Introduction to object recognition

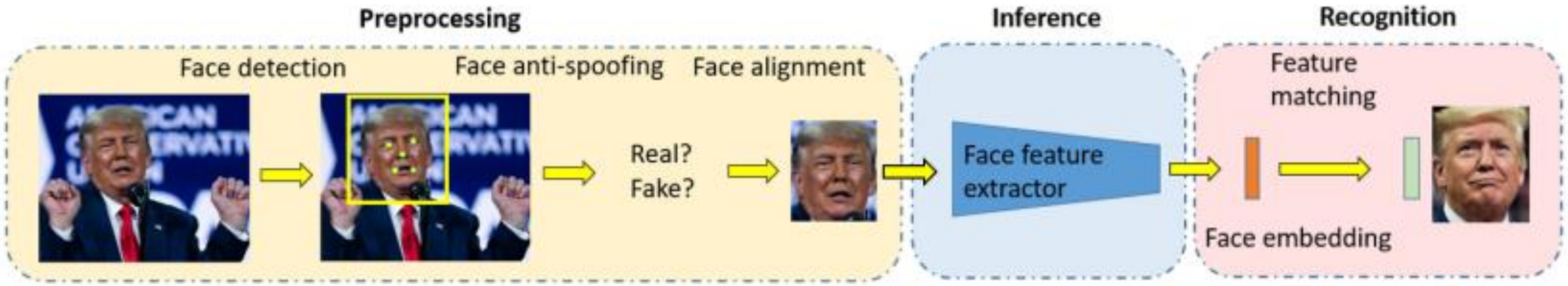
Fiseha B. Tesema, PhD

Recap

- Motion Field and Optical Flow:
- Optical Flow Constraint Equation
  - Aperture problem
- Lukas Kanade Method
- What if we have large Motion
  - Coarse to Fine Flow Estimation
- Dense and Sparse Optical Flow
- Application of Optical Flow

Object recognition

- Object Recognition is a field of artificial intelligence (AI) and computer vision that enables machines to identify, classify, and locate objects within digital images or video frames.
- It involves training algorithms to interpret visual data by recognizing patterns, shapes, textures, and contextual information to distinguish between different objects.



[A Survey of Face Recognition, Xin you etal, 2022]

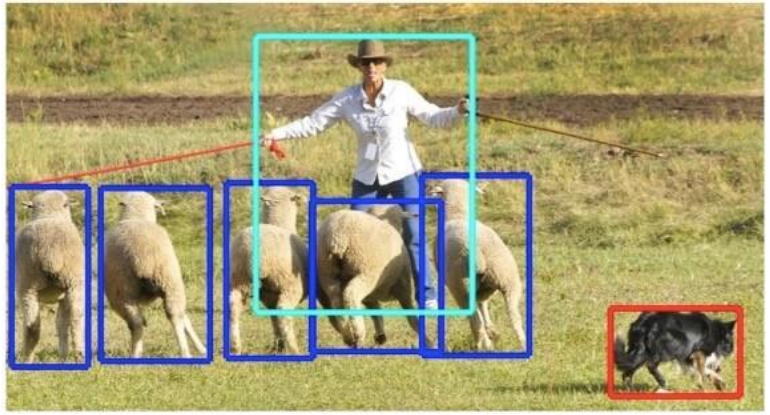
Core components of Object Recognition

- Object Detection – Locates objects in an image/video and draws bounding boxes around them.
- Object Classification – Assigns a label (e.g., "dog," "car") to the detected object.
- Object Localization – Precisely identifies the position of the object within the image.
- Instance Segmentation – Distinguishes between different instances of the same object (e.g., multiple people in a crowd).
- Semantic Segmentation
  - Classifies every pixel in an image into a category (e.g., "road," "sky," "person").

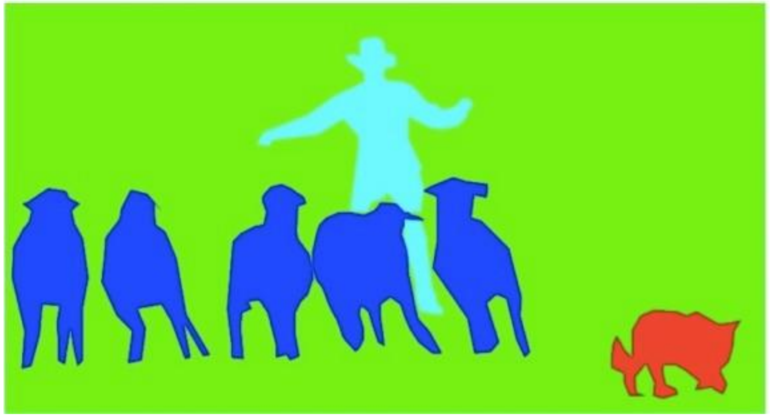
Recognition: What type of output?

Image classification

Object detection

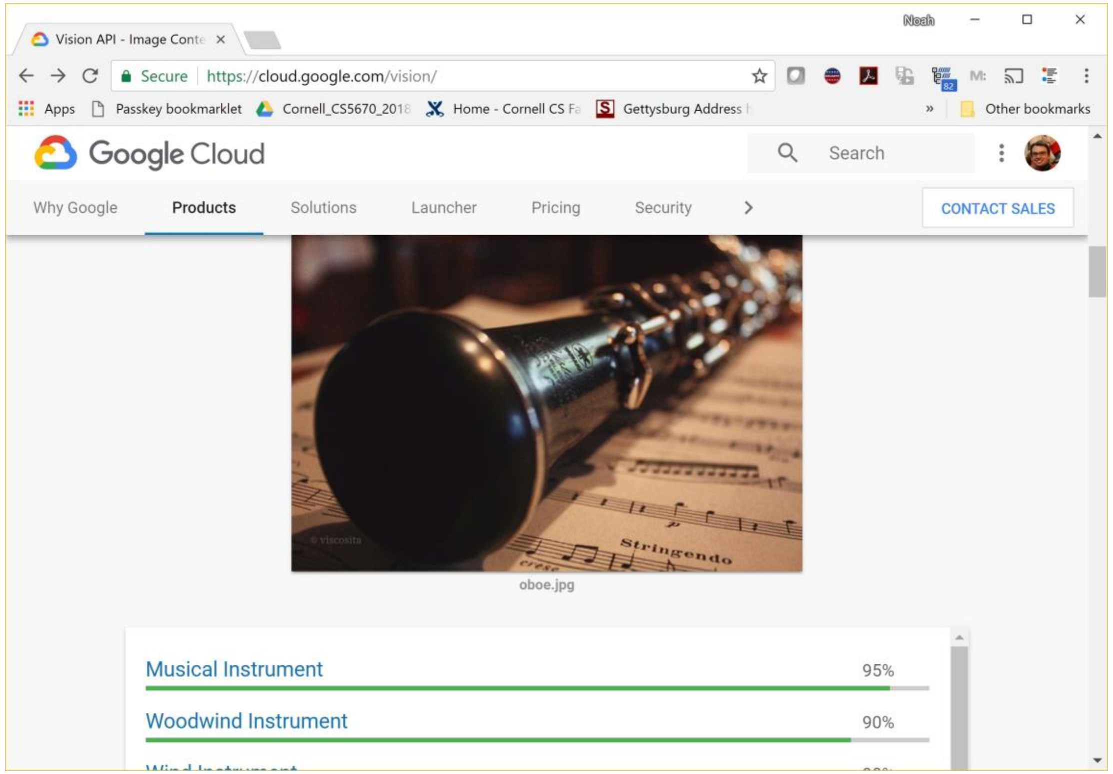


Semantic segmentation



Instance segmentation

Image classification demo



https://cloud.google.com/vision/docs/drag-and-drop

See also:
https://aws.amazon.com/rekognition/
https://www.clarifai.com/
https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/
…

Next few slides adapted from Li, Fergus, & Torralba's excellent [short course on](#) category and object recognition

What is “Recognition”?

Verification: is that a lamp?

What is "Recognition"?

- Detection: where are the people?

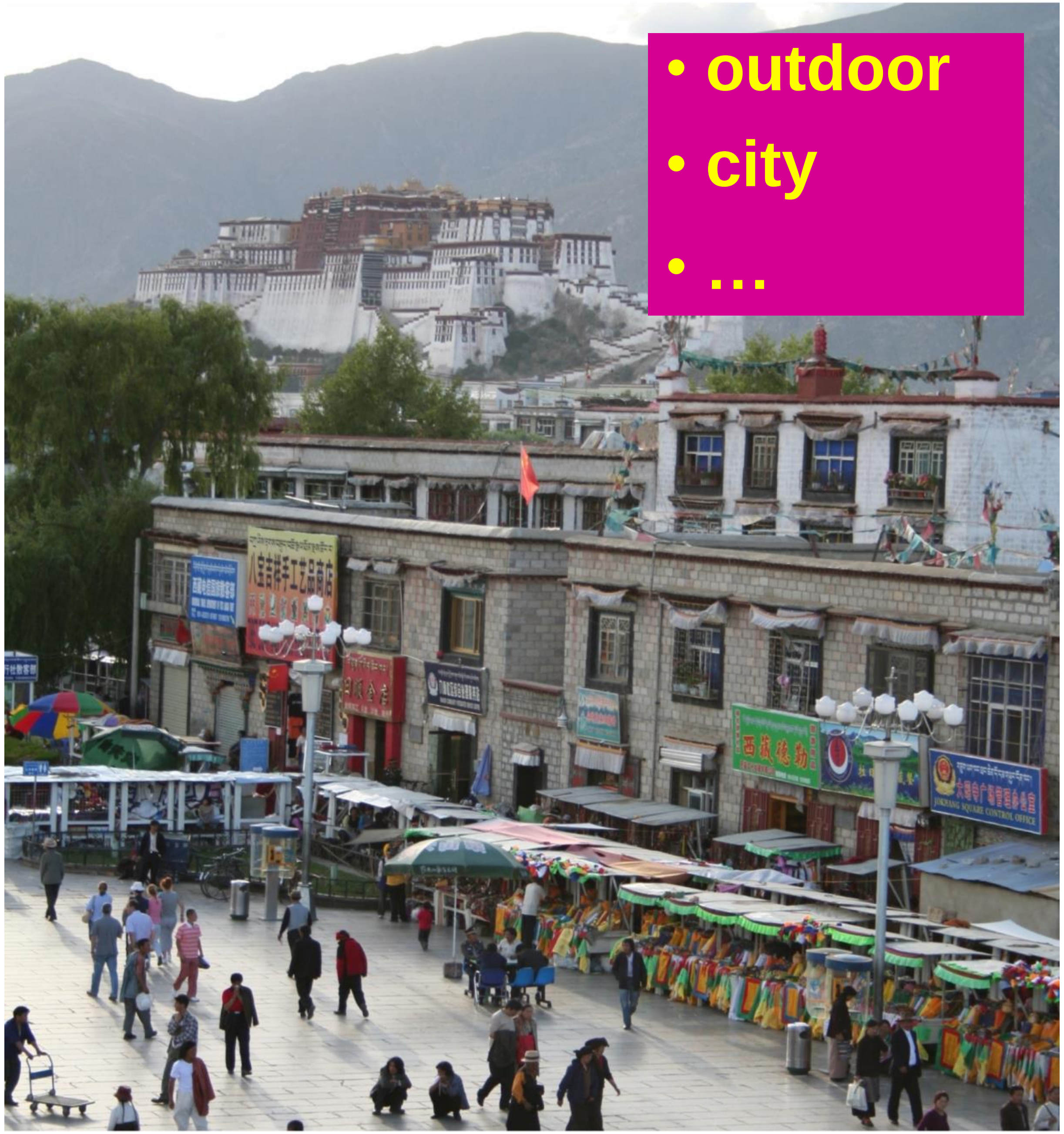
What is “Recognition”?

- Identification: is that Potala Palace?

What is “Recognition”?

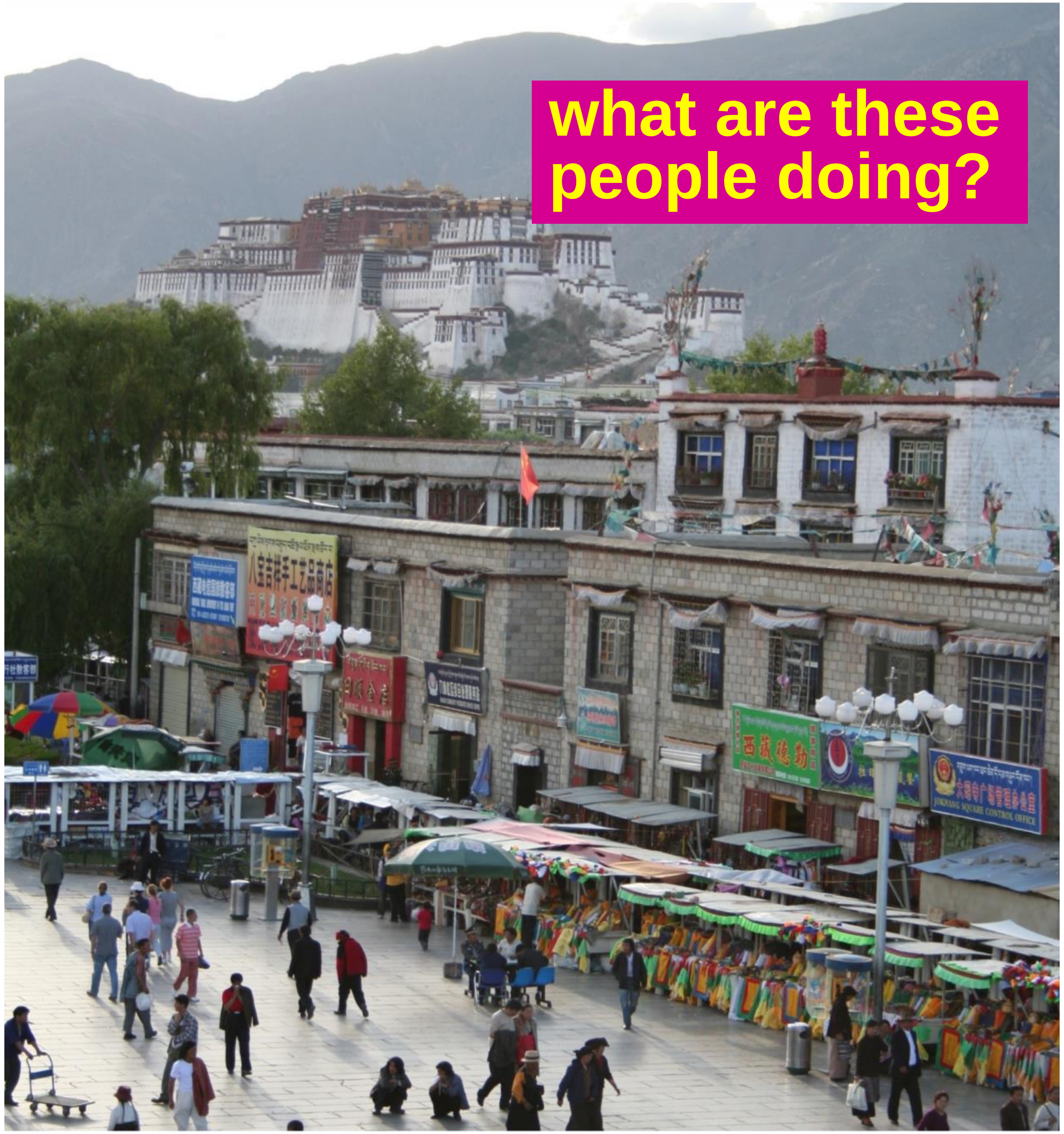
- Object categorization

What is “Recognition”?



- Scene and context categorization

What is “Recognition”?

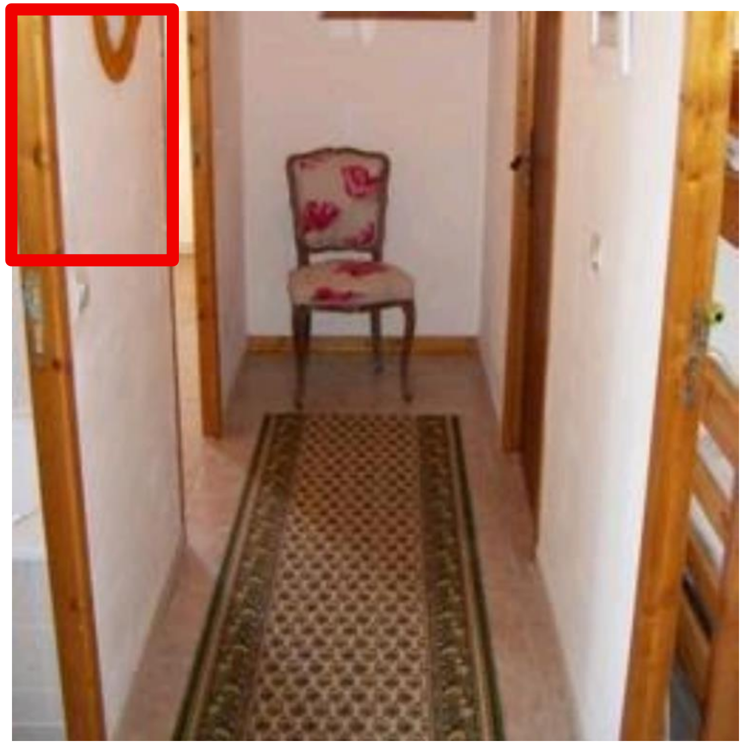


- Activity / Event Recognition

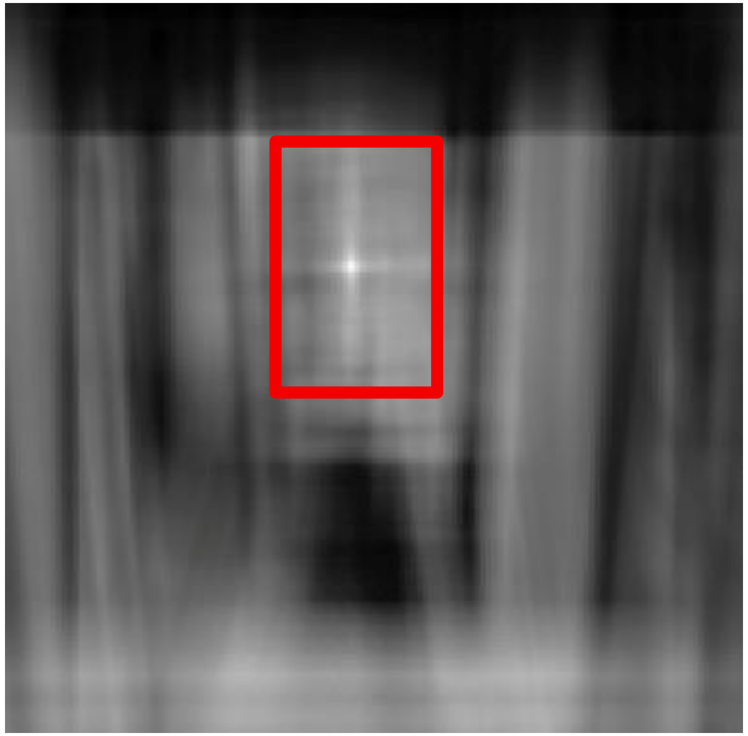
Object recognition: Is it really so hard?

This is a chair

Find the chair in this image

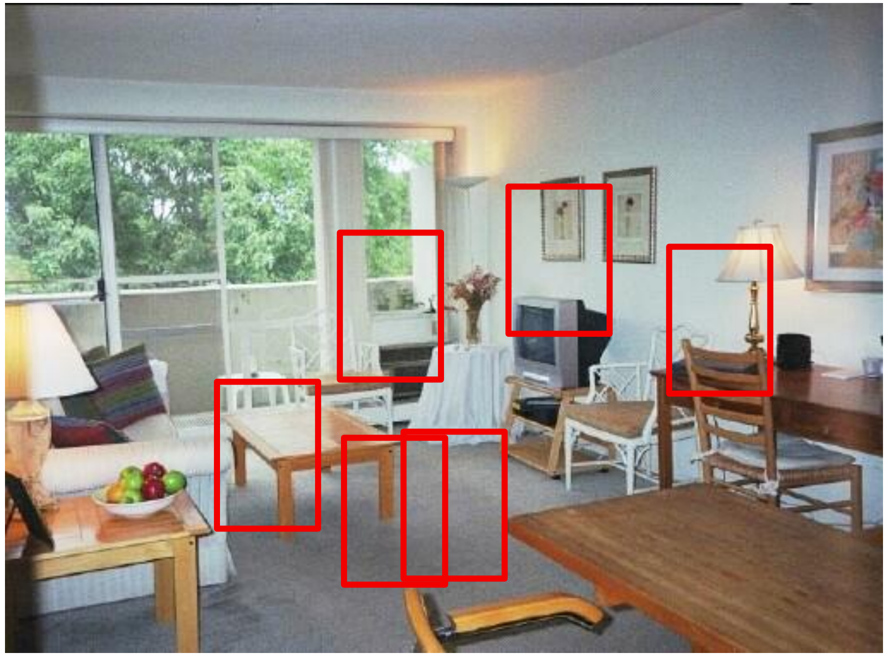


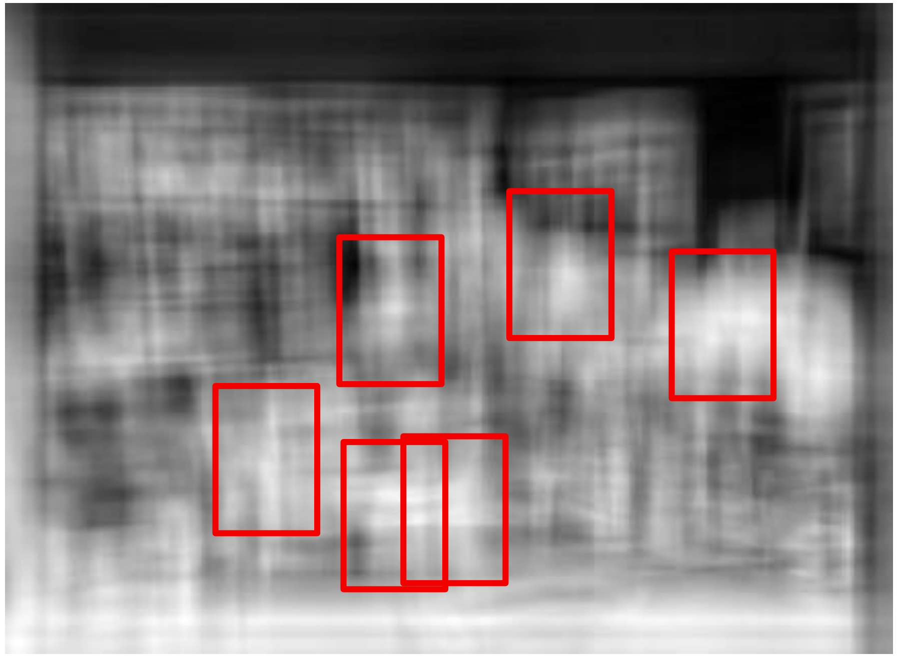
Output of normalized correlation



Object recognition: Is it really so hard?

Find the chair in this image





Pretty much garbage:
Simple template matching is not
going to do the trick

Object recognition:Is it really so hard?

Find the chair in this image

A “popular method is that of template matching, by point to point correlation of a model pattern with the image pattern. These techniques are inadequate for three-dimensional scene analysis for many reasons, such as occlusion, changes in viewing angle, and articulation of parts.” Nivatia & Binford, 1977.

Why not use SIFT matching for everything?

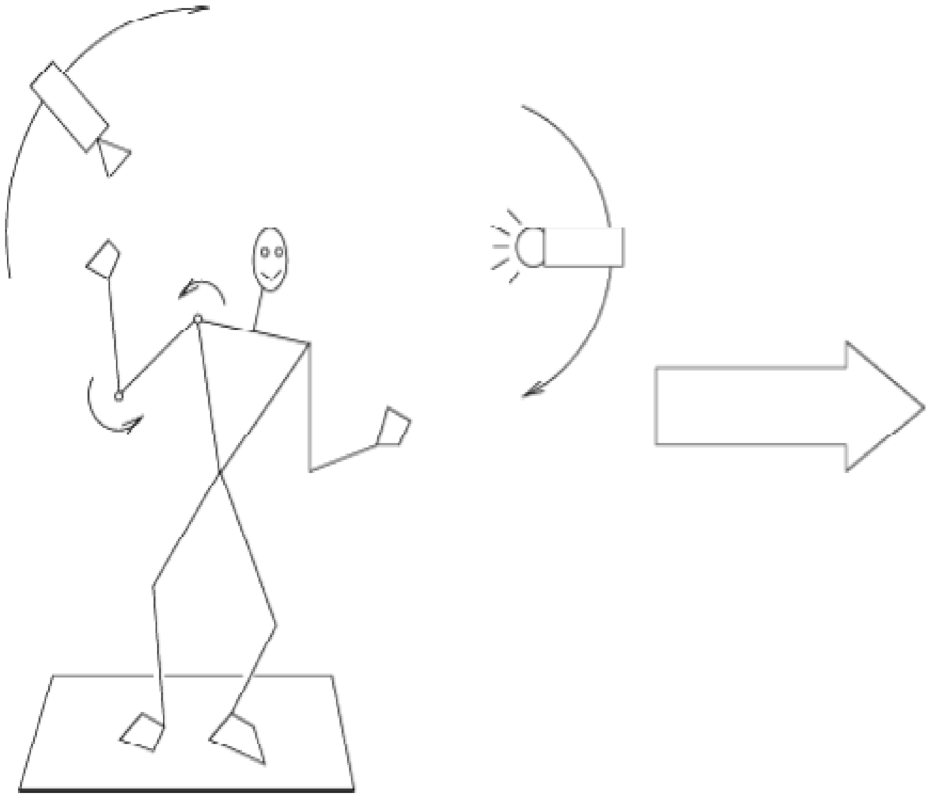
- Works well for object *instances* (or distinctive images such as logos)

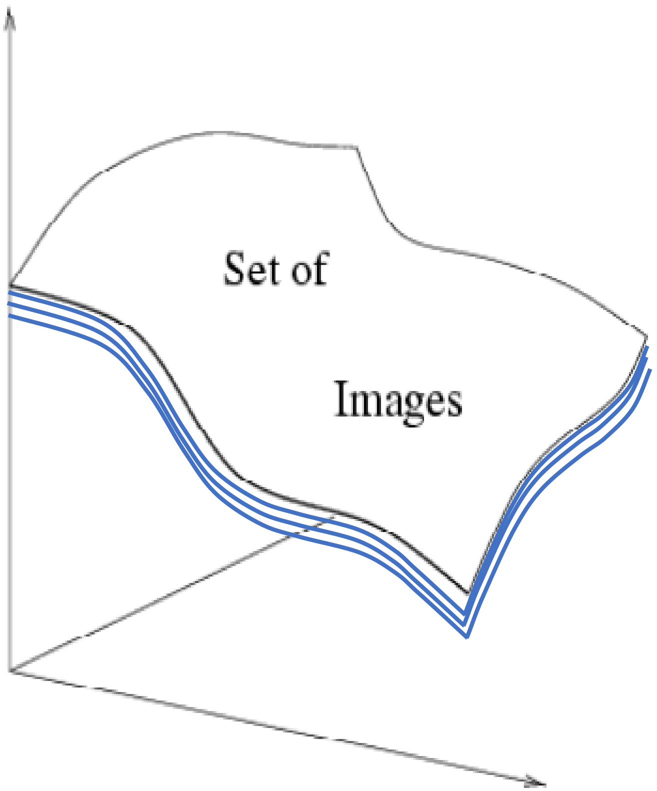
- Not great for generic object *categories*

And it can get a lot harder

Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. J Vis, 3(6), 413-422

Why is recognition hard?



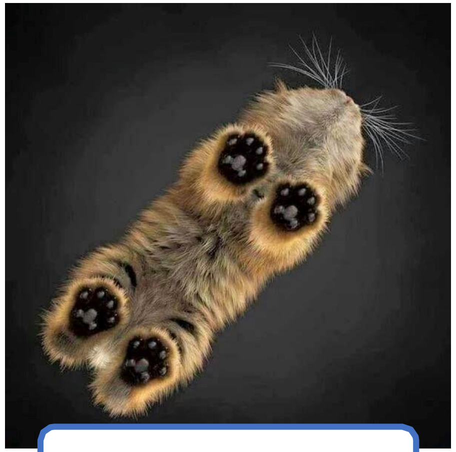


Variability: Camera position,
Illumination,
Shape,
etc…

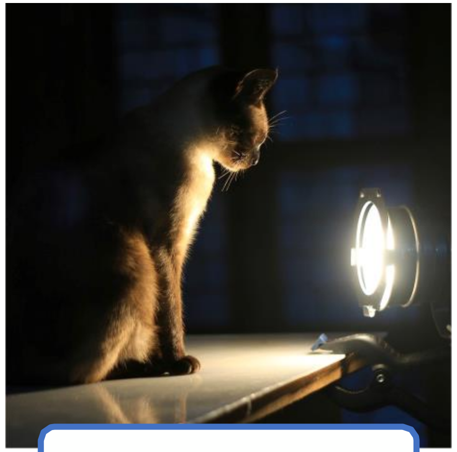
Challenge: lots of potential classes

Variation Makes Recognition Hard

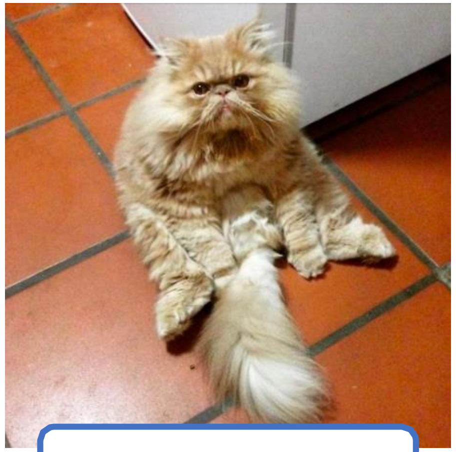
- The same class of object can appear *very* differently in different images



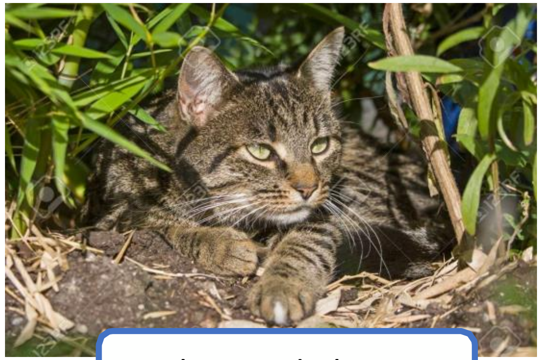
Viewpoint Variation



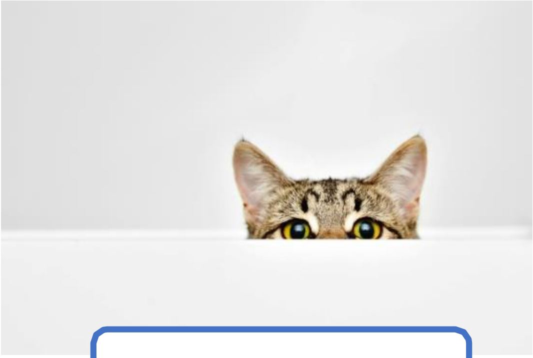
Lighting Variation



Deformation



Background Clutter



Occlusion

The Semantic Gap

What we see

What the computer sees

Image Classifiers in a Nutshell

- Input: an image
- Output: the class label for that image

```
def classifier(image):
        //Do some stuff
        return class_label;
```

- Label is generally one or more of the discrete labels used in training
  - e.g. {cat, dog, cow, toaster, apple, tomato, truck, … }

$f($  $)$ = "Cat"

$f($  $)$ = "Dog"

$f($  $)$ = "Toaster"

The Problem is Under-constrained

- Distinct realities can produce the same image…
- We generally can't compute the "right" answer, but we can compute the most likely one…
- We need some kind of prior to condition on. We can learn this prior from data:

$$f(x) = \underset{\ell_x}{argmax}\ P(\ell_x|data)$$

What Matters in Recognition?

- Data
  - More is always better (as long as it is good data)
  - Annotation is the hard part
- Representation
  - Low level: SIFT, HoG, GIST, edges
  - Mid level: Bag of words, sliding window, deformable model
  - High level: Contextual dependence
  - Deep learned features
- Learning Techniques
  - E.g. choice of classifier or inference method

What Matters in Recognition?

- Data
  - More is always better (as long as it is good data)
  - Annotation is the hard part

24 Hrs in Photos

Flickr Photos From 1 Day in 2011

https://www.kesselskramer.com/project/24-hrs-in-photos/

Data Sets

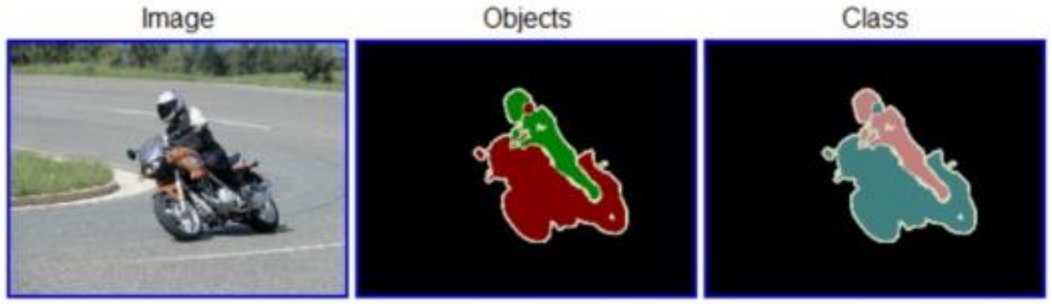
- PASCAL VOC
  - Not Crowdsourced, bounding boxes, 20 categories
- **ImageNet**
  - Huge, Crowdsourced,
- SUN Scene Database, Places
  - Not Crowdsourced, 397 (or 720) scene categories
- LabelMe (Overlaps with SUN)
  - Sort of Crowdsourced, Segmentations, Open ended
- SUN Attribute database (Overlaps with SUN)
  - Crowdsourced, 102 attributes for every scene
- OpenSurfaces
  - Crowdsourced,
- Microsoft COCO
  - Crowdsourced, large-scale objects

… and many more https://paperswithcode.com/datasets?task=image-classification

The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

- 20 object categories (aeroplane to TV/monitor)

- Three challenges:
    - Classification challenge (is there an X in this image?)
    - Detection challenge (draw a box around every X)
    - Segmentation challenge (which class is each pixel?)

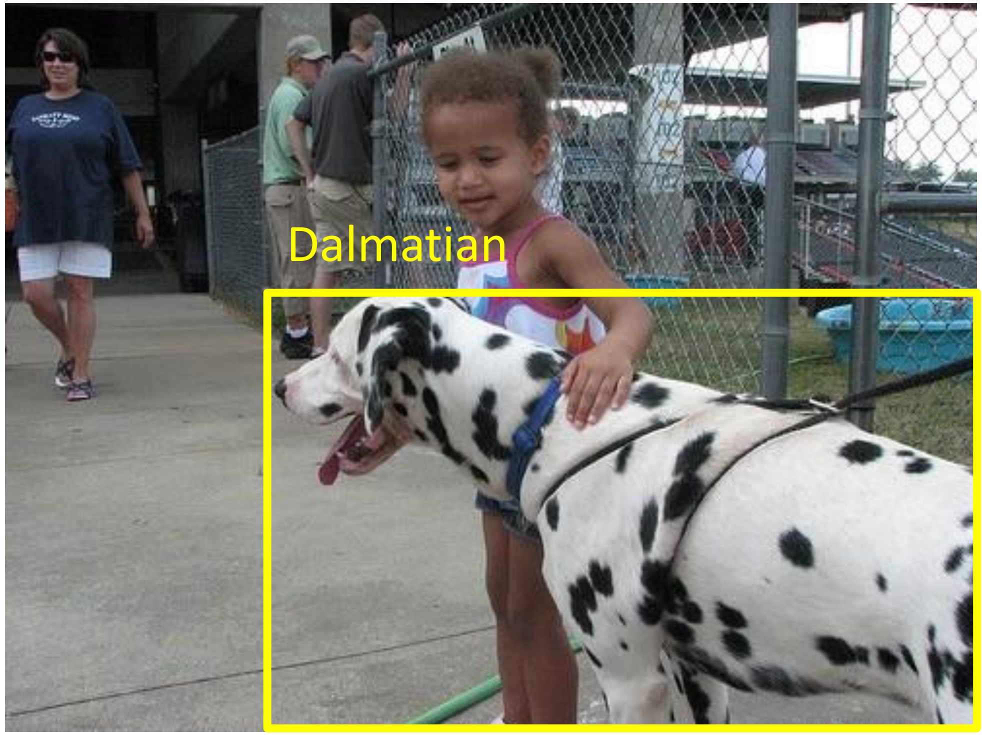


Large Scale Visual Recognition Challenge (ILSCRV)

2010-2017

~~20 object classes~~ ~~22,591 images~~

1000 object classes **1,431,167 images**



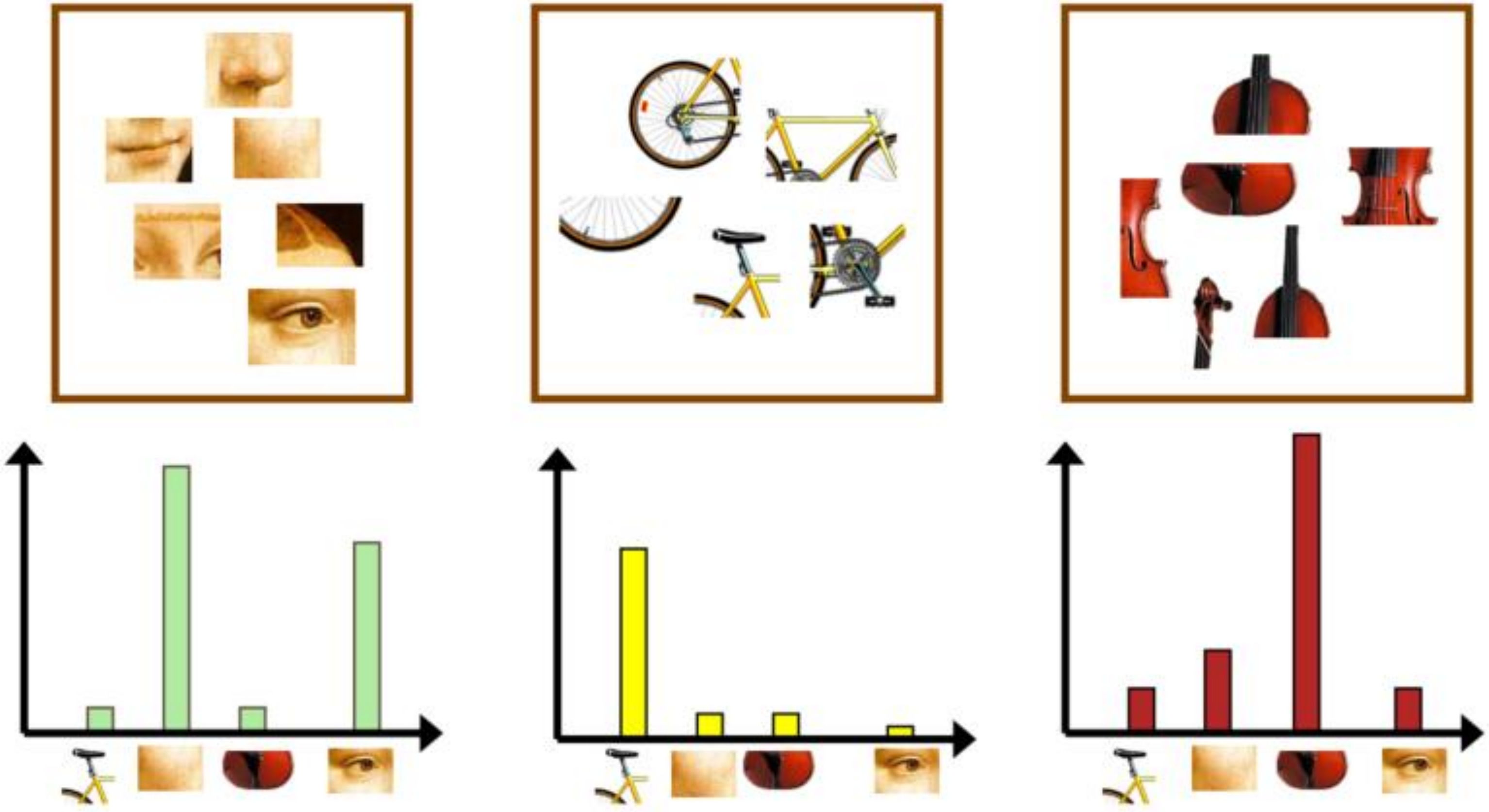
http://image-net.org/challenges/LSVRC/{2010,2011,2012}

Variety of object classes in ILSVRC

What Matters in Recognition?

- Representation
  - Low level: SIFT, HoG,
  - **Mid level**: Bag of words, sliding window, deformable model
  - High level: Contextual dependence
  - **Deep learned features**
    - **CNN**

Bag of wor

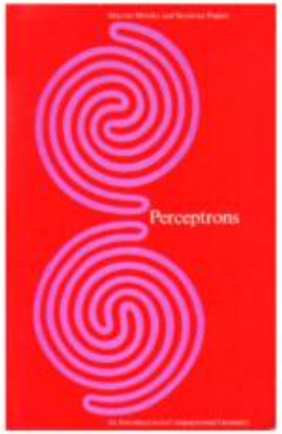


Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

Neural networks

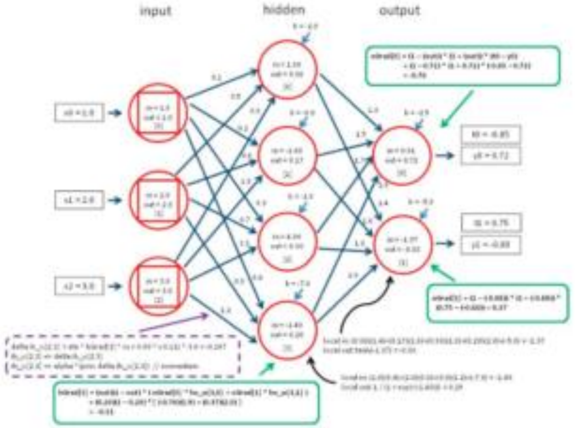
Perceptrons

Rosenblatt (1958)



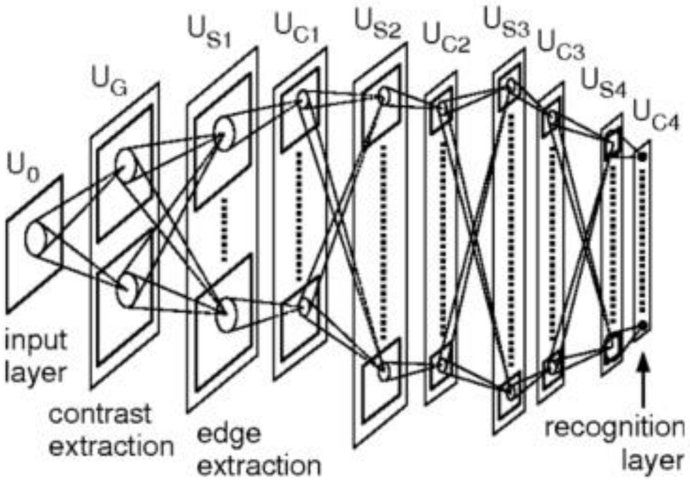
Minsky & Papert (1969)

Back-propagation



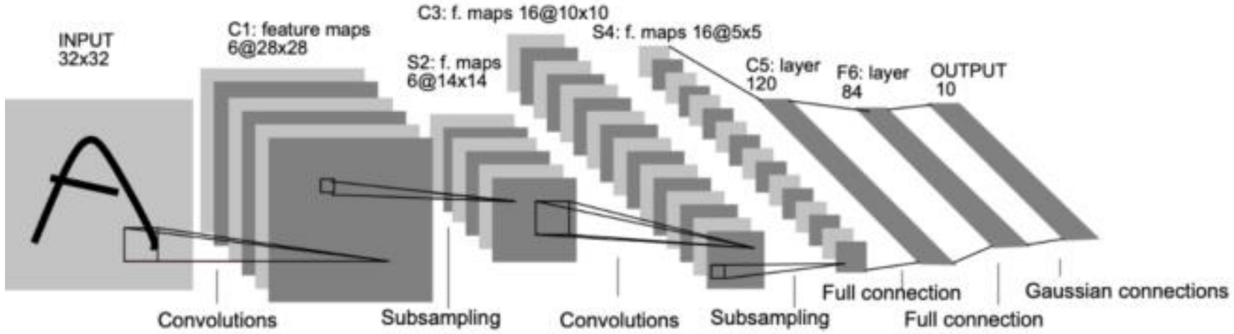
Rumelhart, Hinton & Williams (1986)

Neocognitron



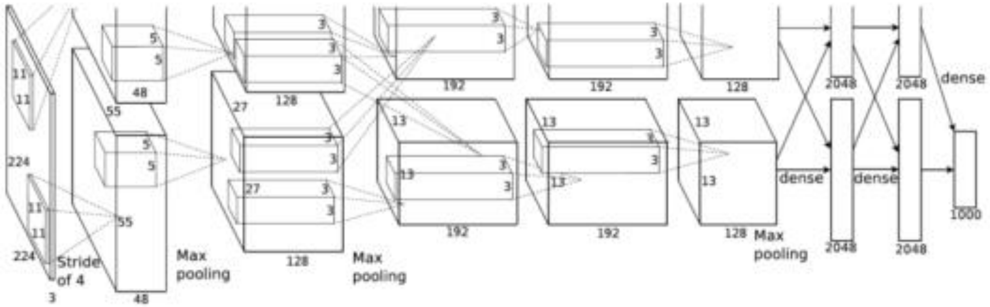
Fukushima (1980)

LeNet-5



LeCun et al. (1998)

AlexNet

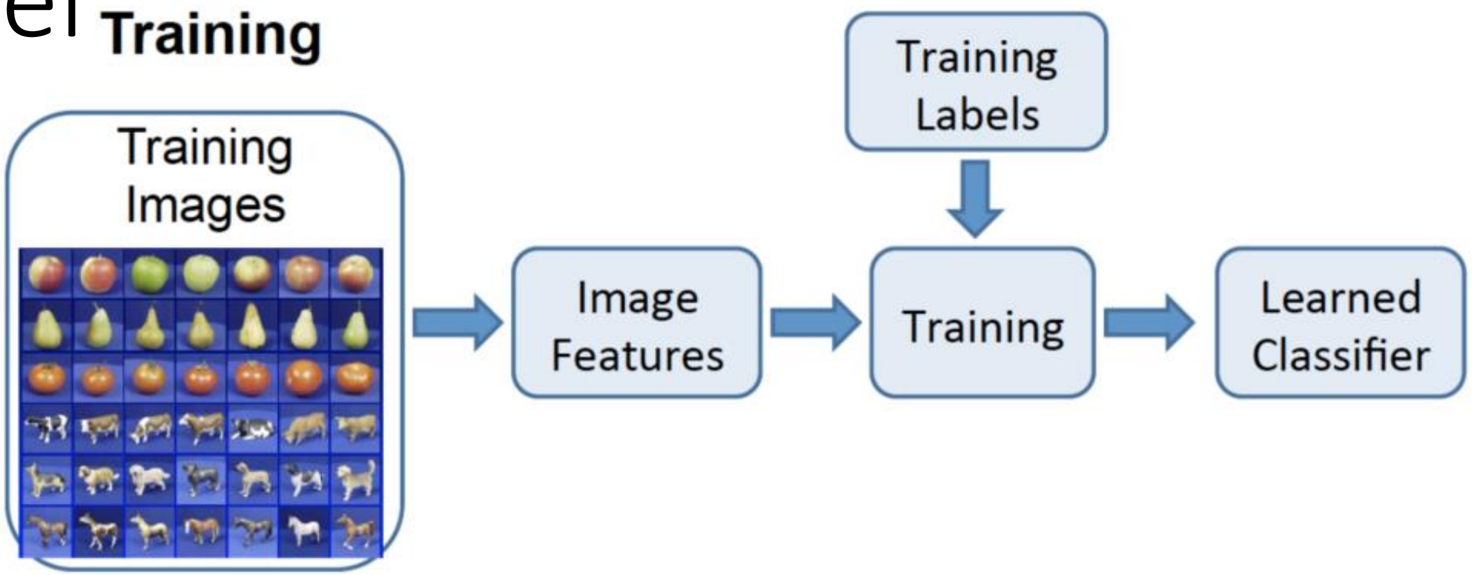


Krizhevsky et al. (2012)

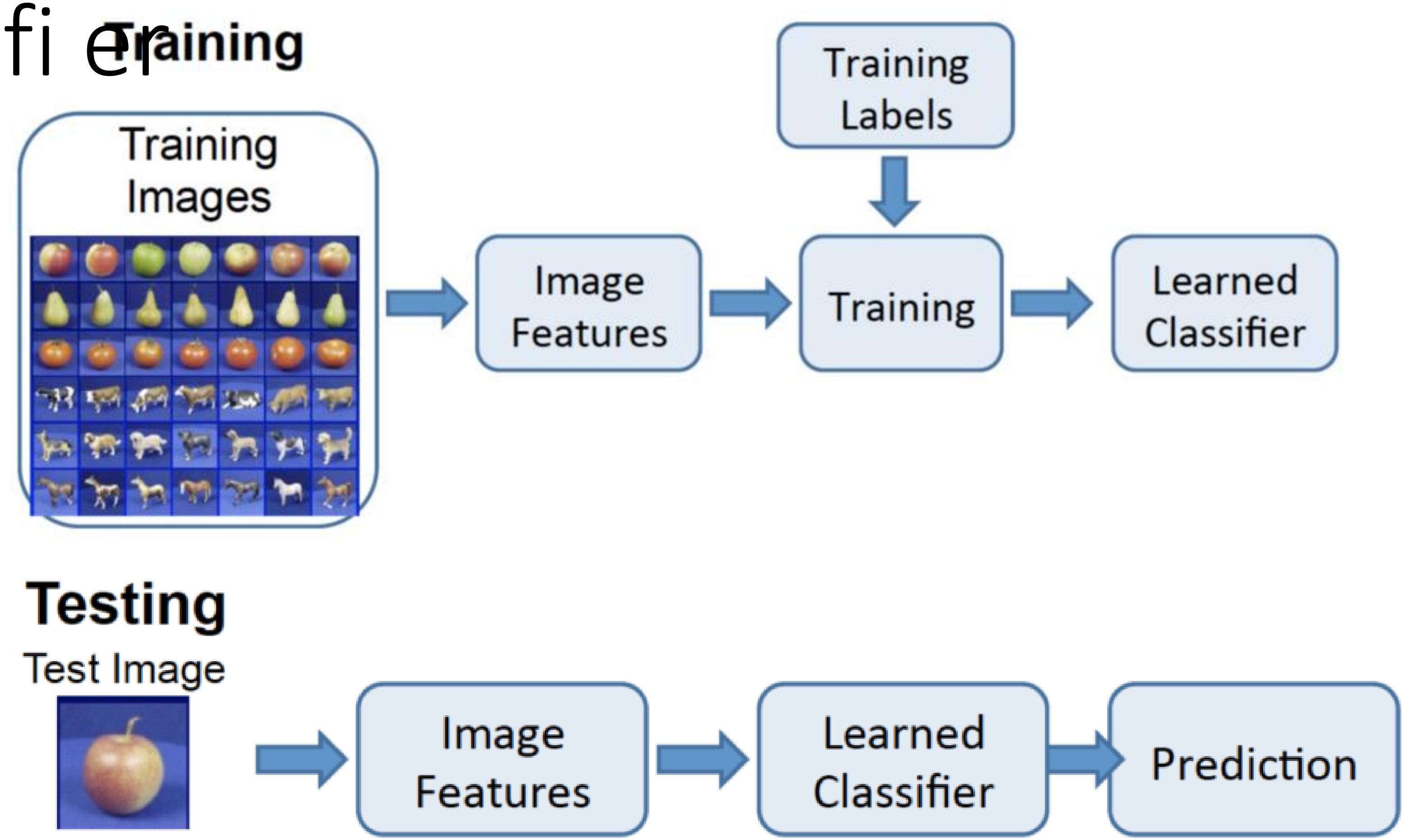
What Matters in Recognition?

- Learning Techniques
  - E.g. choice of classifier or inference method

Training & Testing a Classi fi er



Training & Testing a Classifier



Classifiers

- Nearest Neighbor
- kNN (“k-Nearest Neighbors”)
- Linear Classifier
- **Neural Network**
- **Deep Neural Network**

Next: Bag of features and Vola Jones