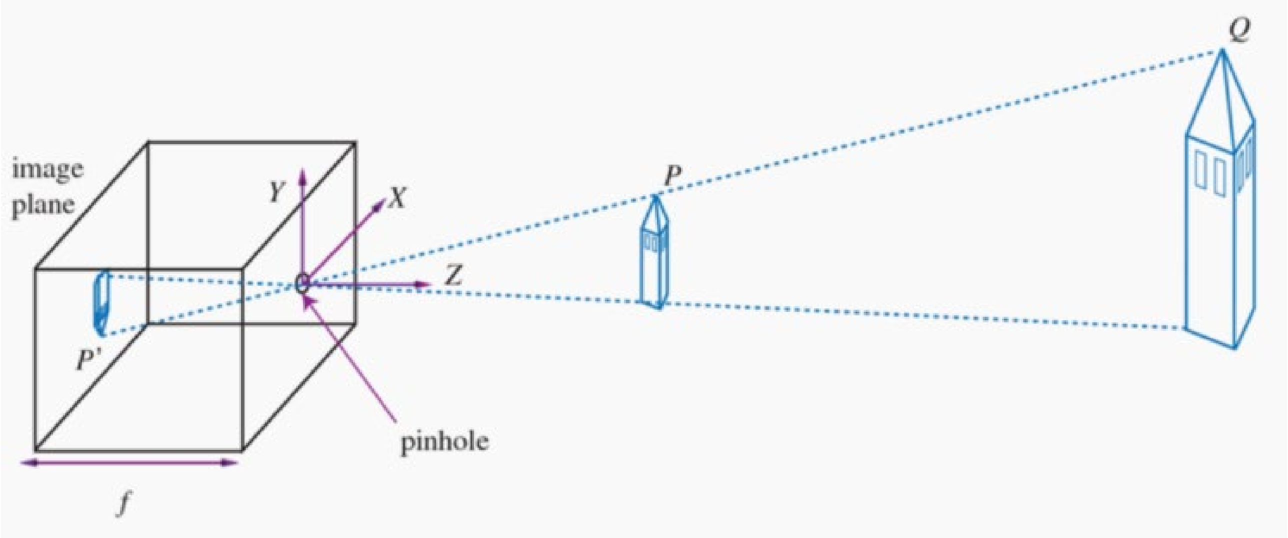
The pinhole camera

- Use a three-dimensional (3D) coordinate system with the origin at O, and consider a point in the scene, with coordinates (X, Y, Z), gets projected to the point in the image plane with coordinates (x, y, z).
- If f is the focal length: the distance from the pinhole to the image plane.
- By similar triangles, define the imaging process of perspective projection：



$$\frac{-x}{f}=\frac{X}{Z},\ \frac{-y}{f}=\frac{Y}{Z} \quad \Rightarrow \quad x=\frac{-fX}{Z},\ y=\frac{-fY}{Z}$$

- Z means that the farther away an object is, the smaller its image will be
- The minus signs mean that the image is inverted, both left–right and up–down, compared with the scene.



Lens systems

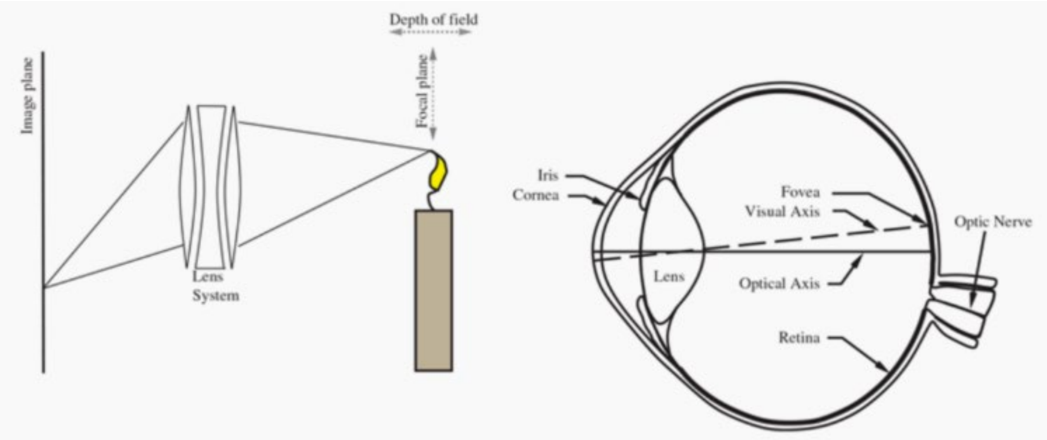
- The deficiency of pinhole camera: the pinhole is small, only a little light will get in, and the image will be dark.

- Improvement： Enlarging the hole (the aperture) will make the image brighter by collecting more light from a wider range of directions.
  - Deficiency： with a larger aperture the light that hits a particular point in the image plane will have come from multiple points in the real world scene, so the image will be defocused

- We need some way to refocus the image.

Lens systems

- **Lens systems**： 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.
- Depth of field and focusing：
  - The lens system focuses only the light in the focal plane within the lens depth range
  - Depth of field： the range of depths for which focus remains sharp enough
    - The larger the lens aperture (opening), the smaller the depth of field.
    - The camera adjusts the focal length by moving the lens element or changing the lens shape



Light and shading

- The brightness of a pixel in the image is a function of the brightness of the surface patch in the scene that projects to the pixel.
- Factors that affect brightness:
  - ambient light: overall light intensity in the scene
  - surface direction: the angle of the surface relative to the light source (whether the point is facing the light or is in shadow)
  - reflection: the amount of light reflected
    - diffuse reflection：Diffuse reflection scatters light evenly across the directions leaving a surface, so the brightness of a diffuse surface doesn’t depend on the viewing direction.（e.g.: paint, rough wood surface, etc）
    - specular reflection：Specular reflection causes incoming light to leave a surface in a lobe of directions that is determined by the direction the light arrived from.（e.g.: mirrors）

Light and shading

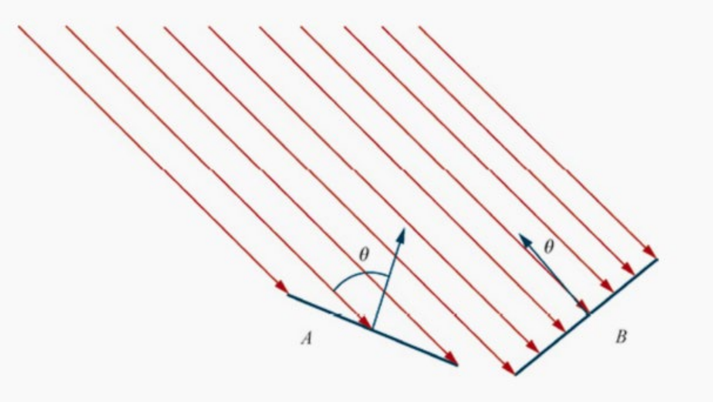
- The brightness of a diffuse patch:

$$I = \rho I_0 \cos \theta$$

I_0 the intensity of the light source

θ the angle between the light source direction and the surface normal

ρ the diffuse albedo



- A and B are two slices of the surface illuminated by a distant light source
- A is tilted away from the light source, receiving less light per unit surface area
- B facing the light source, brighter

Color

- Principle of trichromacy： Human vision relies on three types of colors red, green, and blue (RGB).

- For computer vision applications: for example cameras and screens use the RGB model, where each pixel is represented by three values (R, G, B).

- Color Constancy: The ability of the human visual system to perceive consistent colors under different lighting conditions.

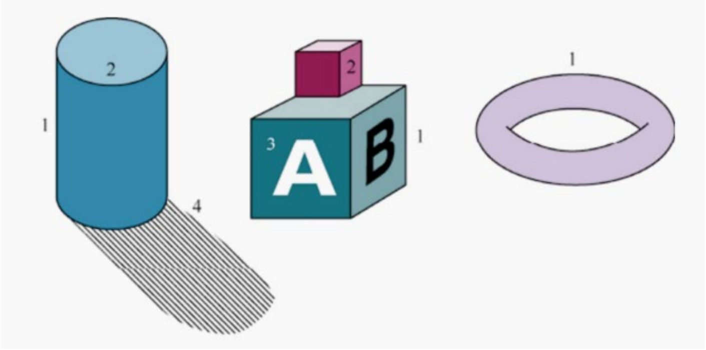
Outline

- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Simple Image Features

- Light reflects off objects in the scene to form an image consisting of, say, twelve million three-byte pixels.
  - There is a lot of data to deal with.
  - The way to get started analyzing this data is to produce simplified representations that expose what’s important, but reduce detail.
- Four properties: edges, texture, optical flow, and segmentation
  - Edges： Local manipulation of images, appears in the local pixel light intensity difference is large.
  - Texture, optical flow and segmentation：involves the processing of larger area images

Edges



Definition：A line or curve with a significant change in the brightness of the image

Different kinds of edges:

(1) Depth discontinuity: A sudden change in the depth of an object

(2) Surface orientation change: When the Angle of the surface changes with respect to the light source

(3) Reflectance change: Different materials reflect different reflectance

(4) Illumination change (shadow): Change in light intensity

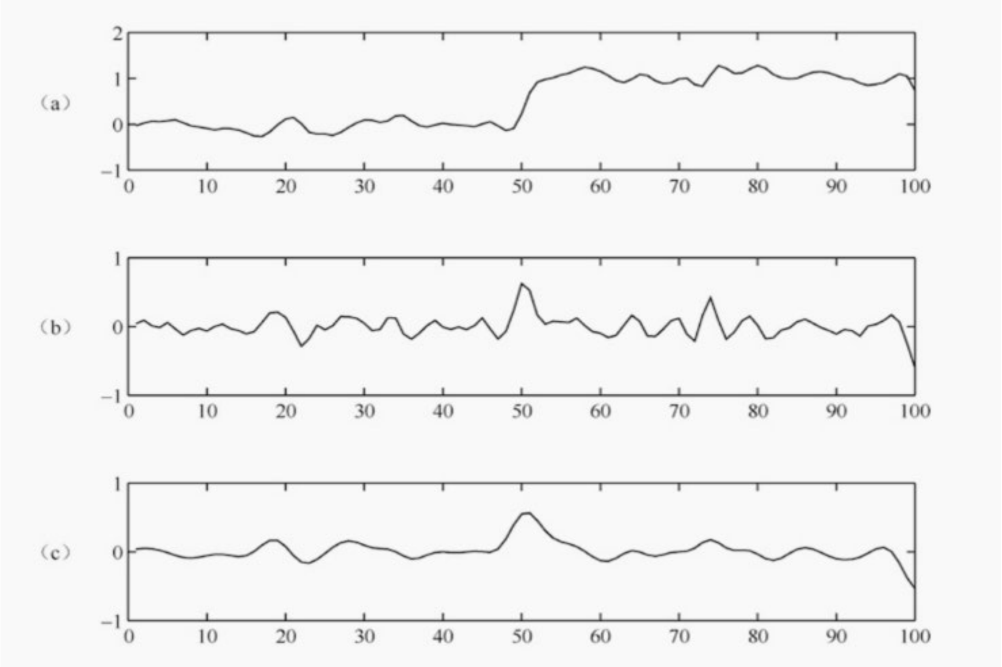
Edge detection method

Gradient-based method: Differentiating the image, finding the position with large derivative to judge the edge

Problem: There may be noise that is mistaken for an edge

Solution： Gaussian filter smooths the image to reduce noise

- Specifically, the weighted sum of nearby pixels is used as a prediction of the "true" value of a pixel, where the nearest pixel has the greatest weight.



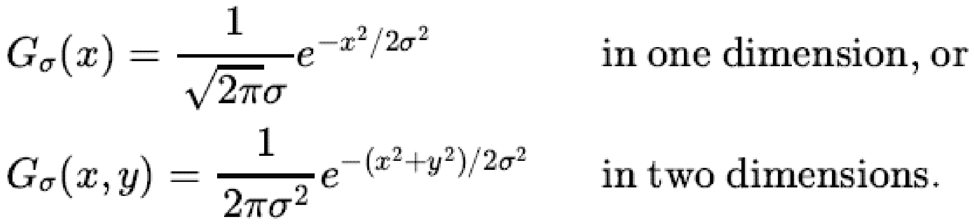
(a): Intensity profile I(x) along a one-dimensional section across a step edge.

(b): The derivative of intensity, I'(x). Large values of this function correspond to edges, but the function is noisy.

(c): The derivative of a smoothed version of the intensity. The noisy candidate edge at x = 75 has disappeared.

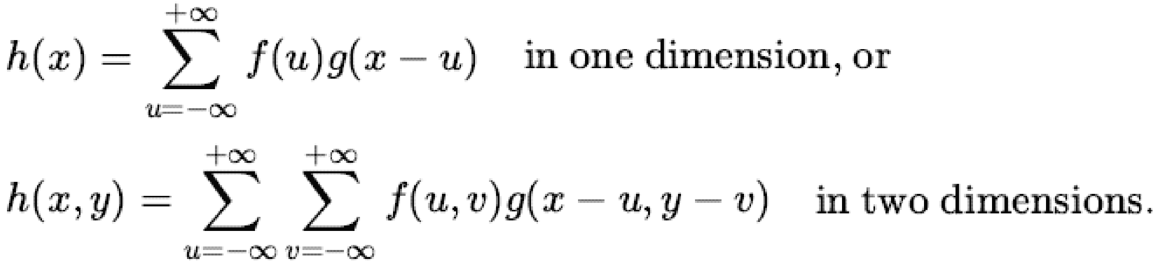
Gaussian Filter

- Smoothing involves using surrounding pixels to suppress noise.
- We will predict the "true" value of our pixel as a weighted sum of nearby pixels, with more weight for the closest pixels.
- A natural choice of weights is a Gaussian filter.



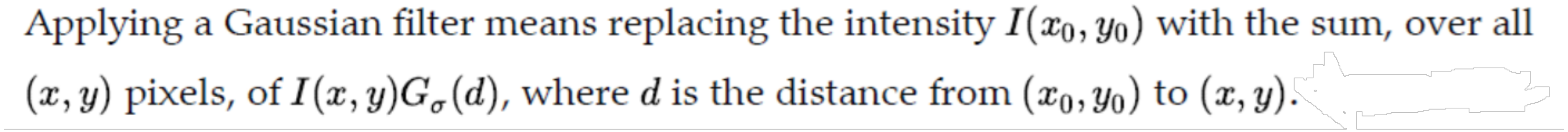
$$G_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \quad \text{in one dimension, or}$$

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad \text{in two dimensions.}$$



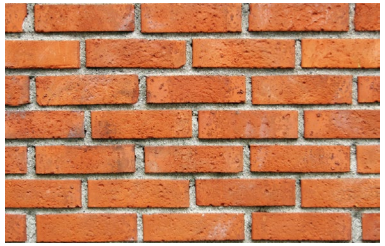
$$h(x) = \sum_{u=-\infty}^{+\infty} f(u)g(x-u) \quad \text{in one dimension, or}$$

$$h(x, y) = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} f(u, v)g(x-u, y-v) \quad \text{in two dimensions.}$$



Applying a Gaussian filter means replacing 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).

Texture



- A visual pattern on a surface characterized by repeating elements or statistical properties.
  - Sometimes the arrangement is quite periodic, as in the stitches on a sweater;
  - in other instances, such as pebbles on a beach, the regularity is only in a statistical sense: the density of pebbles is roughly the same on different parts of the beach.
- Textures are useful for two key tasks：
  - Object recognition: a zebra and horse have similar shape, but different textures
  - Matching: Matching slices in one image with slices in another image
- The basic construction method of texture representation
  - Given an image slice, calculate the gradient orientation of each pixel in the slice, and then use a histogram of the orientations to characterize the slice
- It is no longer usual to construct these descriptions by hand. Instead, convolutional neural networks are used to produce texture representations.

Optical flow

Definition: The apparent motion that occurs in an image when relative motion occurs between one or more objects in a scene

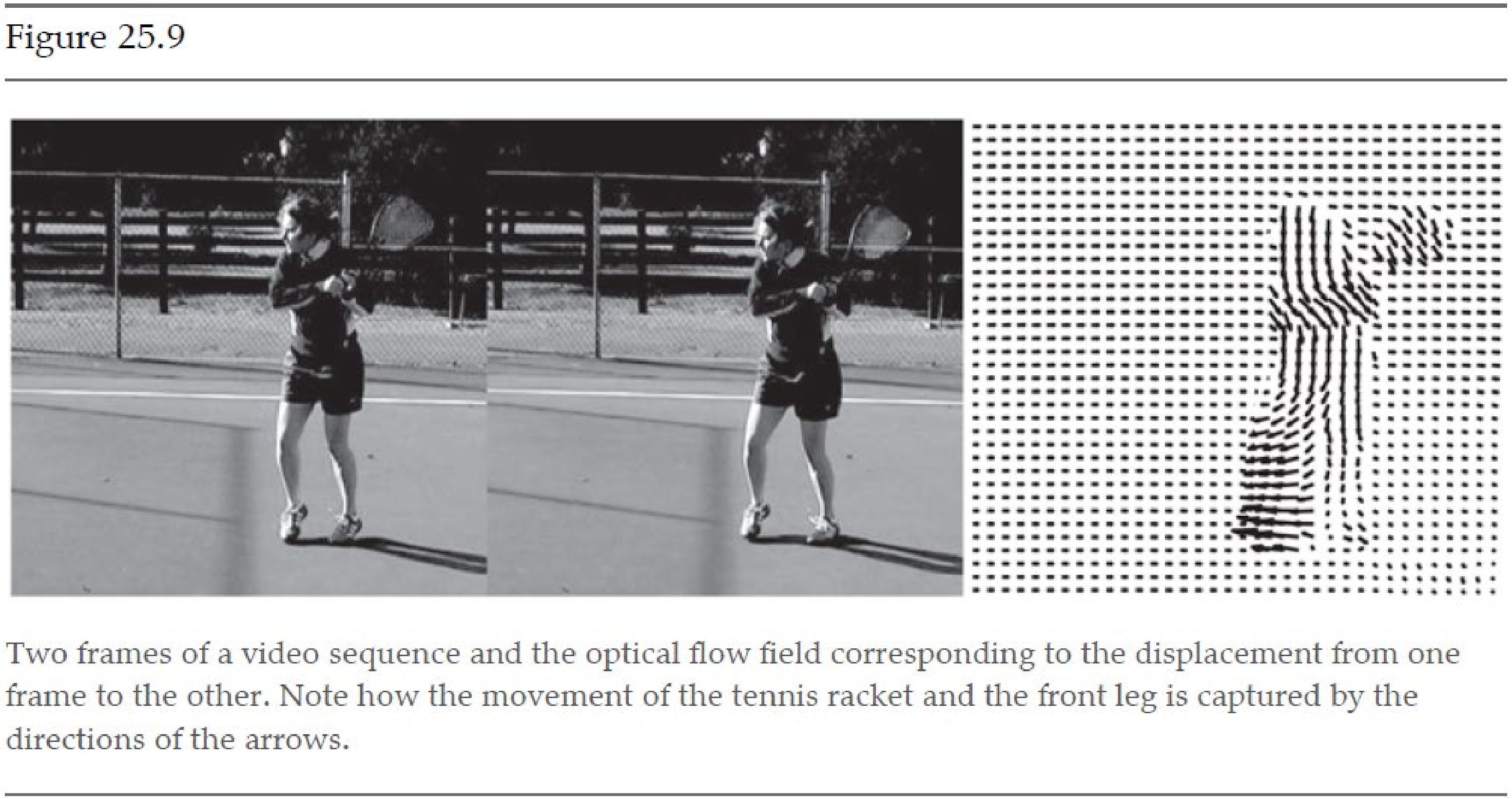


Figure 25.9

Two frames of a video sequence and the optical flow field corresponding to the displacement from one frame to the other. Note how the movement of the tennis racket and the front leg is captured by the directions of the arrows.

Optical Flow

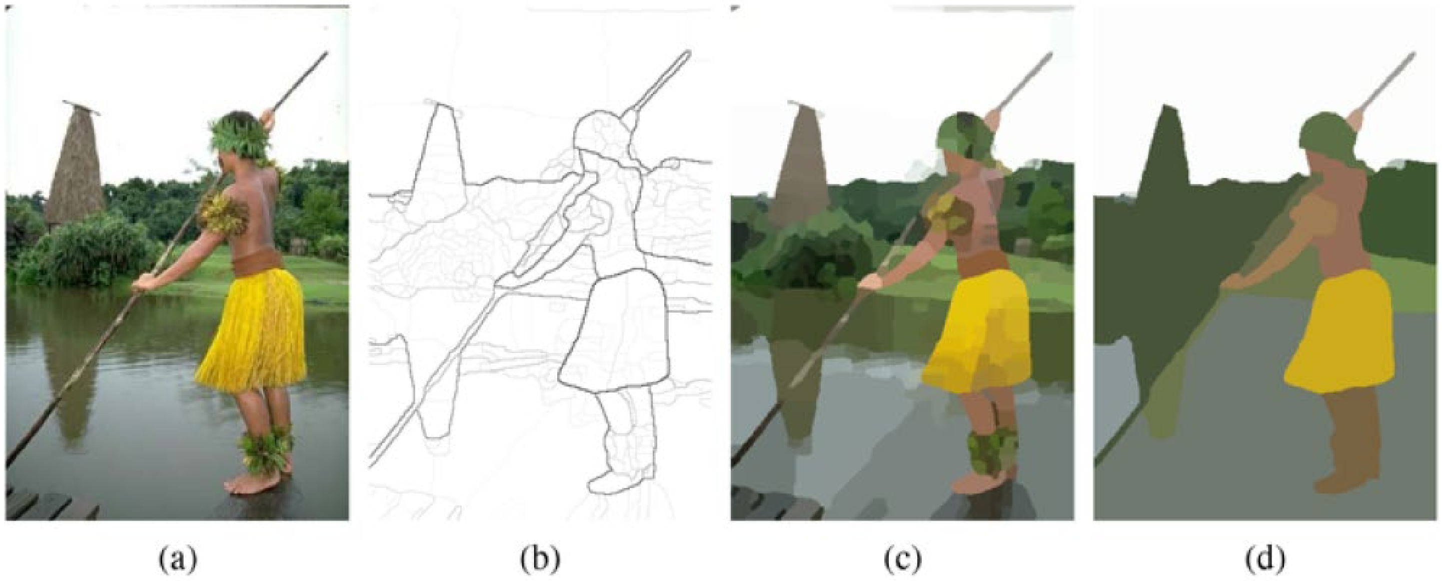
- **Applications**：
  - Motion detection: Identify moving objects in a video.
  - Object tracking: Tracks the movement of objects across frames.
  - Video stabilization: Smooths camera movement in a video sequence.
- Computing:
  - sum of squared differences (SSD): Measure the similarity between pixel blocks

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

  - (x, y) represents the pixel position in the pixel block centered on (x_0, y_0)
  - Find (Dx, Dy) that minimizes the SSD, then optical flow (v_x, v_y) at (x_0, y_0)=(Dx/Dt, Dy/Dt)

Image Segmentation

- Definition: Decomposing an image into several sets of similar pixels.



(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.

Segmentation Techniques

Edge-based segmentation: classification

- The boundary curve at pixel position (x, y) will have direction θ
- The image neighborhood centered at (x, y) will look like a disk and will be segmented into two halves along the diameter direction θ
- We can compute the probability $P_b(x, y, \theta)$ that there is a boundary curve at that pixel along that orientation by comparing features in the two halves.
- The natural way to predict this probability is to train a machine learning classifier using a data set of natural images in which humans have marked the ground truth boundaries—the goal of the classifier is to mark exactly those boundaries marked by humans and no others.

Region-based segmentation: clustering

- Clustering: Pixels are grouped into regions based on brightness, color, texture properties.
- Segmentation: The segmentation is performed by minimizing the sum of weights between adjacent pixels across groups and maximizing the sum of weights within groups.

Outline

- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Image Classification

Two main cases of image classification:

1、The images are of objects

The object is from a given classification category, background information is irrelevant.

Example: In an image of a particular clothing or furniture item, the classifier outputs "cashmere sweater" or "desk chair."

2、Scene with multiple objects:

The image contains multiple objects.

Example: In an image of a savanna, there are both giraffes and lions.

Challenges of Image Classification

Appearance variations of different instances.

Color and texture variations: e.g., cats of different colors and textures.

Viewpoint variations: e.g., donuts from different angles.

Occlusion: Partial occlusion of objects, such as a cup handle disappearing.

Deformation: Changes in the shape of the object, such as an athlete's actions.

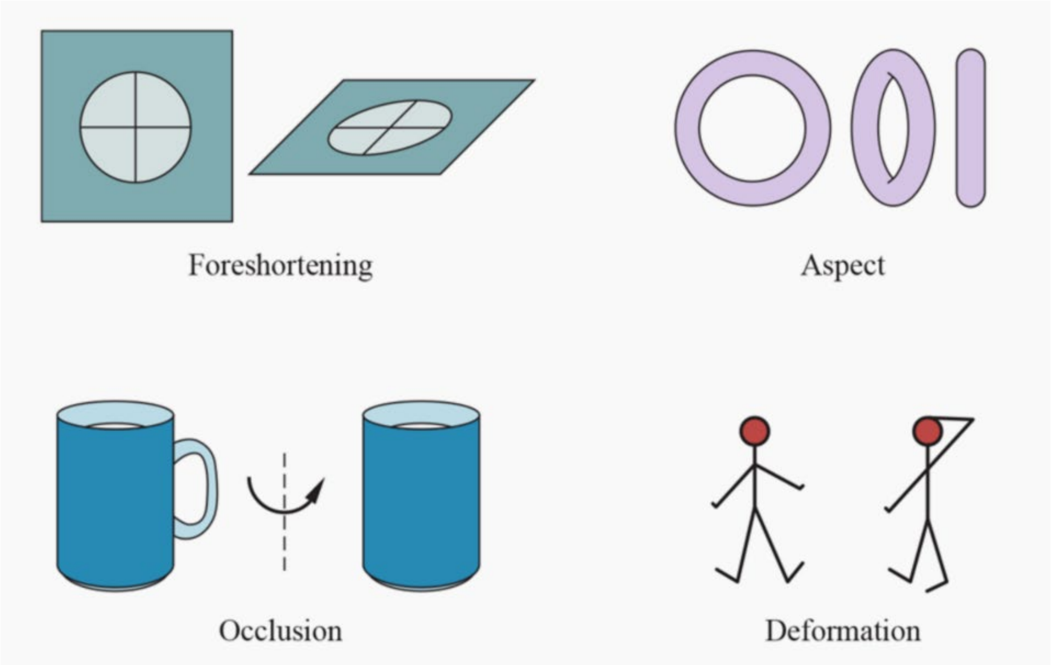
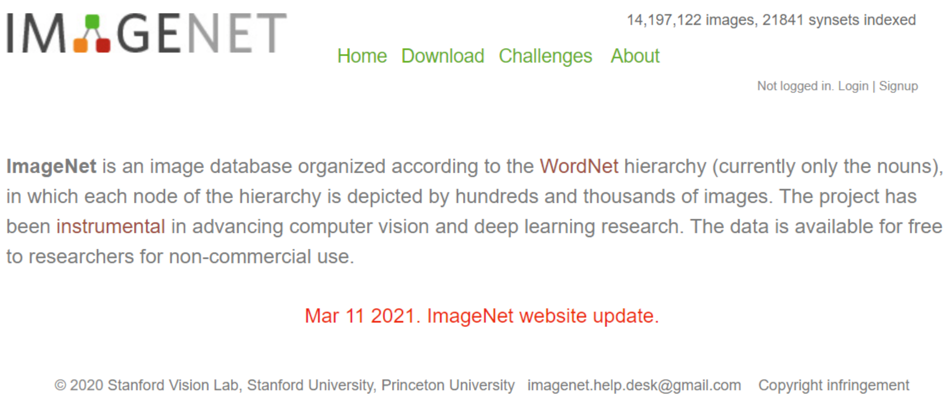


Image Classification with Convolutional Neural Networks

- Convolutional neural networks (CNNs) are spectacularly successful image classifiers.
- The ImageNet data set played a historic role in the development of image classification systems by providing them with over 14 million training images, classified into over 30,000 fine-grained categories. ImageNet also spurred progress with an annual competition.



IMAGENET

14,197,122 images, 21841 synsets indexed

Home Download Challenges About

Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. The project has been instrumental in advancing computer vision and deep learning research. The data is available for free to researchers for non-commercial use.

Mar 11 2021. ImageNet website update.

© 2020 Stanford Vision Lab, Stanford University, Princeton University imagenet.help.desk@gmail.com Copyright infringement

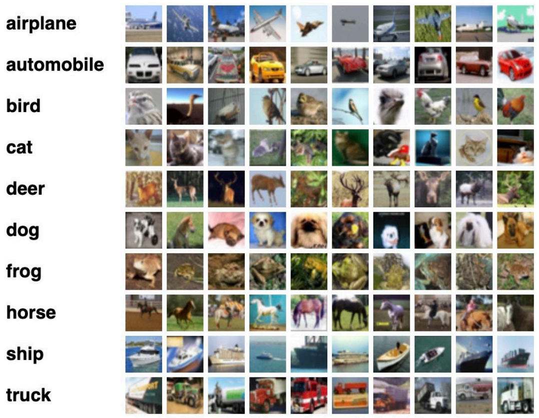


Image Classification Based on Convolutional Neural Networks

ImageNet dataset: A large-scale visual database

Content：

Number of images: Over 14 million images.

Number of categories: More than 20,000 categories.

Annotation information: Each image has a corresponding category label, some images also have bounding box annotations for object detection tasks.

Hierarchical structure: There are semantic relationships between categories, such as "animal" containing sub-categories "dog" and "cat.“

Evaluation criteria:

In image classification tasks, focus mainly on Top-1 accuracy and Top-5 accuracy. In object detection tasks, focus on mean Average Precision (mAP).

Systems are evaluated by both the classification accuracy of their single best guess and by top-5 accuracy, in which systems are allowed to submit five guesses—for example, *malamute, husky, akita, samoyed, eskimo dog*. ImageNet has 189 subcategories of *dog*, so even dog-loving humans find it hard to label images correctly with a single guess.

In the first ImageNet competition in 2010, systems could do no better than 70% top-5 accuracy. The introduction of convolutional neural networks in 2012 and their subsequent refinement led to an accuracy of 98% in top-5 (surpassing human performance) and 87% in top-1 accuracy by 2019. The primary reason for this success seems to be that the features that are being used by CNN classifiers are learned from data, not hand-crafted by a researcher; this ensures that the features are actually useful for classification.

Image Classification Based on Convolutional Neural Networks

Convolutional neural networks (CNNs) are spectacularly successful image classifiers.

Structure：

Convolutional layer: Uses multiple filters to perform convolution operations on the input image to extract local features.

Activation function: Introduces non-linear features. Common activation function ReLU, solves the vanishing gradient problem.

Pooling layer: Reduces the computational load and parameter count while retaining key features.

Fully connected layer: Similar to traditional neural networks, each neuron is connected to all neurons in the previous layer.

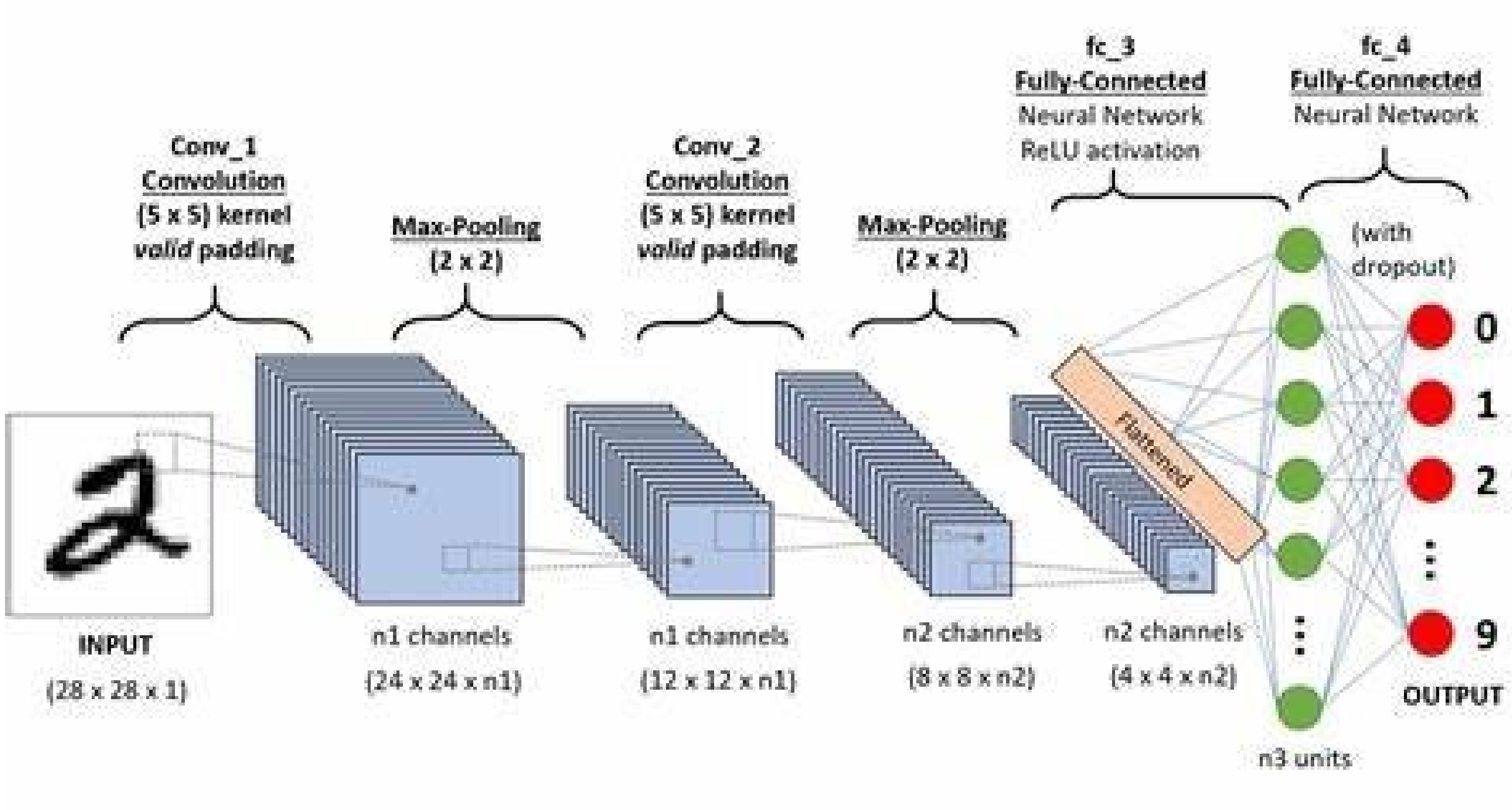
Advantages:

Local connections and weight sharing: Reduces the number of parameters, lowers computational complexity.

Step-by-step extraction of image features from low to high levels, which helps in recognizing complex patterns.

Classic models based on CNN:

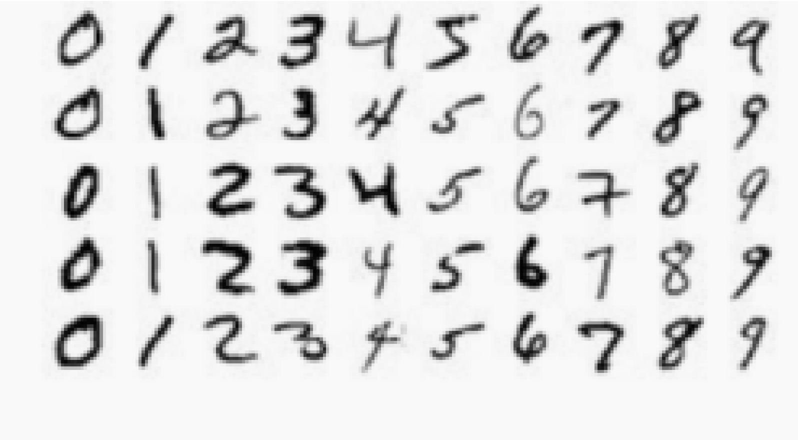
- LeNet: Used for handwritten digit recognition.
- AlexNet: First used deep convolutional networks in the ImageNet competition, significantly improving classification performance.
- VGG: Uses smaller convolutional kernels, but the network structure is deeper.


fc_3
Fully-Connected
Neural Network
ReLU activation
fc_4
Fully-Connected
Neural Network
Conv_1
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
Conv_2
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
(with
dropout)
Flattened
0
1
2
⋮
9
INPUT
(28 x 28 x 1)
n1 channels
(24 x 24 x n1)
n1 channels
(12 x 12 x n1)
n2 channels
(8 x 8 x n2)
n2 channels
(4 x 4 x n2)
OUTPUT
n3 units

Why Convolutional Neural Networks Are Effective for Image Classification

MNIST dataset:

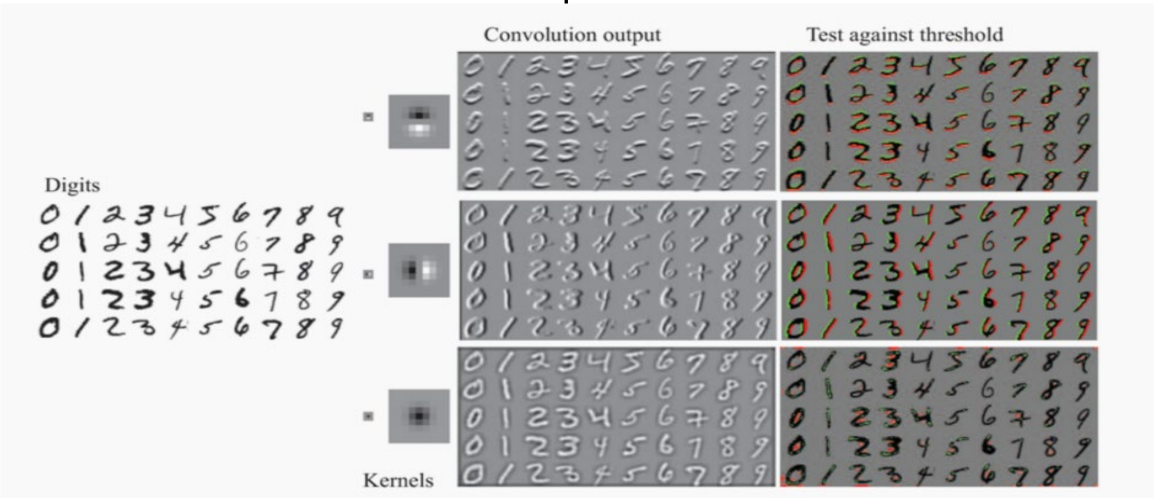
- Number of images: 70,000 grayscale handwritten digit images.
- Training set: 60,000 images.
- Test set: 10,000 images.
- Image size: Each image is 28x28 pixels, single channel.
- Labels: Each image corresponds to a label representing a handwritten digit from 0 to 9, with a total of 10 categories.



Why Convolutional Neural Networks Are Effective for Image Classification

1、Local patterns in images can provide a lot of information, and convolutional neural networks are good at capturing local information.

- A convolution with a ReLU activation function composite can be seen as a local pattern detector.
- The convolution measures the similarity between each local window of the image and the kernel pattern.
- The ReLU activation function sets low-score windows to zero and highlights high-score windows.
- Convolution with multiple filters can find various patterns.



Why Convolutional Neural Networks Are Effective for Image Classification

2、Multi-layer pattern detectors

- **The first layer**: detects local features
  - Convolution with multiple kernels finds multiple patterns
- **The second layer**: will be influenced by pixels within a larger window compared to the first layer
  - Composite patterns can be detected by applying another layer to the output of the first layer.
  - “patterns of patterns”

Each layer extracts different levels of information, which is conducive to image classification

Outline

- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Object Detection

- **Definition:** An object detector finds multiple objects in an image, determines the category of each object, and reflects its position by adding a **bounding box** around it.

- **Specifically:** build an object detector by looking at a small sliding window onto the larger image—a rectangle. At each spot, we classify what we see in the window, using a CNN classifier. We then take the high-scoring classifications—a cat over here and a dog over there—and ignore the other windows. After some work resolving conflicts, we have a final set of objects with their locations.

- **Applications:** Security surveillance, autonomous driving, medical image analysis, etc.

Application of Convolutional Neural Networks in Object Detection

Sliding window：

- Apply a window to the upper left corner of the image
- Slide the window step by step, covering the entire image

Based on CNN：

- Use the contents within each sliding window as input, pass it to the convolutional neural network
- CNN extracts features and classifies the window contents, outputting the results
- Determine whether the target object is contained in the window based on the classification result

Region Proposal Network (RPN)

- Searching all possible windows isn’t efficient—in an n by n pixel image there are O(n^4) possible rectangular windows.
- It makes sense to have a mechanism that scores “objectness”—whether a box has an object in it, independent of what that object is.
  - Classify the object for just those boxes that pass the objectness test.
- Region Proposal Network (RPN): A network that finds regions with objects
  - Consider several possible boxes (anchor boxes) for each center point and determine regions of interest (ROI) through object detection scores.
  - Use anchor boxes of various sizes and aspect ratios to cover objects of different sizes and shapes.

Which window to report

Conflict issue: Sliding windows may overlap, causing the same object to be detected multiple times.

Solution: Non-maximum suppression (NMS)

- Sort all windows' detection scores
- Select the window with the highest score, delete other overlapping windows
- Repeat the above process until all windows are processed

Bounding Box Regression

Definition: Predicts the improvement of trimming the window to the appropriate bounding box based on features calculated by the classifier

Specific process: High-score window - non-maximum suppression - predict the exact location.

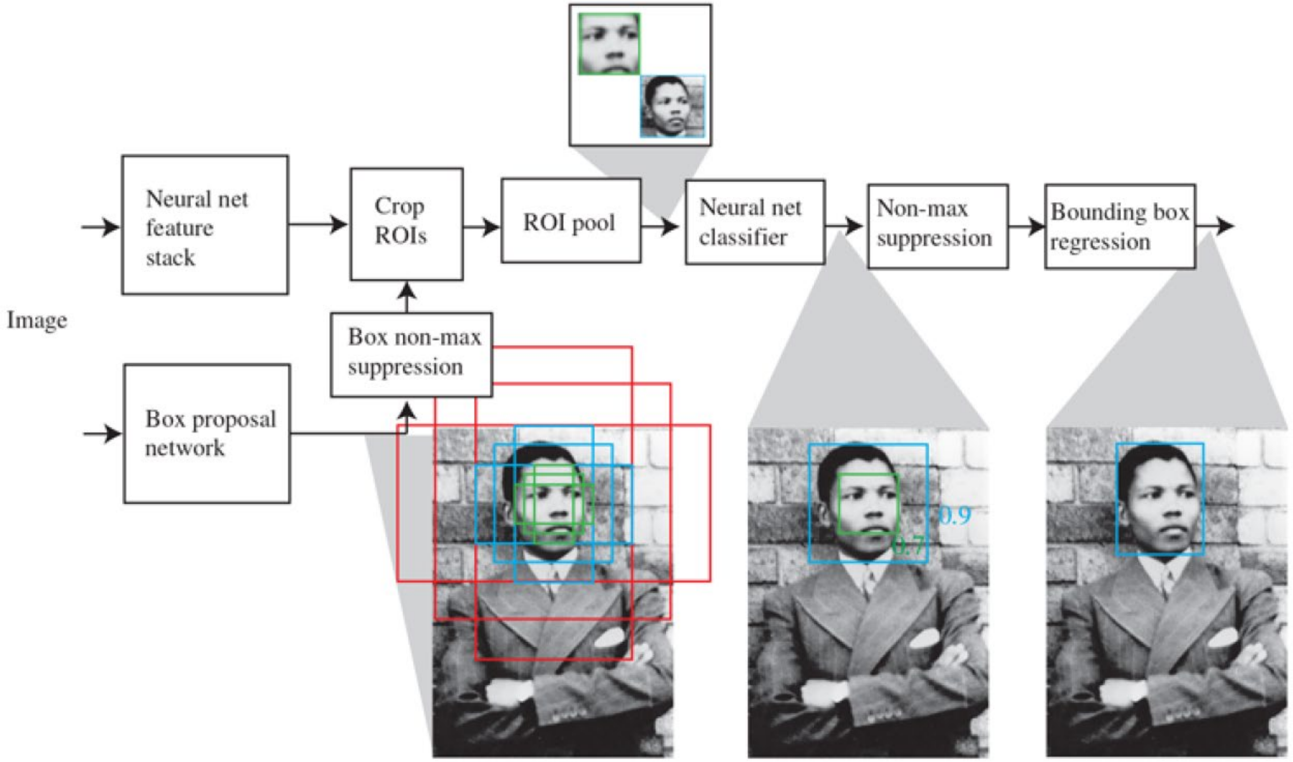
Advantage: Improves the accuracy of bounding box positioning, accurately reflects the position and size of the object.

Evaluating object detectors

- First, we need a test set: a collection of images with each object in the image marked by a ground truth category label and bounding box.
    - Usually, the boxes and labels are supplied by humans.
- Then we feed each image to the object detector and compare its output to the ground truth.
    - We should be willing to accept boxes that are off by a few pixels, because the ground truth boxes won’t be perfect.
- The evaluation score should balance recall (finding all the objects that are there) and precision (not finding objects that are not there).

Fast RCNN

Figure 25.13



Faster RCNN uses two networks. A picture of a young Nelson Mandela is fed into the object detector. One network computes "objectness" scores of candidate image boxes, called "anchor boxes," centered at a grid point. There are nine anchor boxes (three scales, three aspect ratios) at each grid point. For the example image, an inner green box and an outer blue box have passed the objectness test. The second network is a feature stack that computes a representation of the image suitable for classification. The boxes with highest objectness score are cut from the feature map, standardized in size with ROI pooling, and passed to a classifier. The blue box has a higher score than the green box and overlaps it, so the green box is rejected by non-maximum suppression. Finally, bounding box regression the blue box so that it fits the face. This means that the relatively coarse sampling of locations, scales, and aspect ratios does not weaken accuracy.

Outline

- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Multiple Views and 3D Clues

- Images show a 2D picture of a 3D world. But this 2D picture is rich with cues about the 3D world.
  - One kind of cue occurs when we have multiple pictures of the same world, and can match points between pictures.
- Advantages of multiple views:
  - Two images taken from different angles can provide more geometric information.
  - By calculating point correspondences between views, a 3D model can be reconstructed.
- Two Cases of 3D Model Reconstruction
  - If the camera parameters are known, a 3D model can be directly constructed.
  - If the camera parameters are unknown, a 3D model can still be constructed through point correspondences
- The key problem is to establish which point in the first view corresponds to which in the second view.

Multiple Views and 3D Clues

Establishing point correspondences:

- Use simple texture features for point matching
- Use geometric constraints and surface smoothness for matching

Methods of obtaining multiple views

- Use two cameras or binocular views
- Move the camera to obtain different perspectives

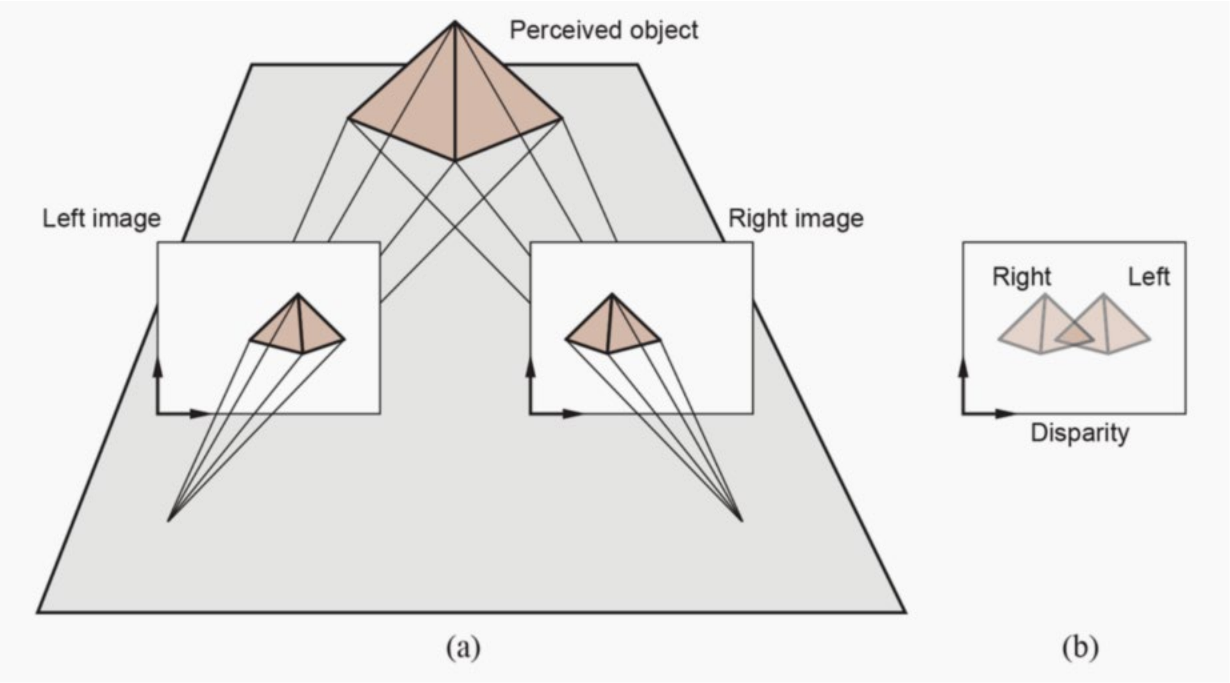
Binocular Stereo Vision

Disparity: The position change from the left view to the right view

Disparity measurement: Use the sum of squared differences to match the left pixel block with the right pixel block

（a）The difference in position is an indication of the depth of the object

（b）Superimpose the left and right images, and parallax will be observed



3D Clues from Moving Cameras

- Assume we have a camera moving in a scene.
- Label the left image “Time t” and the right image “Time t+1.”
- The geometry has not changed, so all the material from the discussion of stereopsis also applies when a camera moves.
- What we called disparity in that section is now thought of as apparent motion in the image, and called optical flow.

3D Clues from a Single View

Occlusion

- When one object partially occludes another, the occluded object is usually perceived as farther away
- Occlusion provides depth relationships between objects

Texture

- Due to perspective effects, texture elements appear smaller and denser in the distance
- for example, pebbles on a beach—it may not be uniform in image—the farther pebbles appear smaller than the nearer pebbles

Perspective Foreshortening

- Patterns on a plane extending into the distance appear to shrink in the image.

Texture Gradient

- Nearby texture elements are larger and sparser, while distant texture elements are smaller and denser.
- By analyzing texture gradients, the 3D shape and relative distance of object surfaces can be inferred.

3D Clues from a Single View

Shading

- Shading is determined by the scene geometry and surface reflection properties
- Shading variations allow us to perceive the 3D shape of objects.

Familiar Objects

- Known shapes and sizes of objects can help determine their position and orientation in the image

Spatial Relationships between Objects

- Example: The size and relative positions of pedestrians in an image can infer their distance from the camera.

Perspective Relationships

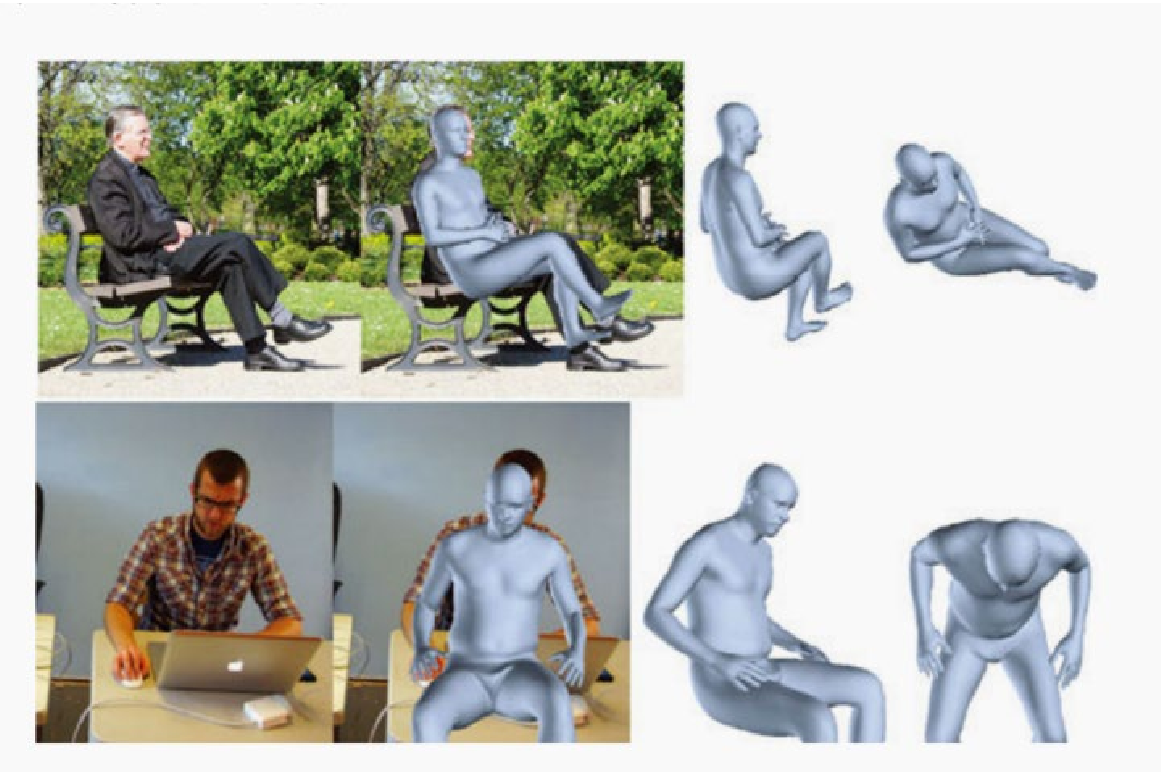
- Objects in the scene are arranged according to perspective rules, with distant objects appearing smaller and higher

Outline

- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision

Applications of Computer Vision in Human Behavior Understanding

Understanding human behavior through video analysis, establishing human-computer interaction programs, and observing and reacting to human behavior.



Example: Describing how to predict body joint positions through images and videos.

Both rows of images show 3D body shape reconstruction based on a single image.

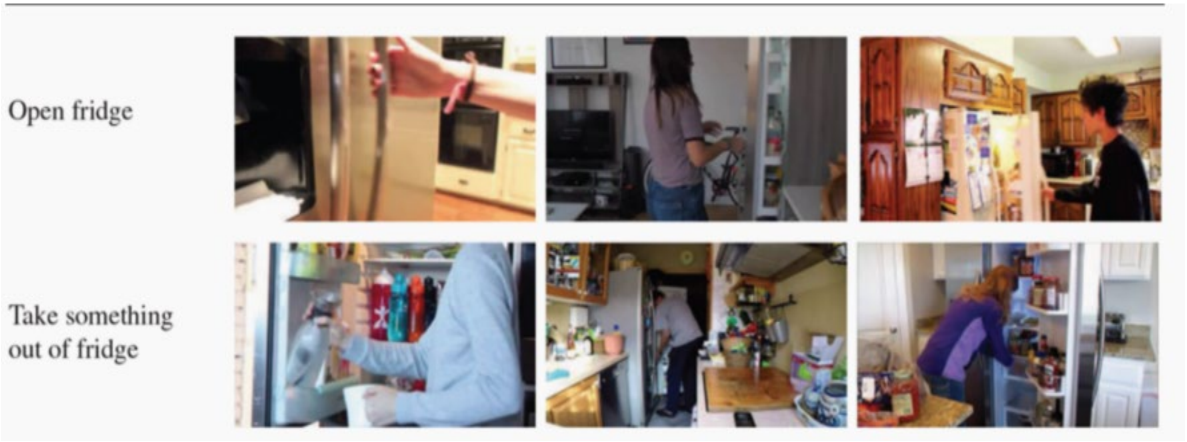
Challenges of Behavior Classification

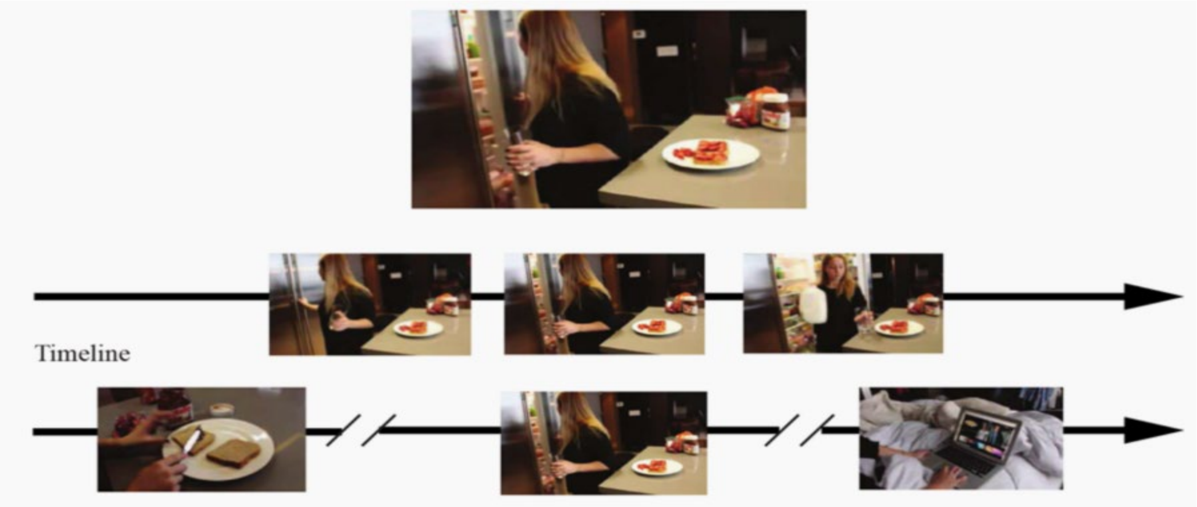
Complexity of behavior classification

Left image: Visual comparison of similar and different behaviors,

Impact of time scale on behavior understanding

Right image: Behavior classification at different time scales





Why is it difficult?

- Learned classifiers are guaranteed to behave well only if the training and test data come from the same distribution.
- We have no way of checking that this constraint applies to images, but empirically we observe that image classifiers and object detectors work very well.
- But for activity data, the relationship between training and test data is more untrustworthy because people do so many things in so many contexts. For example, suppose we have a pedestrian detector that performs well on a large data set. There will be rare phenomena (for example, people mounting unicycles) that do not appear in the training set, so we can’t say for sure how the detector will work in such cases.
- The challenge is to prove that the detector is safe whatever pedestrians do, which is difficult for current theories of learning.

Linking Pictures with Words

A baby eating a piece of food in his mouth

A young boy eating a piece of cake

A small bird is perched on a branch

A small brown bear is sitting in the grass

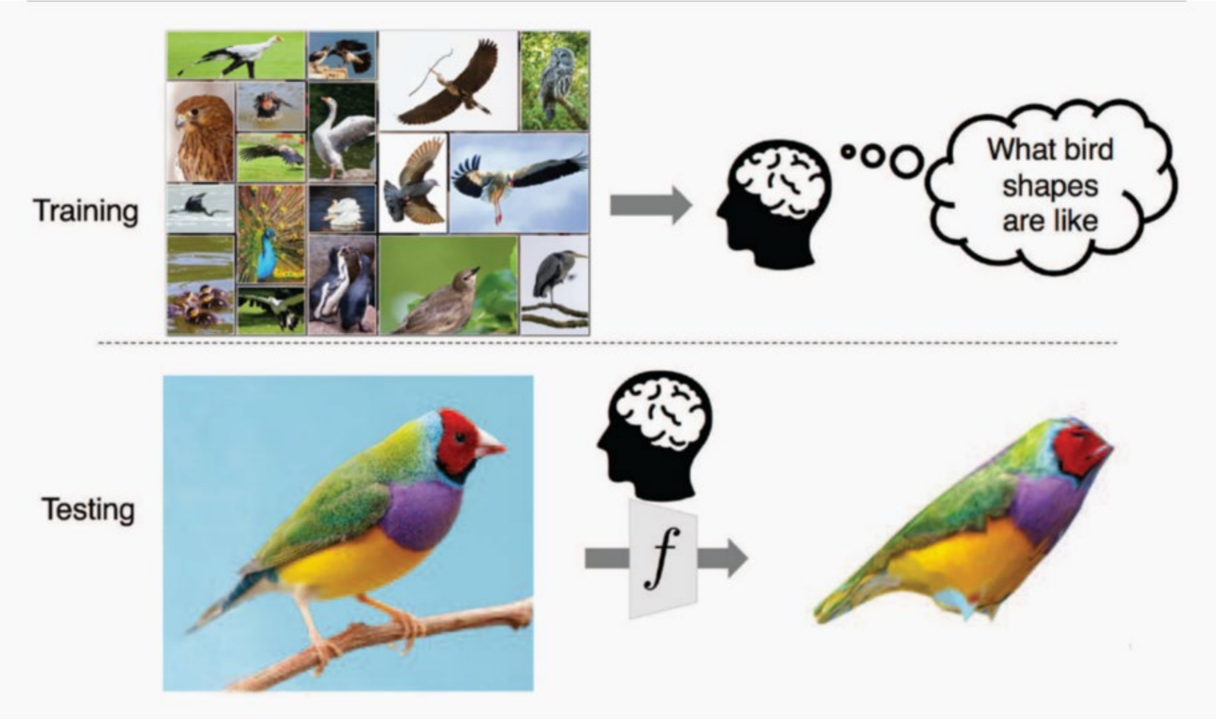
- Tagging system: tag images with related words
  - Apply image classification and object detection methods and tag the image with the output words.
  - Limitation: It might label "cat," "street," "trash can," "fishbone," but cannot describe "a cat pulling out a fishbone from the trash can.“
- Captioning system: Generates natural language sentences to describe image content.
  - couple a convolutional network (to represent the image) to a recurrent neural network or transformer network (to generate sentences), and train the result with a data set of captioned images.
  - The COCO (Common Objects in Context) data set is a comprehensive collection of over 200,000 images labeled with five captions per image.

Reconstruction from many views

- Reconstructing a set of points from many views
    - which could come from video or from an aggregation of tourist photographs
    - is similar to reconstructing the points from two views, but there are some important differences.
- More work to be done to establish correspondence between points in different views.
- But more views mean more constraints on the reconstruction and on the recovered viewing parameters, so usually more accurate.
- Reconstruction proceeds by
    - matching points over pairs of images, extending these matches to groups of images, coming up with a rough solution for both geometry and viewing parameters, then polishing that solution.
    - Polishing means minimizing the error between points predicted by the model (of geometry and viewing parameters) and the locations of image features.
- The detailed procedures are too complex to cover fully, but are now very well understood and quite reliable.

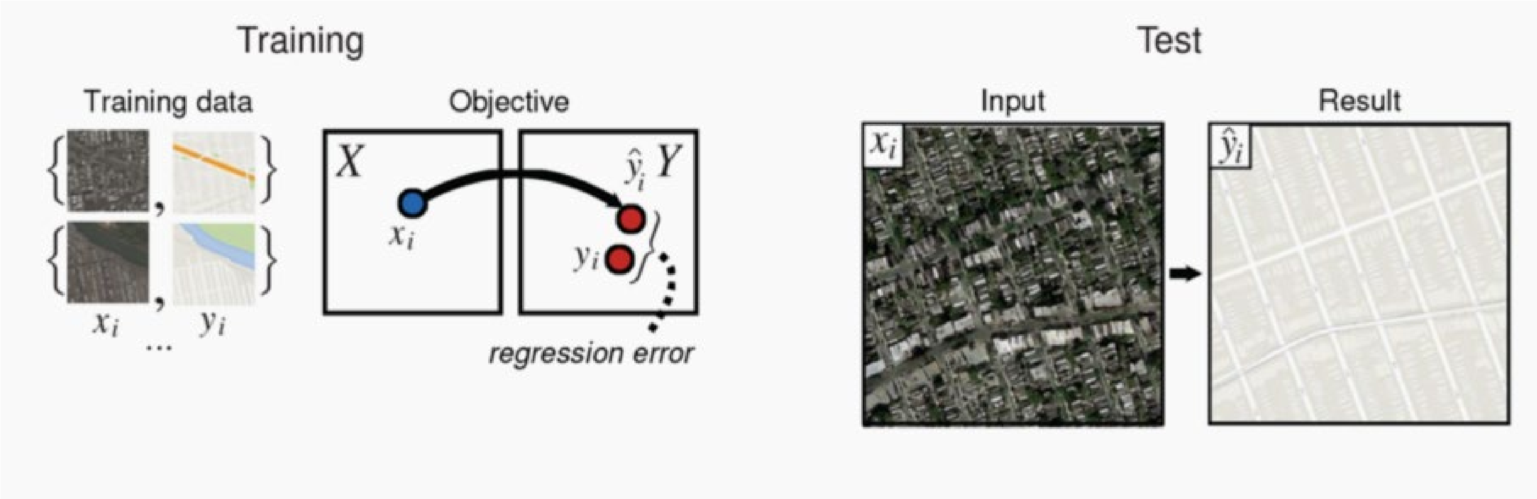
Geometry from a single view

Geometric representation: Predict a depth map from a single image.
Question: How to reconstruct the sparrow's pose and estimate its geometric model from this image?



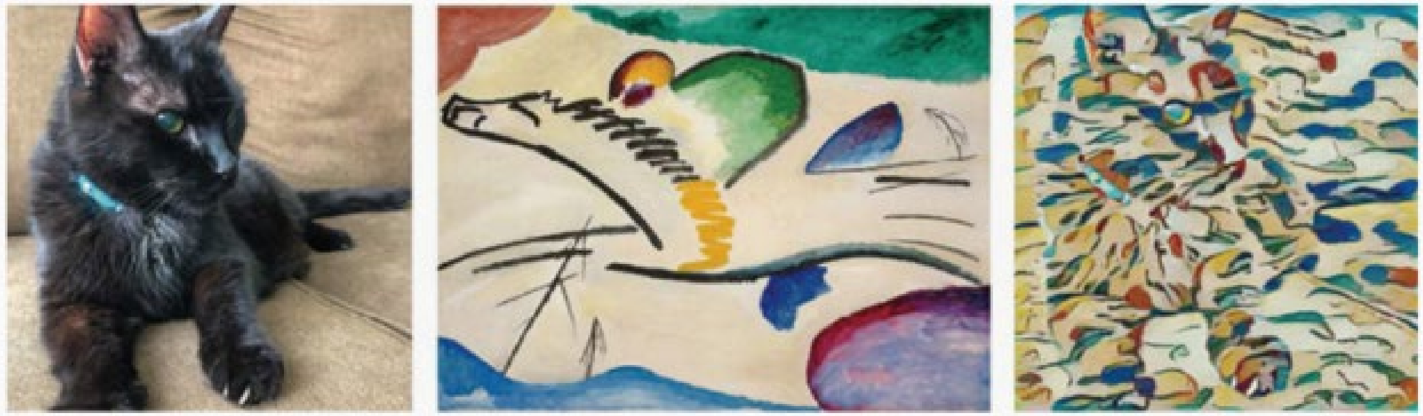
Making Pictures

Image transformation: Train neural networks to map X-type images (e.g., blurry images, aerial images of towns, or hand-drawn sketches of new products) to Y-type images (e.g., deblurred images, city road maps, or a product image).



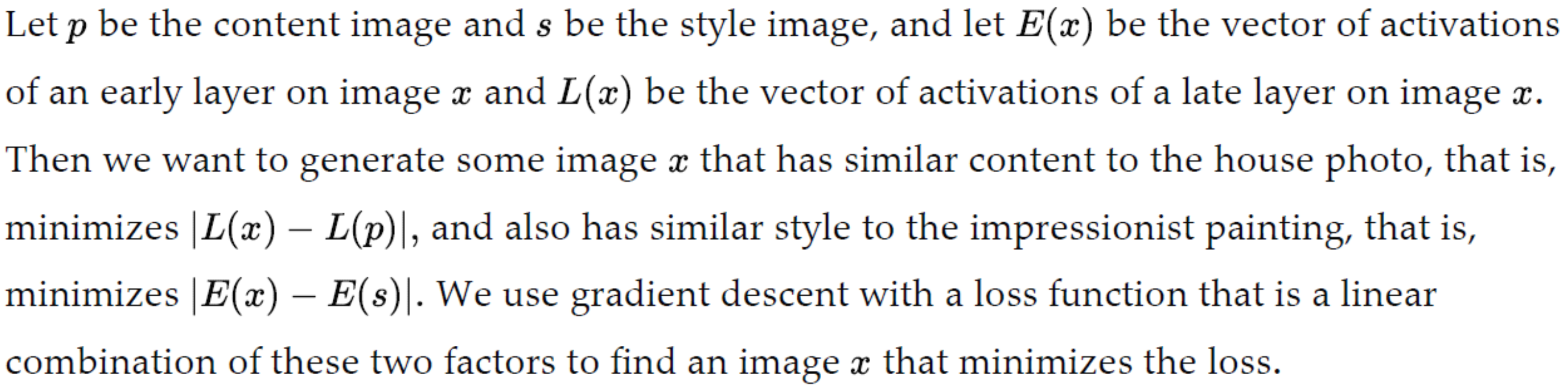
Conversion of pairs of images where the input consists of an aerial image and a corresponding road map. The goal is to train a network that generates road maps from aerial images

Style transfer



Style transfer

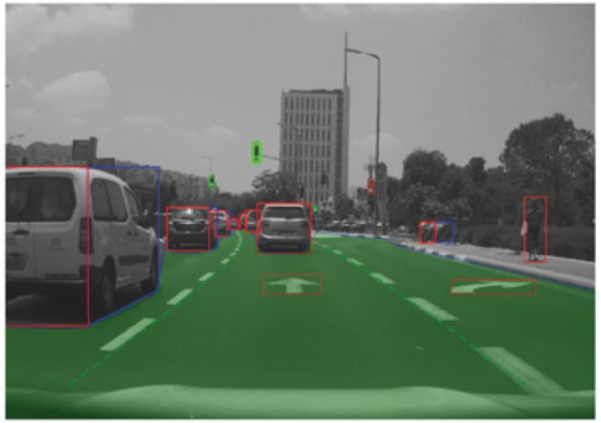
- Input consists of two images: content (e.g., a photo of a cat) and style (e.g., an abstract painting).
- The output is an image of the cat rendered in the abstract style.
- CNN: early layers tend to represent the style of a picture, and the late layers represent the content.



Let p be the content image and s be the style image, and let $E(x)$ be the vector of activations of an early layer on image x and $L(x)$ be the vector of activations of a late layer on image x. Then we want to generate some image x that has similar content to the house photo, that is, minimizes $|L(x) - L(p)|$, and also has similar style to the impressionist painting, that is, minimizes $|E(x) - E(s)|$. We use gradient descent with a loss function that is a linear combination of these two factors to find an image x that minimizes the loss.

Vision Applications in Autonomous Driving

One of the ma manipulating
objects (pickinc and obstacle
avoidance navic



Navigation t ıtdoor
environment

(1) Map buildi ncluding the
robot's position

(2) Path plann the current
position to the target position.

Lecture 14 ILOs

- Image Formation
- Simple Image Features
- Classifying Images
- Detecting Objects
- The 3D World
- Using Computer Vision