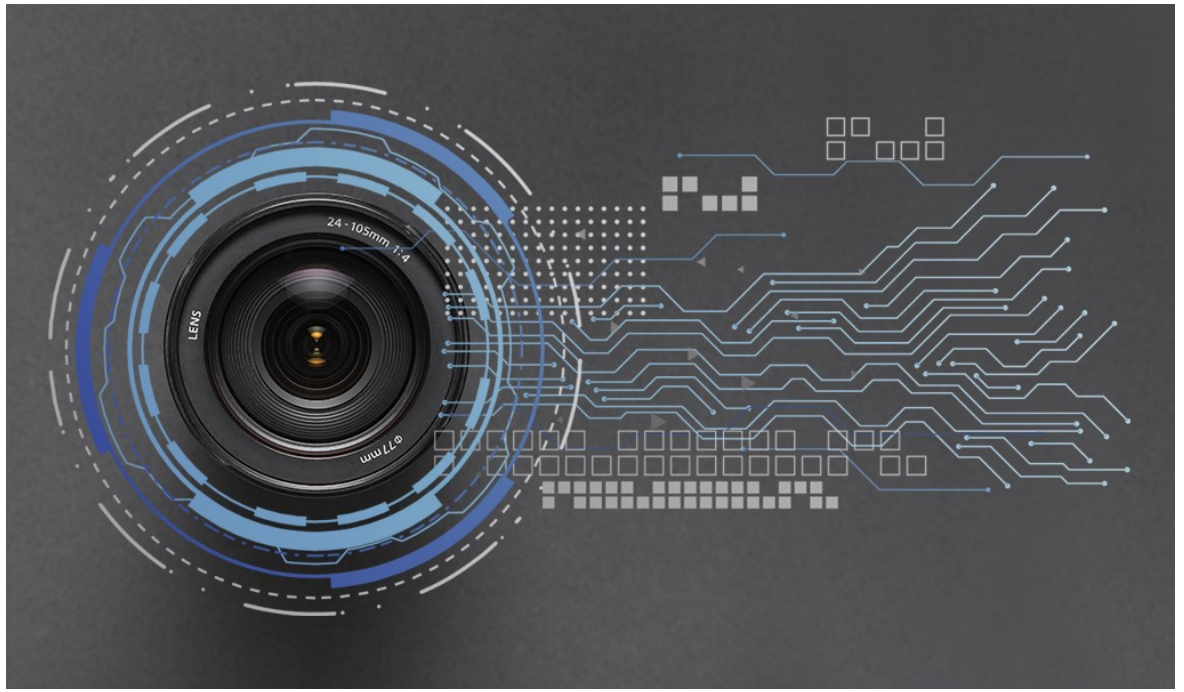
Computer vision and deep learning

Lecture 1

What is Computer Vision (CV)?

Goal: build computer-based vision systems which perform the same functions as the human visual system

CV builds the theoretical and algorithmic basis by which useful information about the world can be extracted and analysed from an image, a set of images (e.g. stereo vision) or an image sequence



Computer vision – multidisciplinary subject

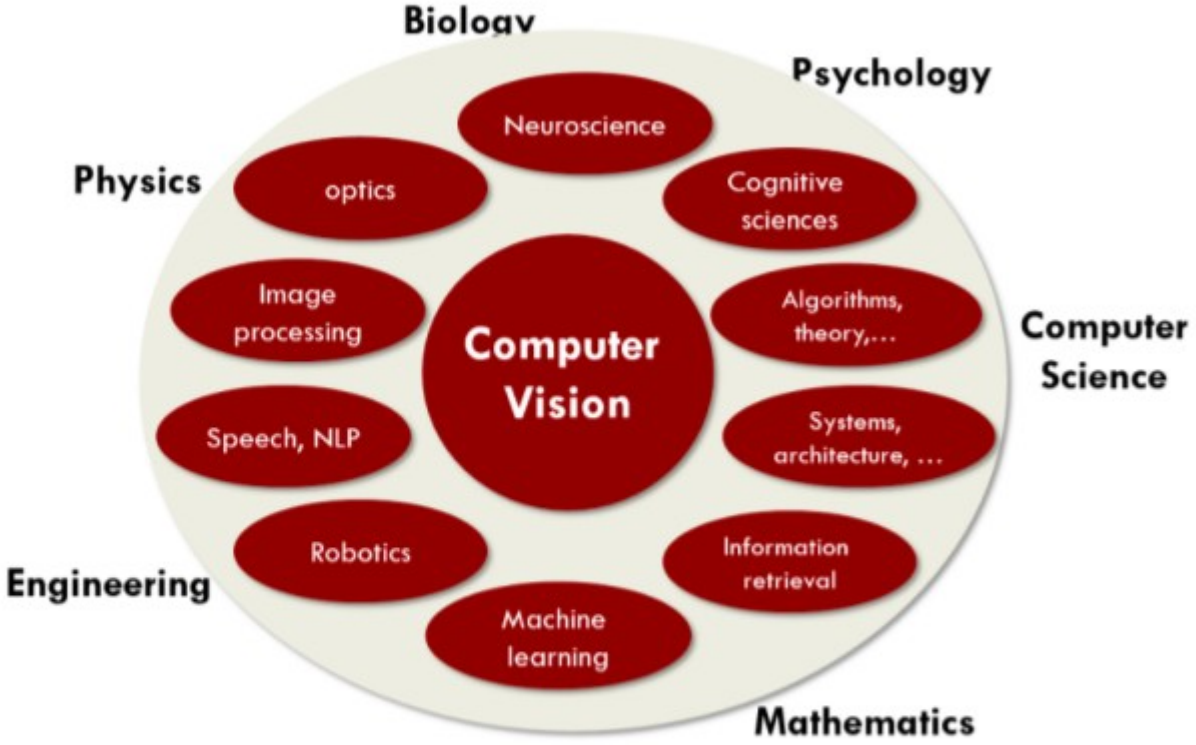


Figure 1: Computer vision at the intersection of multiple scientific fields

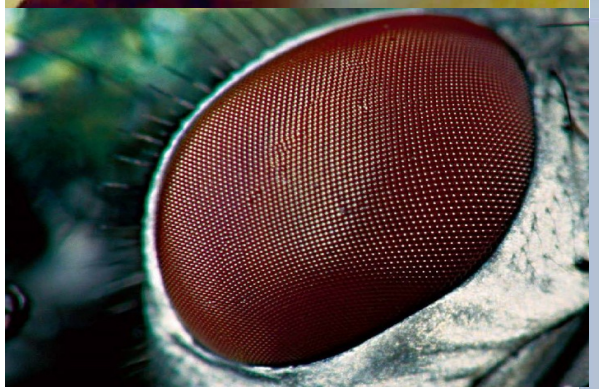
http://vision.stanford.edu/teaching/cs131_fall1718/files/cs131-class-notes.pdf

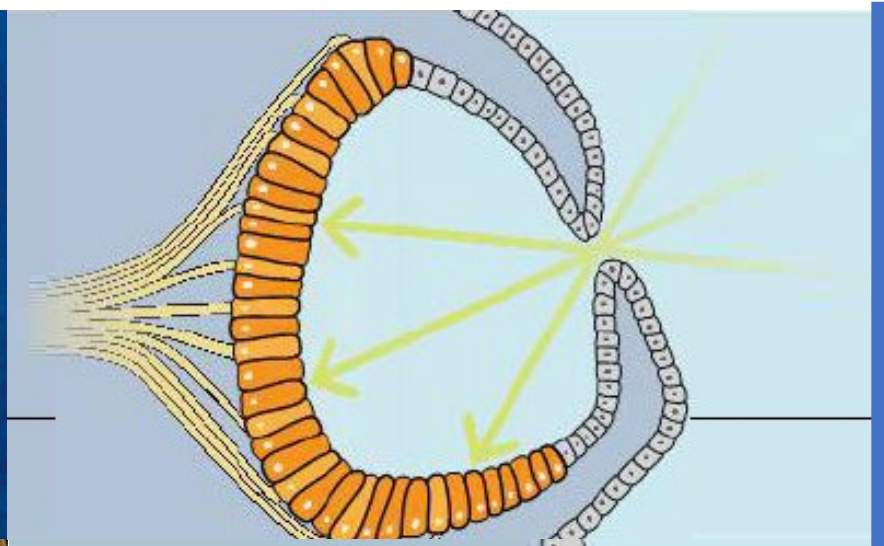
Image source: *Le miracle de la vision,* **Christian Schwagerl, Rainer Harf** et **Henning Engeln**

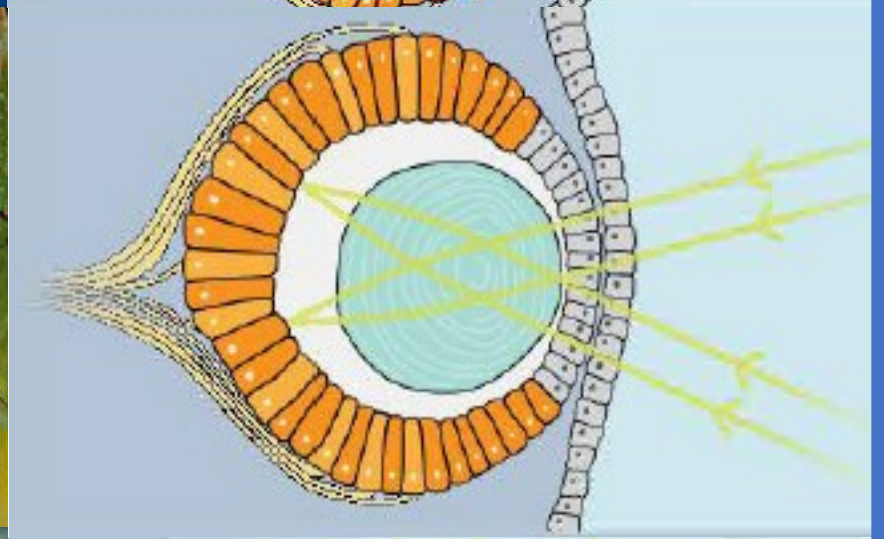
Evolution of the eye

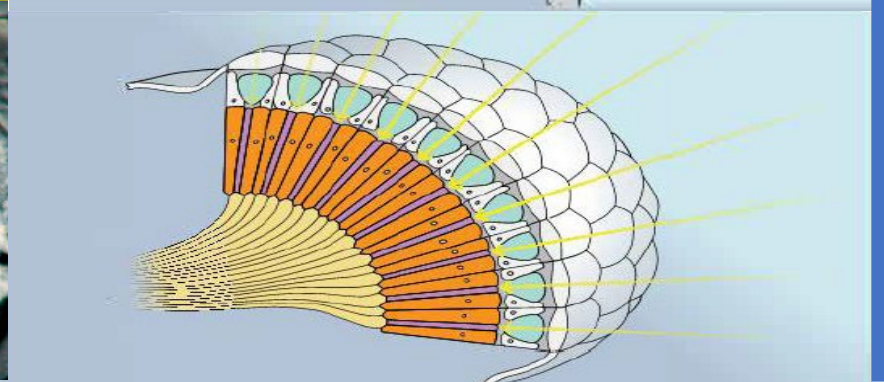
Cambrian Explosion: 541 million years ago

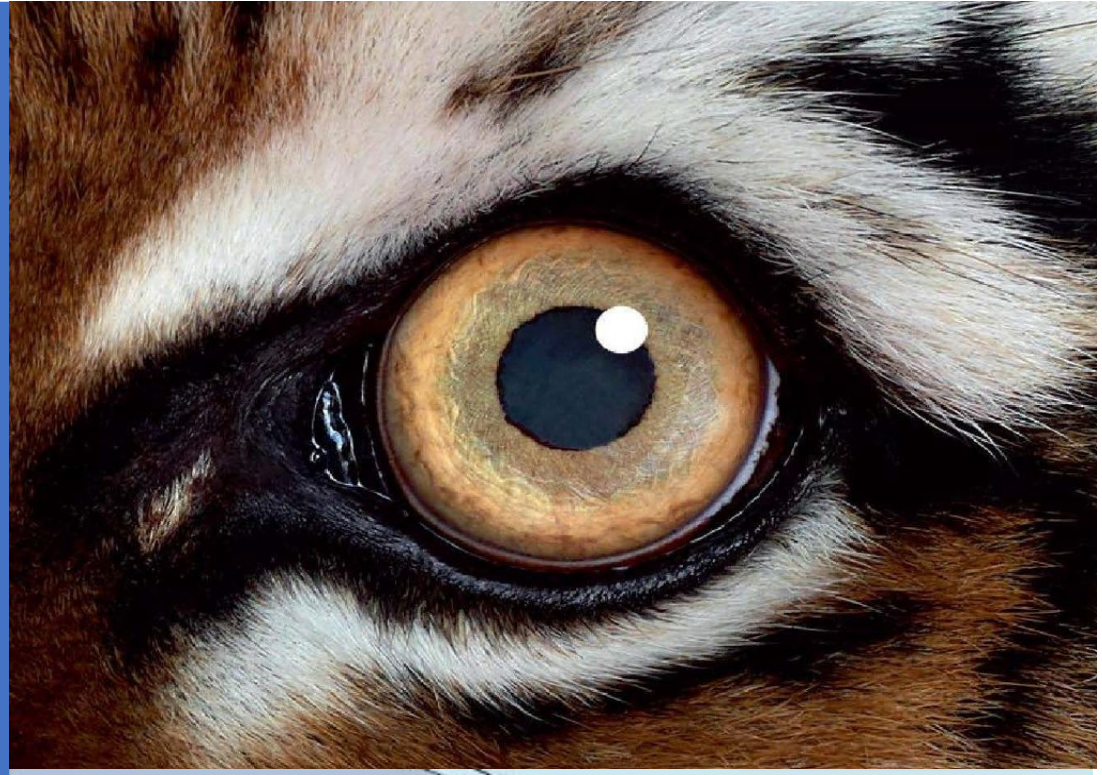
Image source: https://en.wikipedia.org/wiki/Cambrian_explosion

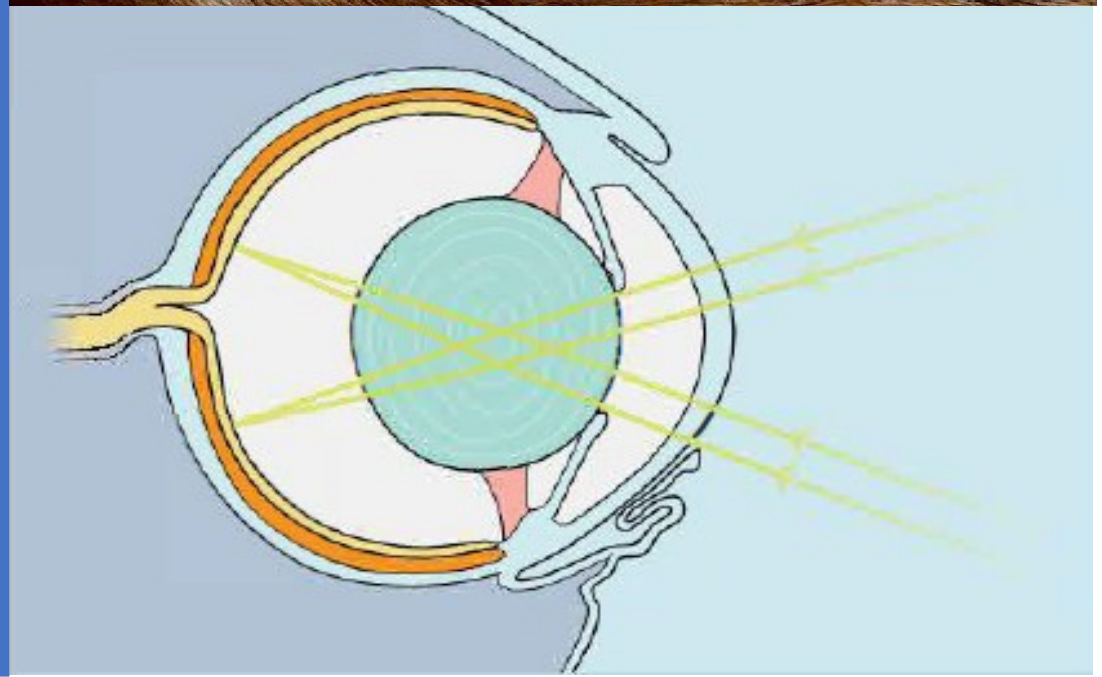






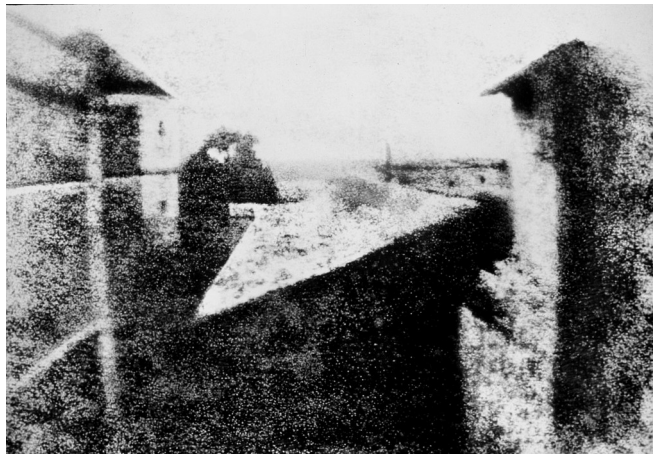




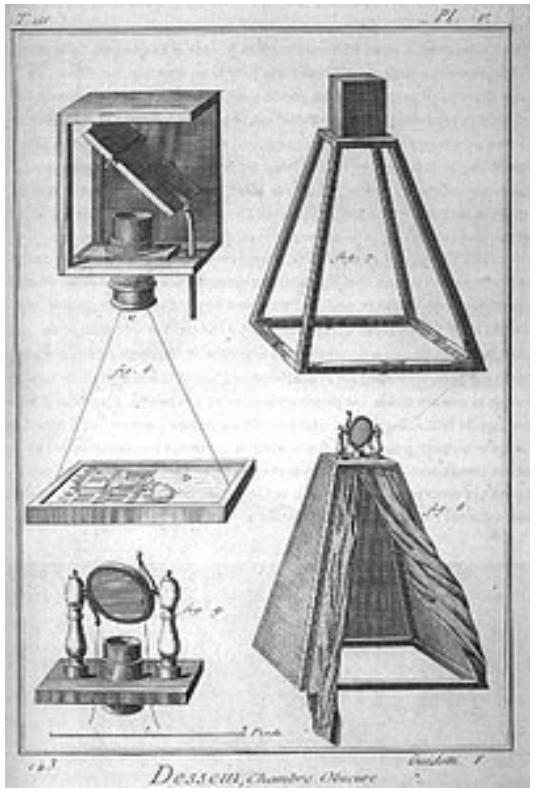


History of the camera

- Chinese philosopher *Mozi* (470 – 391 BC) first person in history to describe the first to describe the physical principle behind the camera
- Alhazen (Ibn al-Haytham) (965–1040 AD) – considered the inventor of the pinhole camera
- First permanent photograph (1825) by Joseph Nicéphore Niépce (8h exposure time)



View from the Window at Le Gras, 1827
Oldest surviving camera photograph



18th century description of a pinhole camera
Encyclopedia

Evolution of the camera

Kyocera Visual Phone VP-210 - 1999

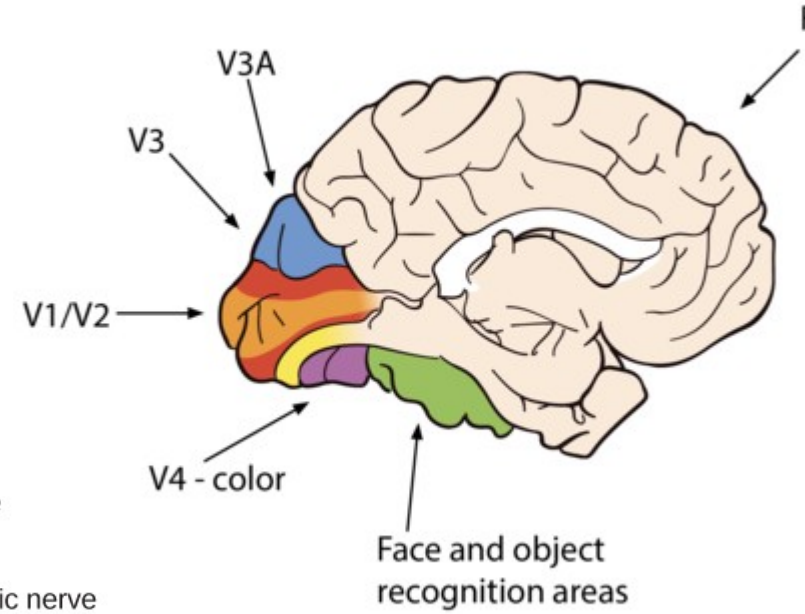
Image source: https://www.precisioncamerarepairs.com/blog/the-evolution-of-the-camera

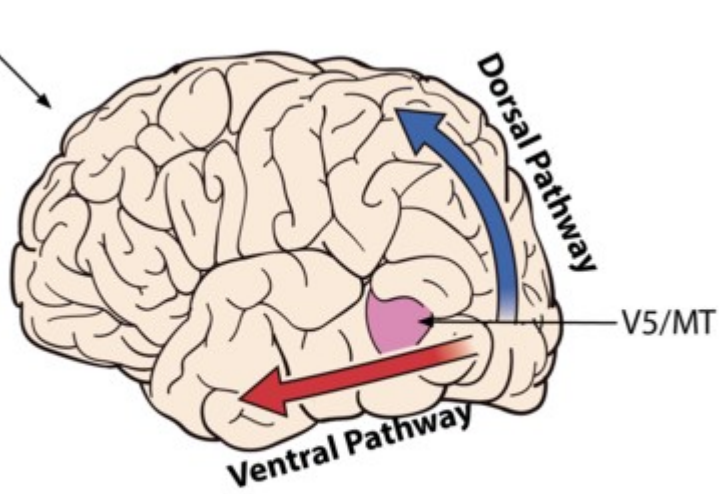
Nowadays…

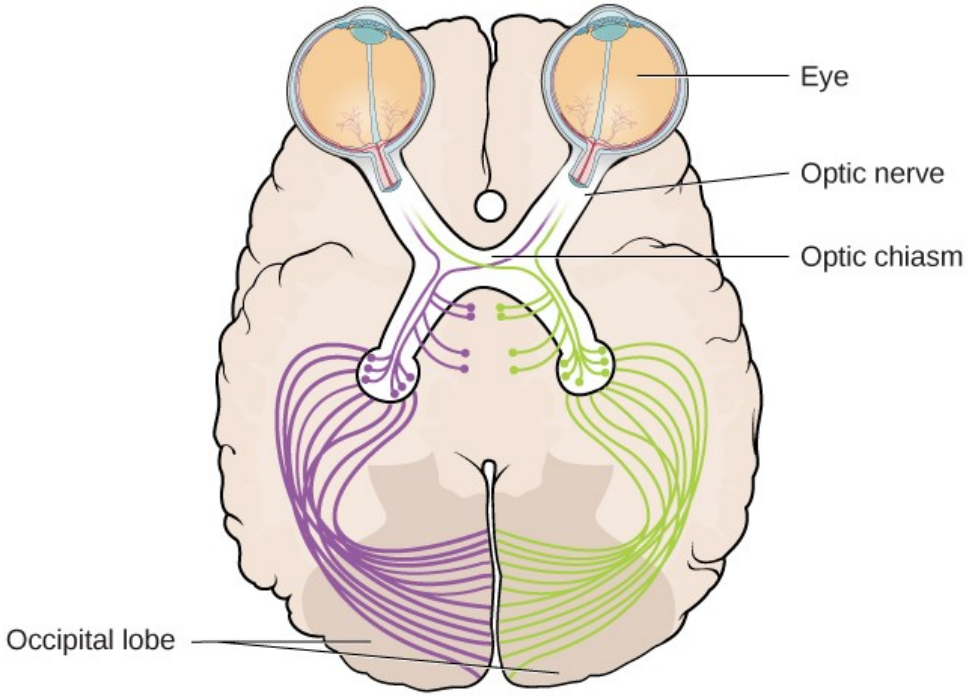
- ~ 1.436.300.000.000 (> 1.4 trillion) photos will be taken in 2020 (https://focus.mylio.com/tech-today/how-many-photos-will-be-taken-in-2020)
  - If you were to take all these pictures alone, and you took one pic per second = 45.544 years
- As of May 2019, > 500 hours of video are uploaded to YouTube every minute! (https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/)
- an estimate of 770 million surveillance cameras installed around the world today

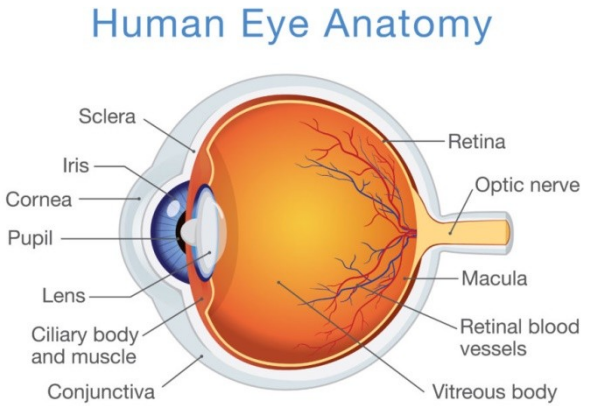
Visual cortex

Ventral pathway: “what” pathway
Dorsal pathway: “where/how” pathway



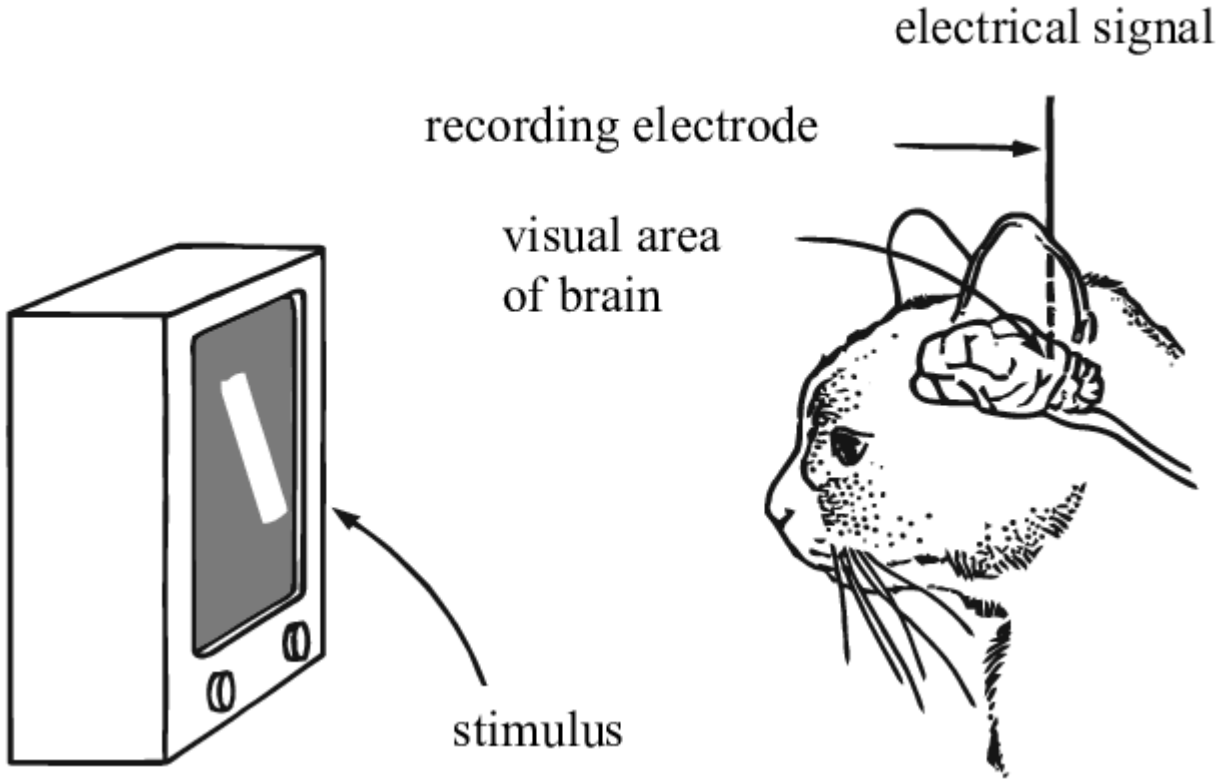






https://courses.lumenlearning.com/wmopen-psychology/chapter/outcome-vision/

Understanding the visual cortex



Hubel and Wiesel, 1959

https://www.youtube.com/watch?v=IOHayh06LJ4&ab_channel=PaulLester

Nobel Prize for Physiology or Medicine in 1981:
David Hubel and Torsten Wiesel

Simple cells:
orientation, position

Complex cells:
orientation, motion, direction

"Hypercomplex" cells:
orientation, motion, direction, length

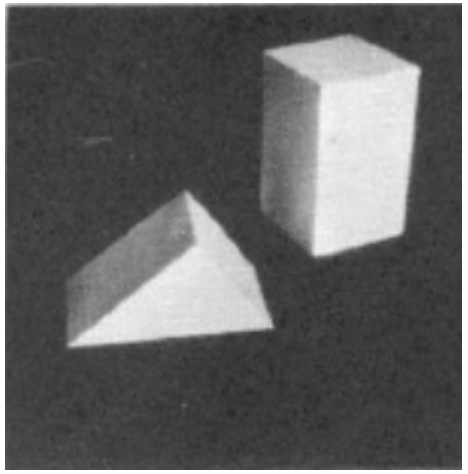
mage source: https://www.researchgate.net/publication/335707980_Understanding_Neural_Networks_via_Feature_Visualization_A_Survey

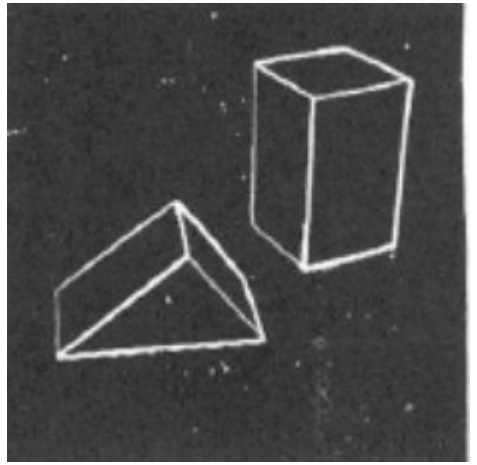
1963

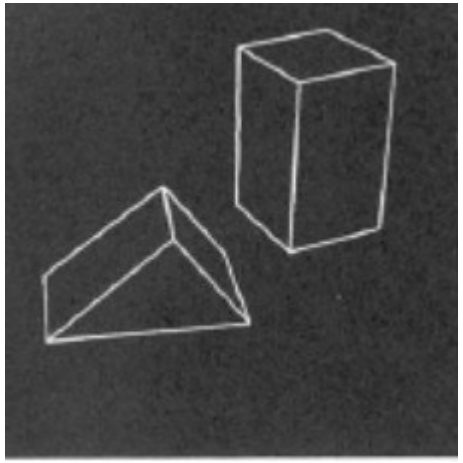
Larry Roberts

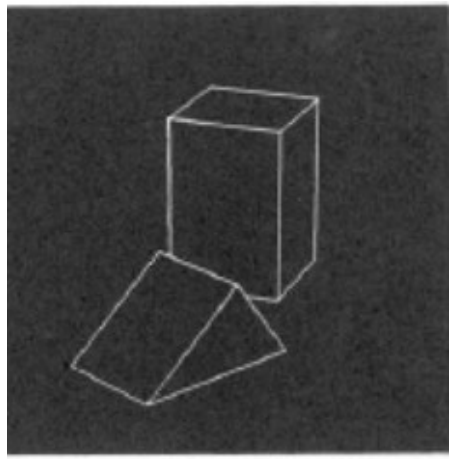
First PhD thesis in the field of Computer Vision: “Machine Perception of Three Dimensional Solids”, Larry Roberts, 1963, MIT Press

- extract 3D information about solid objects from 2D photographs of line drawings









1966

The Summer Vision Project

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

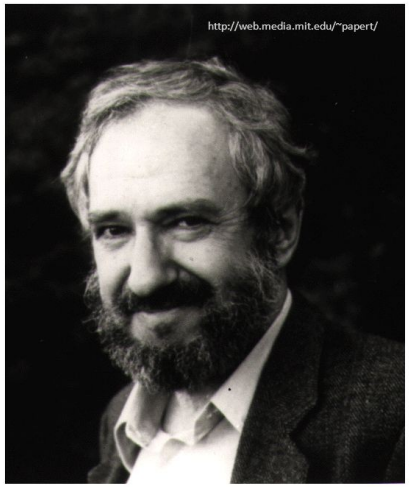
PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

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".



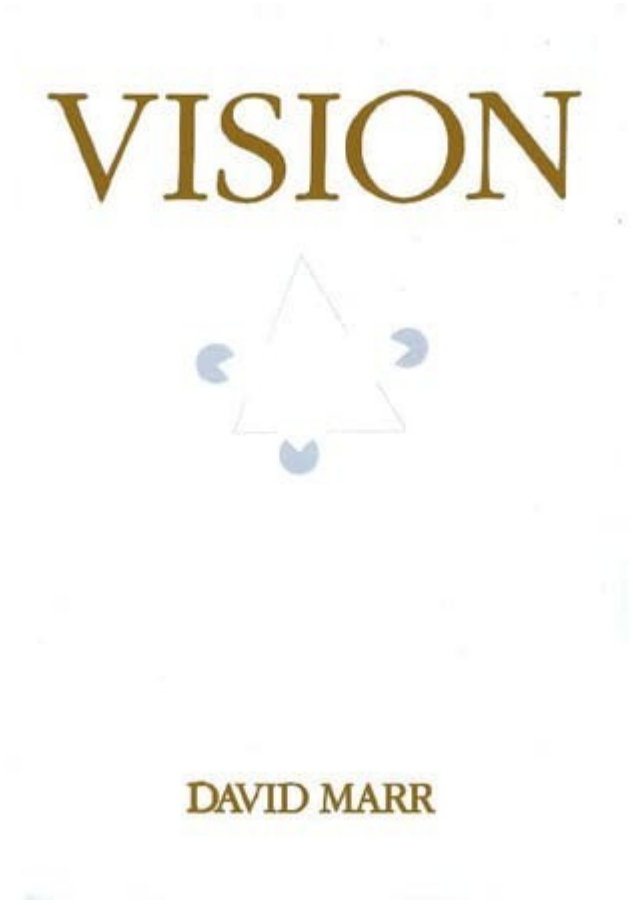
"I am convinced that the best learning takes place when the learner takes charge."

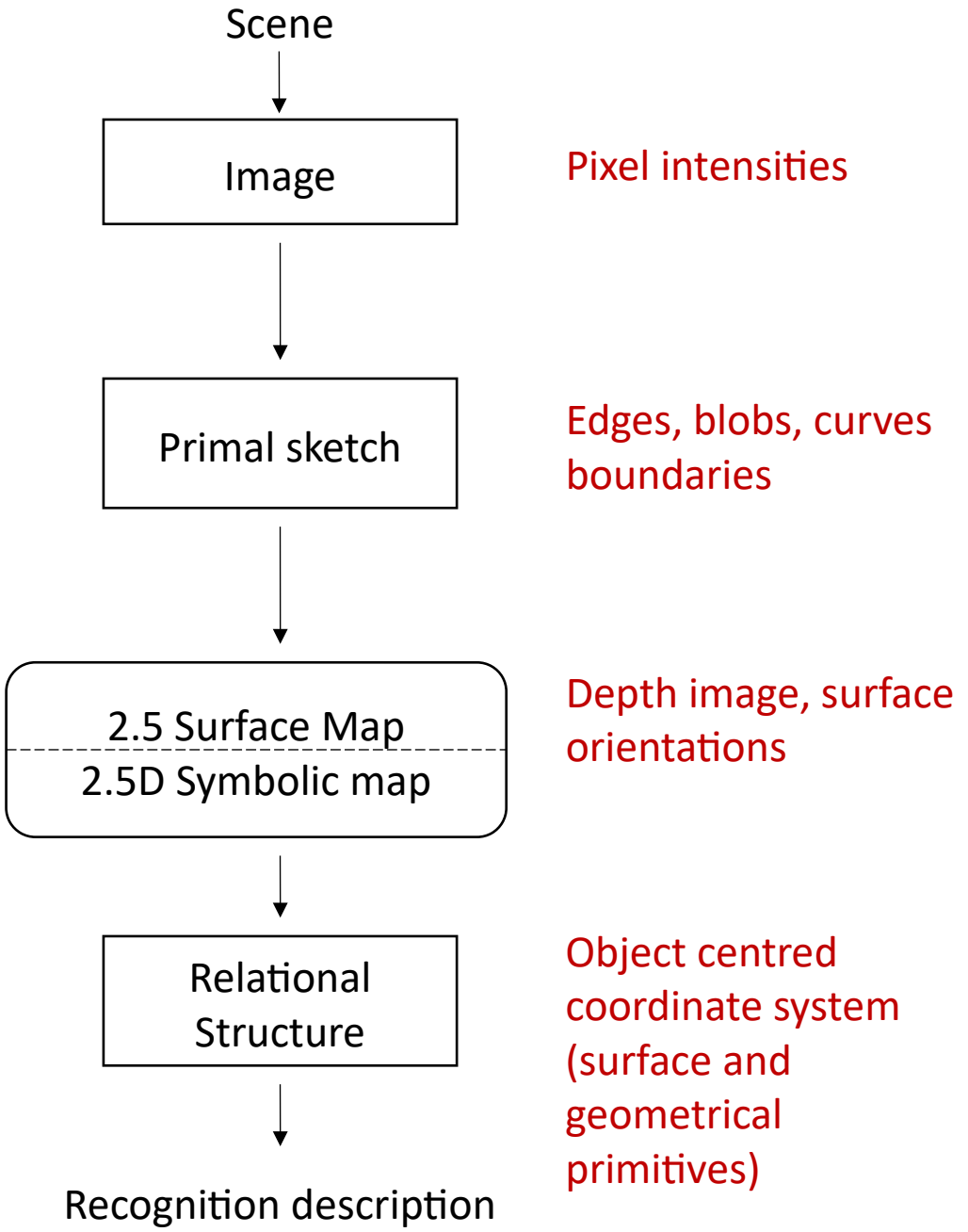
– Seymour Papert

The primary goal of the project is to construct a system of programs which will divide a vidisector picture into regions such as likely objects, likely background areas and chaos. We shall call this part of its operation FIGURE-GROUND analysis. It will be impossible to do this without considerable analysis of shape and surface properties, so FIGURE-GROUND analysis is really inseparable in practice from the second goal which is REGION DESCRIPTION. The final goal is OBJECT IDENTIFICATION which will actually name objects by matching them with a vocabulary of known objects.

1981

David Maar





1986

Canny edge detector




Images from https://towardsdatascience.com/canny-edge-detection-step-by-step-in-python-computer-vision-b49c3a2d8123

1997

Normalized Cuts and Image Segmentation

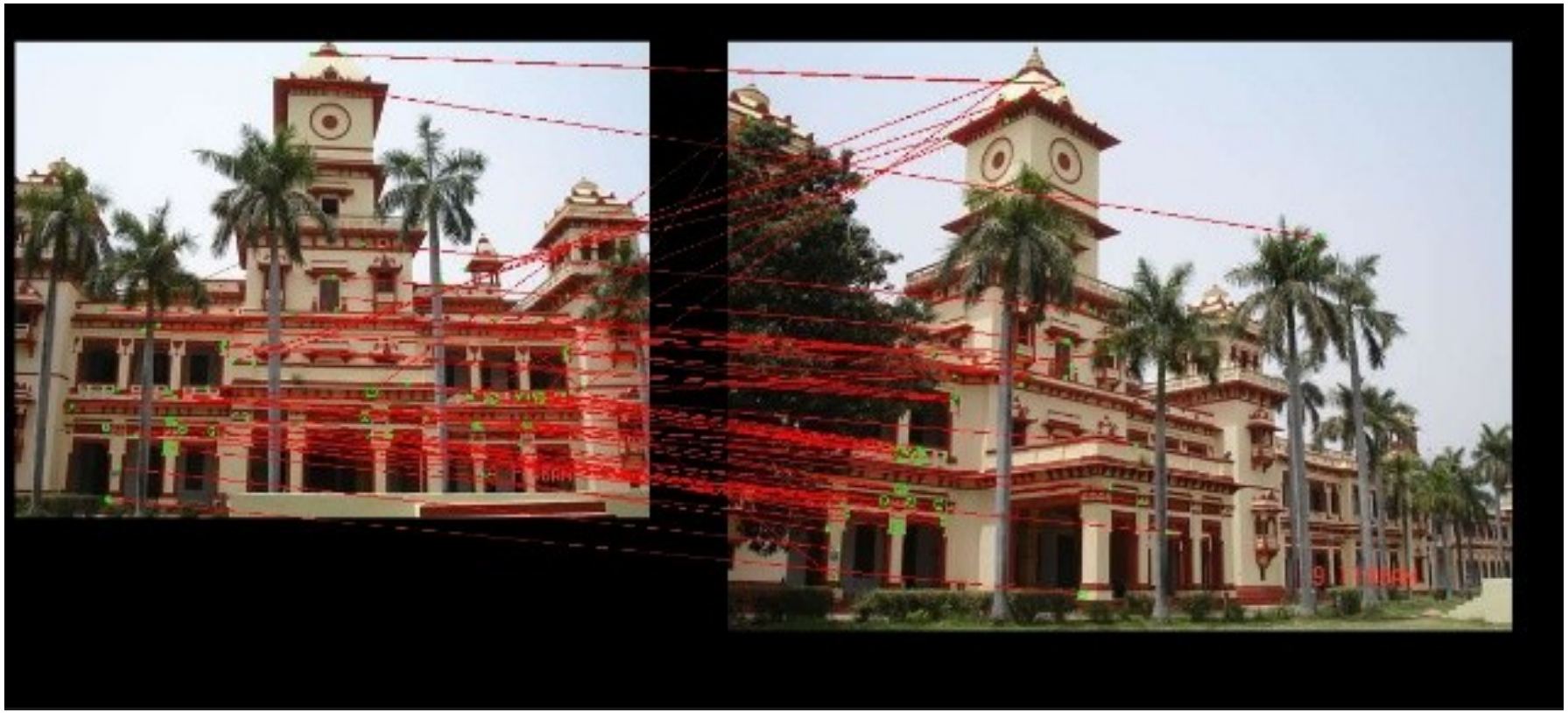
Jianbo Shi and Jitendra Malik

1999

Scale Invariant Feature Transform

David G. Lowe

Local image features invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection



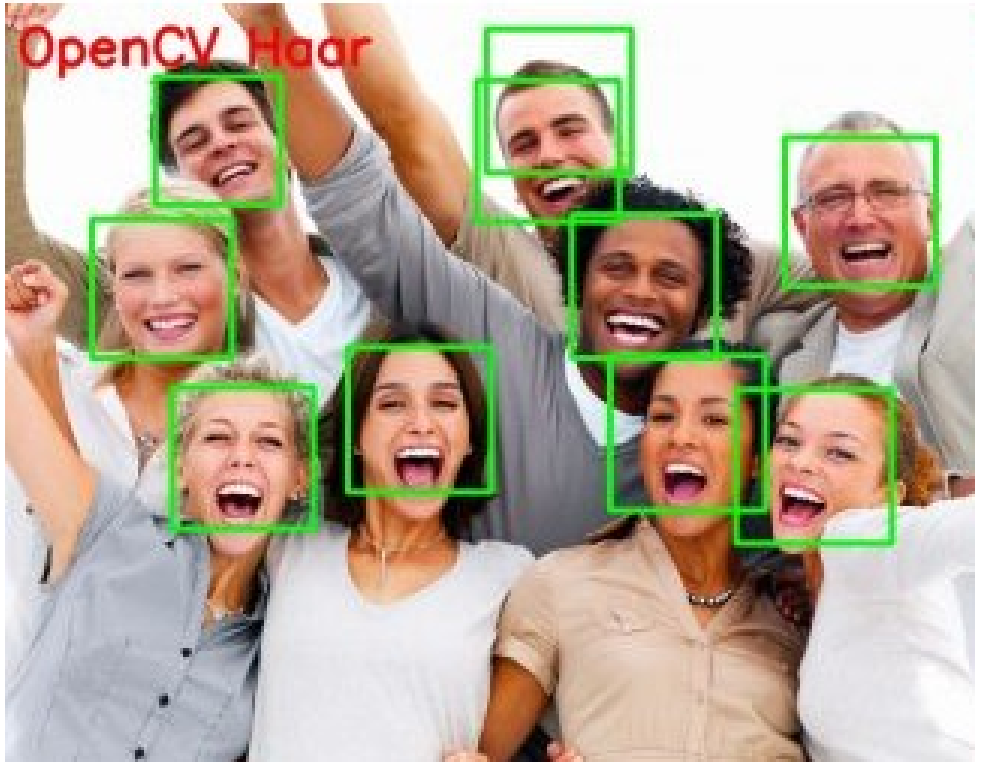
2001

Rapid object detection using a boosted cascade of simple features

Paul Viola and Michael Jones

Abstract

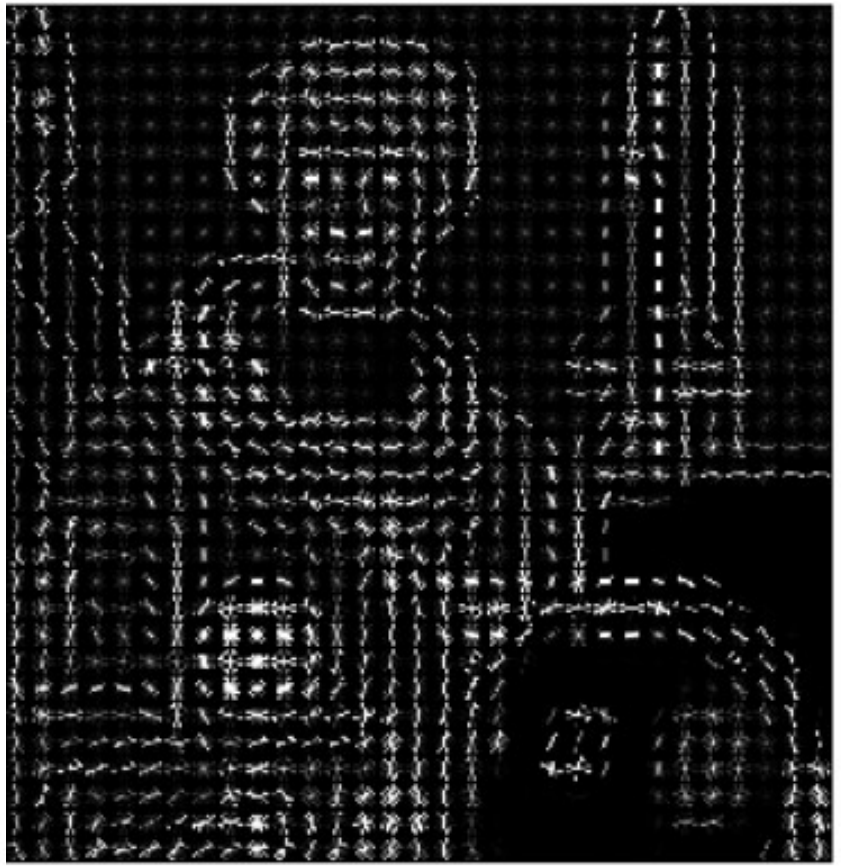
This paper describes a machine learning approach for visual object detection which is capable of processing images extremely rapidly and achieving high detection rates. This work is distinguished by three key contributions. The first is the introduction of a new image representation called the "Integral Image" which allows the features used by our detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features from a larger set and yields extremely efficient classifiers[6]. The third contribution is a method for combining increasingly more complex classifiers in a "cascade" which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. The cascade can be viewed as an object specific focus-of-attention mechanism which unlike previous approaches provides statistical guarantees that discarded regions are unlikely to contain the object of interest. In the domain of face detection the system yields detection rates comparable to the best previous systems. Used in real-time applications, the detector runs at 15 frames per second without resorting to image differencing or skin color detection.



Histogram of Oriented Gradients

2005

Navneet Dalal and Bill Triggs



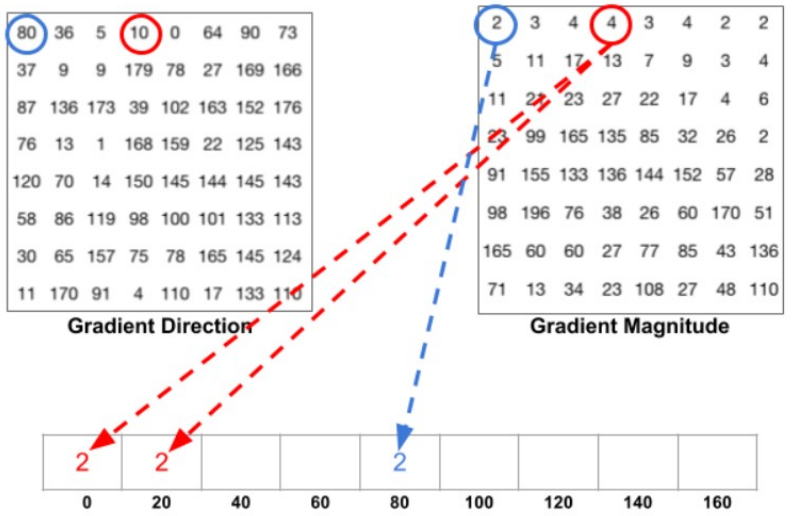
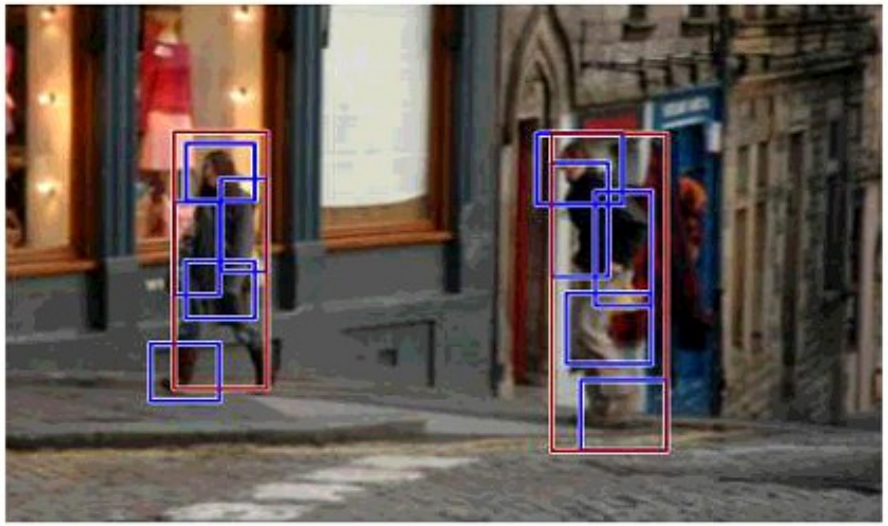


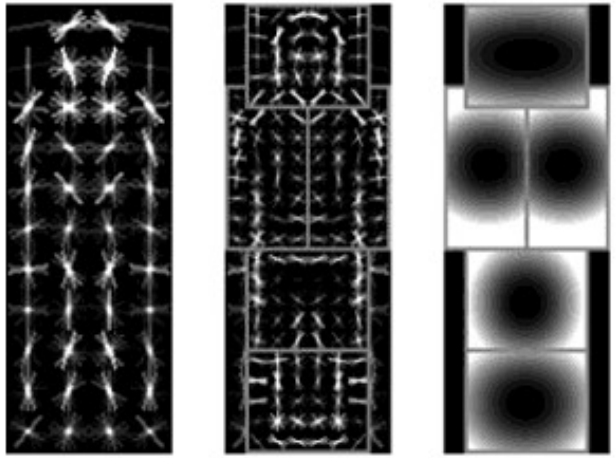
Image source: https://iq.opengenus.org/object-detection-with-histogram-of-oriented-gradients-hog/

2009

Deformable part models

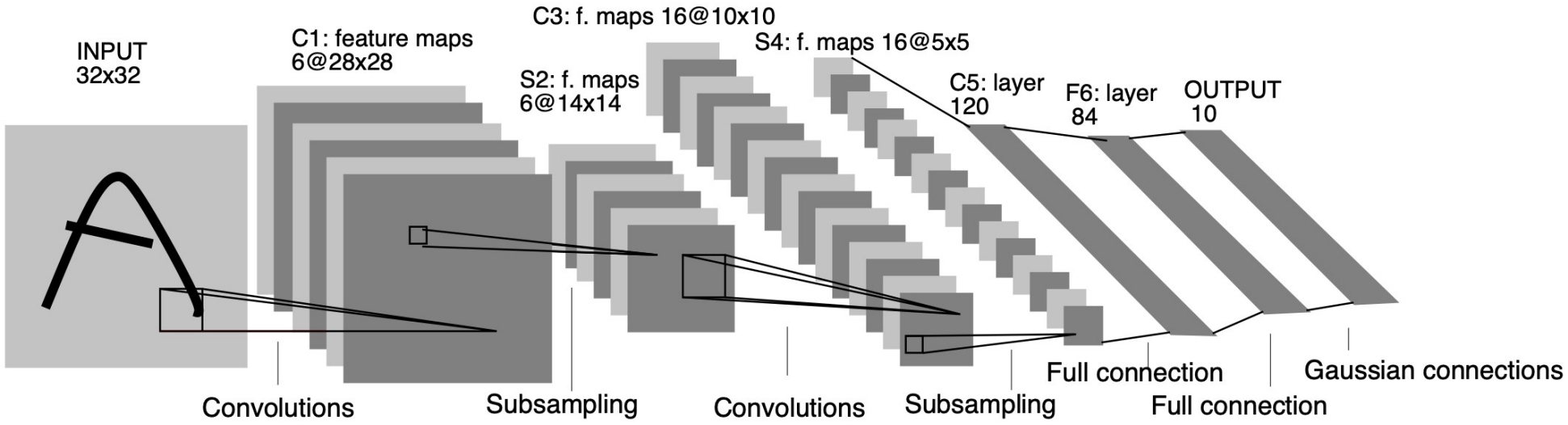
P. Felzenszwalb,R. Girshick, D. McAllester, D. Ramanan



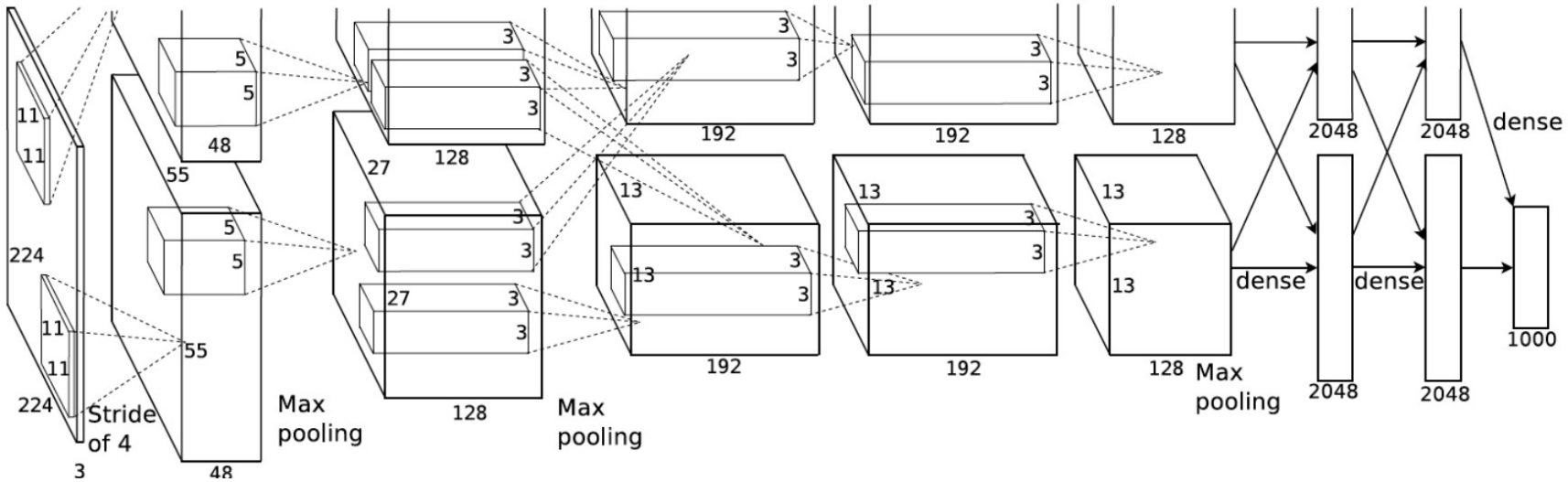


2012

LeNet (1998)

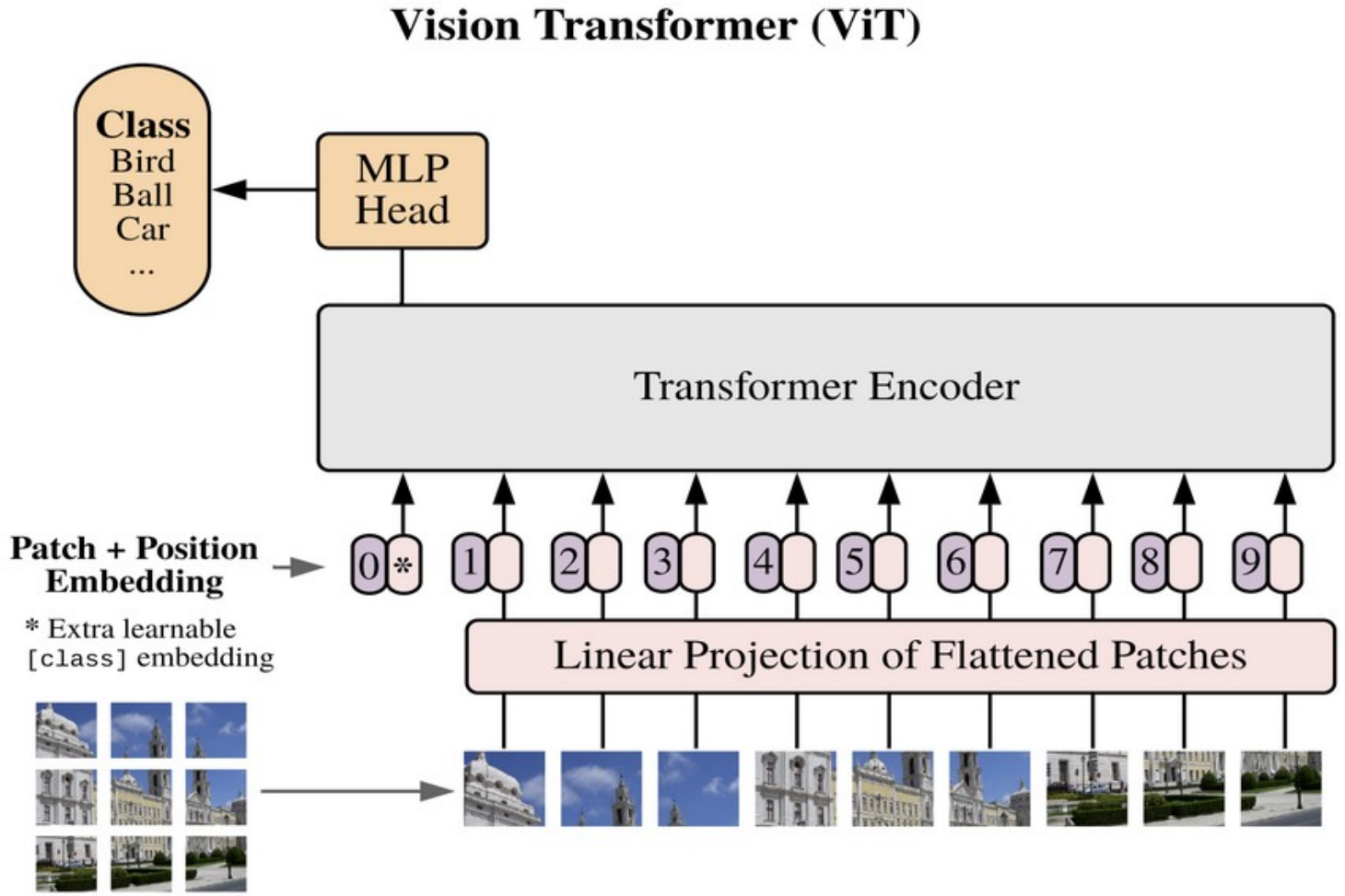


AlexNet (2012)



2021

Visual transformers



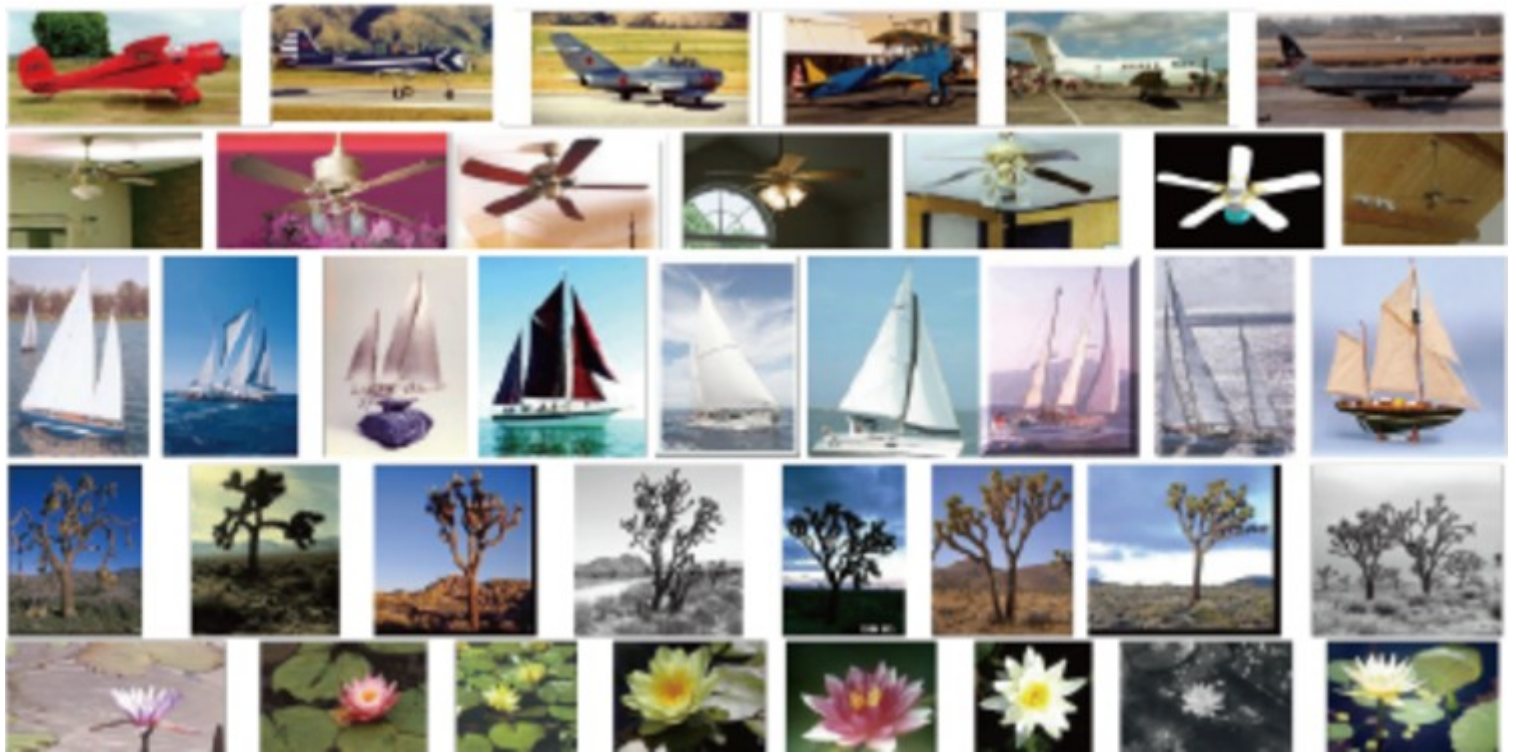
https://arxiv.org/abs/2010.11929

Data

2003

Caltech 101

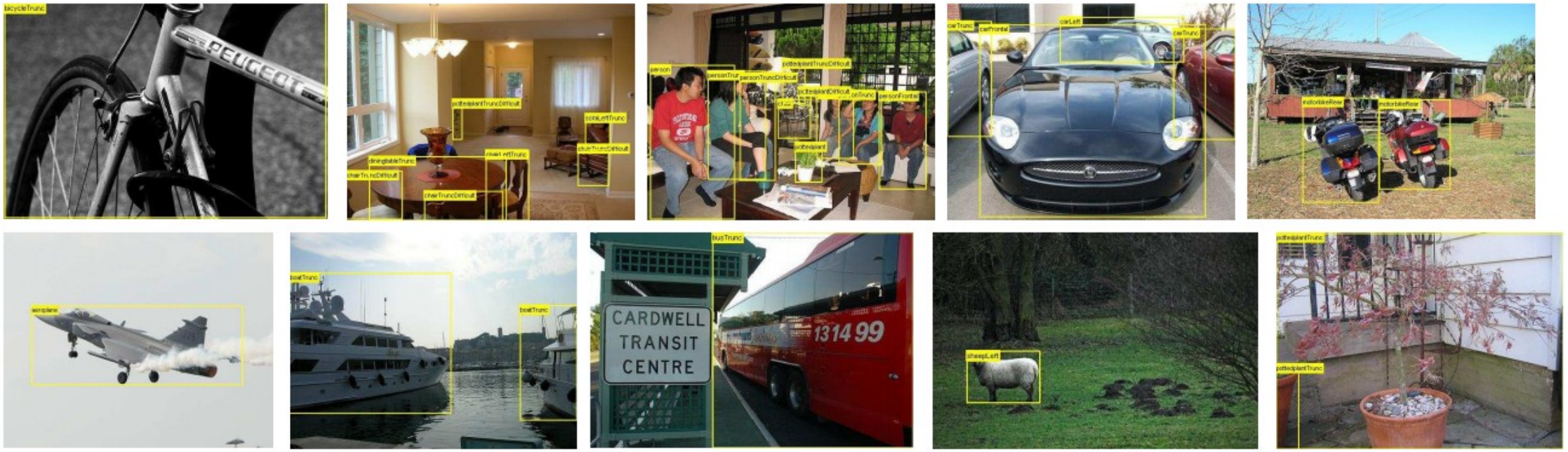
- 9,146 images, split between 101 distinct object categories and a background category
- 40 to 800 images per category



2007

Pascal VOC

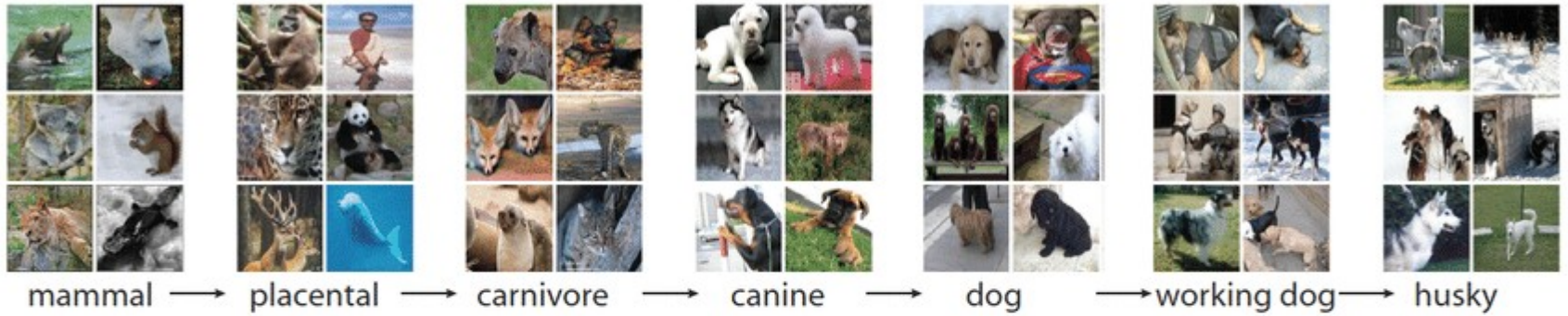
- 9,963 images containing 24,640 object split into 20 classes:
  - airplane, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, train, TV



2010

ImageNet

- > 14 million images organized into over 20,000 categories
- organized according to the WordNet hierarchy (nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images



2017

JFT

- internal dataset used by Google to train image classification algorithms
- the dataset has 300M images and 375M labels (multiple labels per image, on average each image has 1.26 labels)
- *noisy labels*: no human processing is involved
- labeled using an algorithm that uses complex mixture of raw web signals, connections between webpages and user feedback

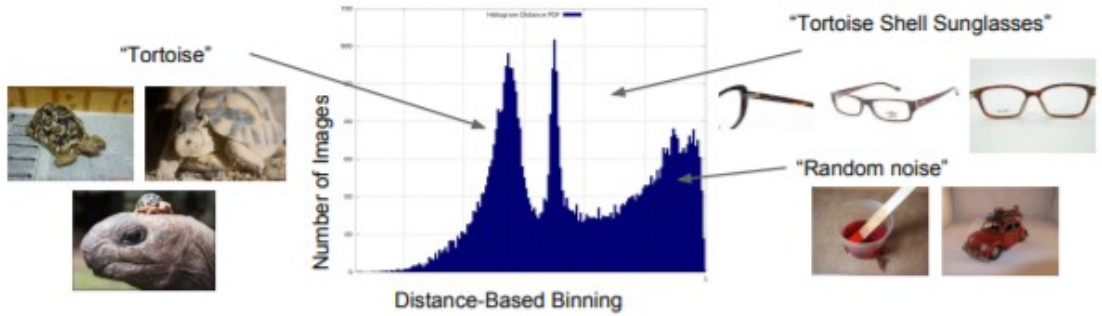


Figure 2. JFT-300M dataset can be noisy in terms of label confusion and incorrect labels. This is because labels are generated via a complex mixture of web signals, and not annotated or cleaned by humans. x-axis corresponds to the quantized distances to K-Means centroids, which are computed based on visual features.

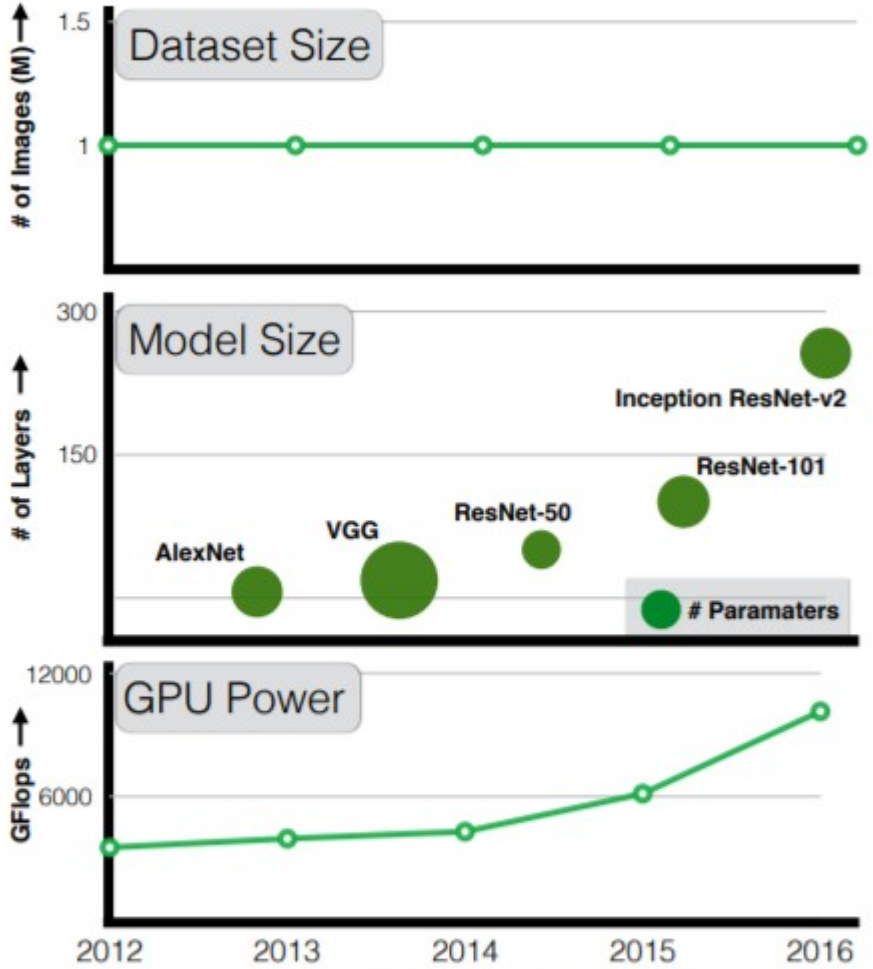


Figure 1. The Curious Case of Vision Datasets: While GPU computation power and model sizes have continued to increase over the last five years, size of the largest training dataset has surprisingly remained constant. Why is that? What would have happened if we have used our resources to increase dataset size as well? This paper provides a sneak-peek into what could be if the dataset sizes are increased dramatically.

https://arxiv.org/pdf/1707.02968v2.pdf

Computer vision- Where are we now?

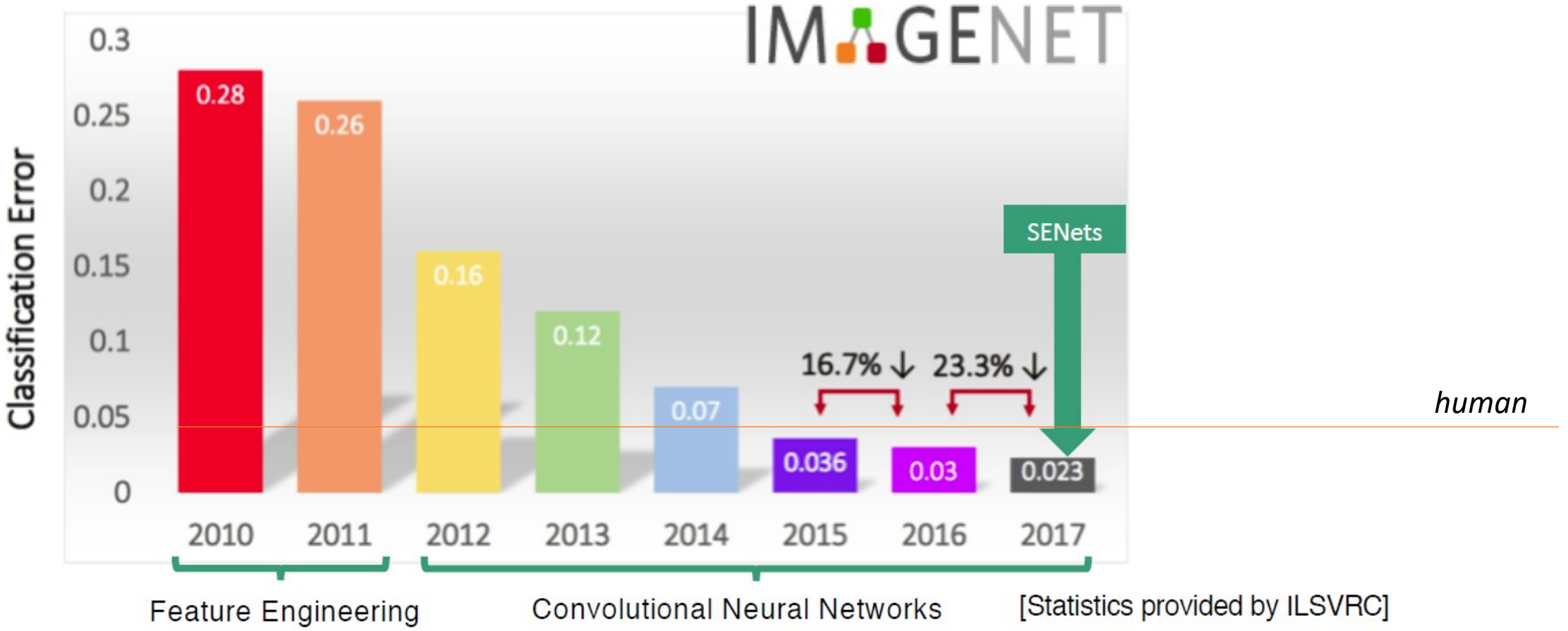
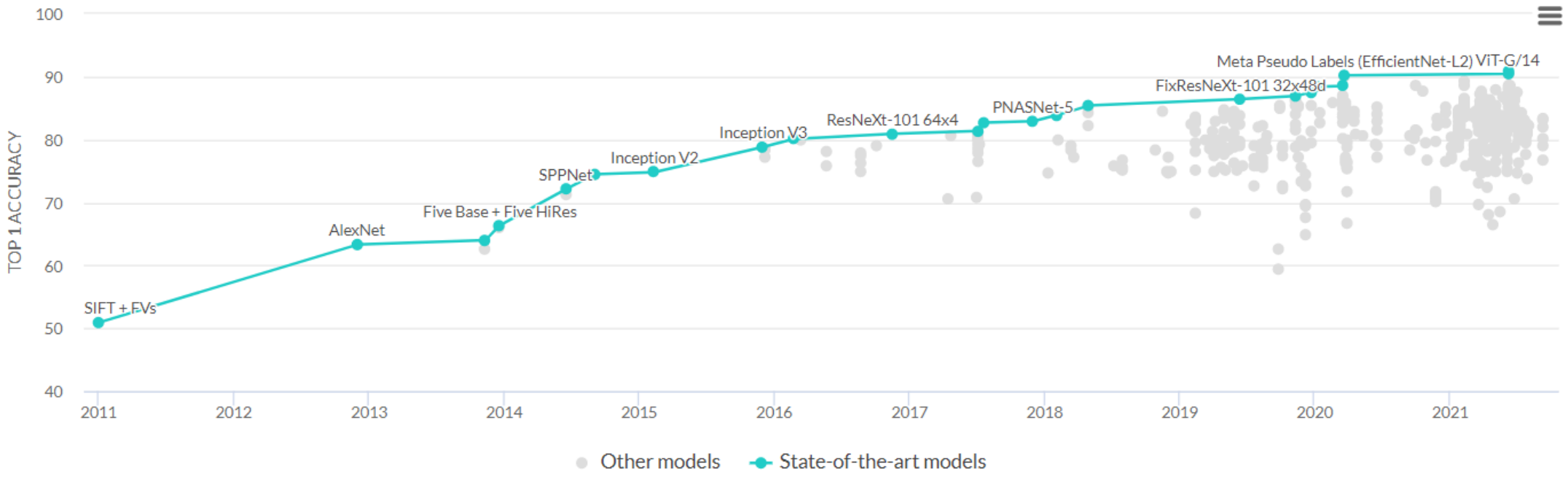


Image source: https://towardsdatascience.com/review-senet-squeeze-and-excitation-network-winner-of-ilsvrc-2017-image-classification-a887b98b2883

ImageNet accuracy

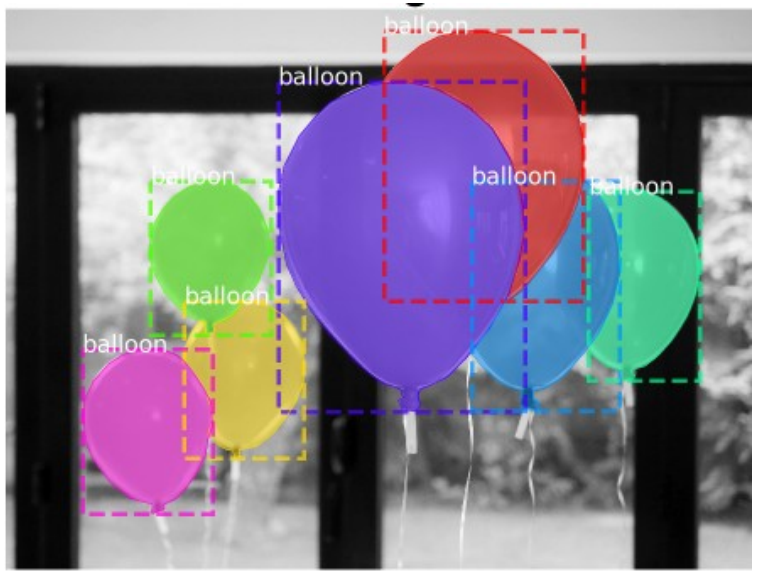


Computer vision applications I

Classification

Object Detection

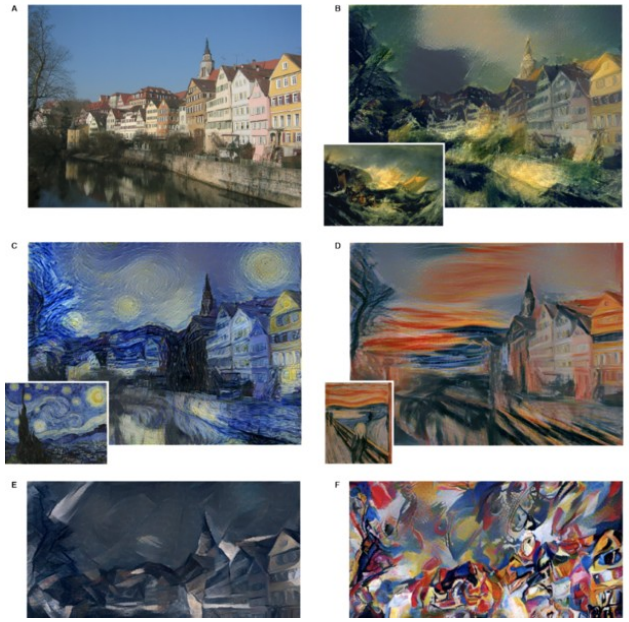
Semantic segmentation



Instance segmentation

Computer vision applications II

Image captioning



Style transfer

Action recognition

Computer vision applications III

Image generation

https://thispersondoesnotexist.com/

https://thiscatdoesnotexist.com/

https://edition.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html

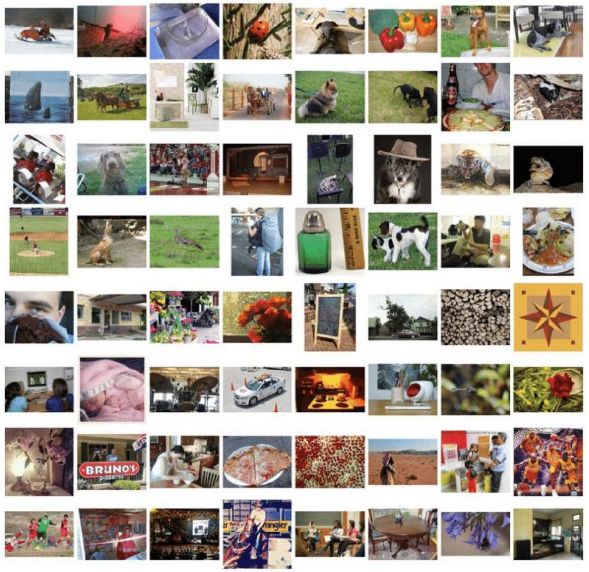
Computer vision applications IV

Image translation

Talking head models

Computer vision surpassed human vision

Captcha recognition



Object recognition

Face recognition

“We should all be playing games and let AI do all the work, instead AI plays games and we do all the work”

Andrej Karpathy

Computer vision vs. human vision

Example from Andrej Karpathy's blog: http://karpathy.github.io/2012/10/22/state-of-computer-vision/

Stanford AI index 2021

Industry has become, by far, the largest consumer of AI talent. 65% of graduating North American PhDs in AI went into industry—up from 44.4% in 2010.

AI investment in drug design and discovery increased significantly

Globally, investment in AI start-ups: from a total of $1.3B raised in 2010 to over $40.4B in 2018 (with $37.4B in 2019 as of November 4th), funding has increased at an average annual growth rate of over 48%

PRIVATE INVESTMENT in FUNDED AI COMPANIES, 2015-20
Source: CapIQ, Crunchbase, and NetBase Quid, 2020 | Chart: 2021 AI Index Report

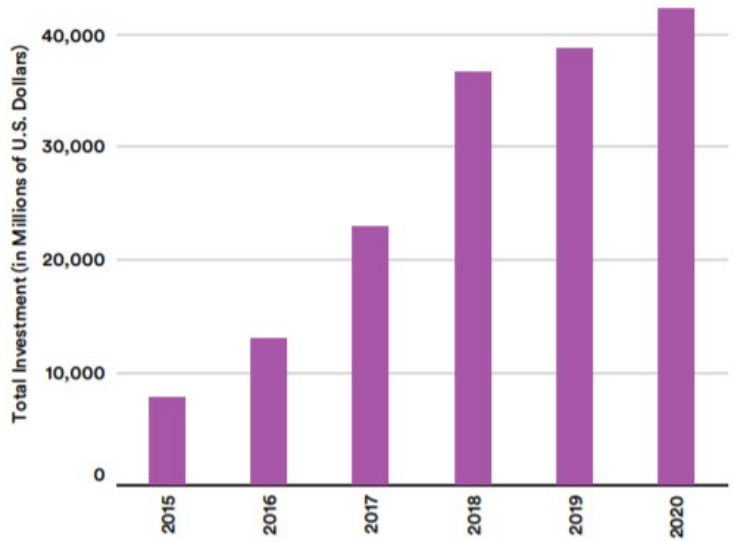
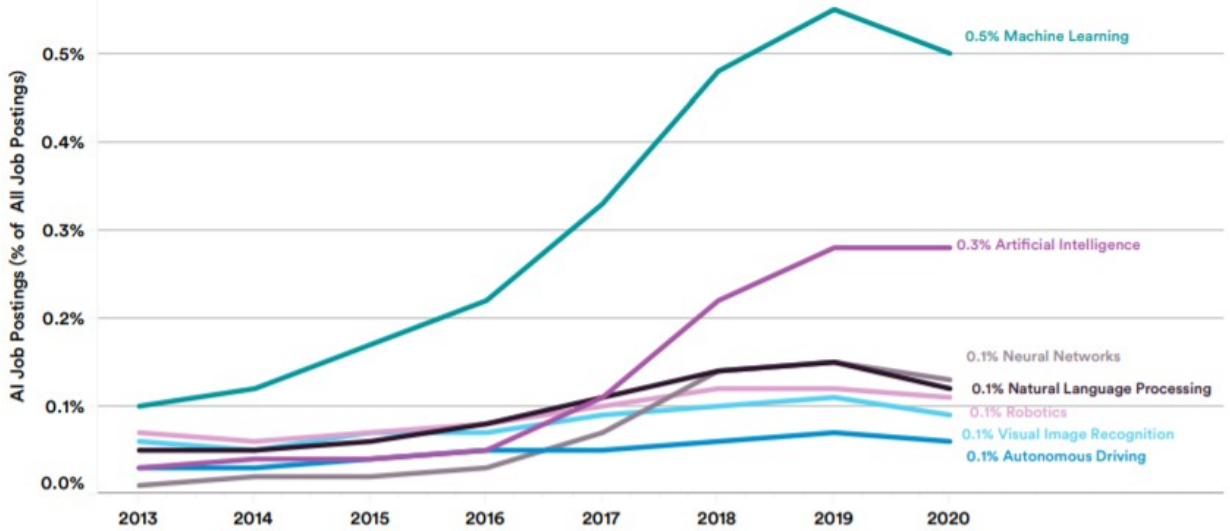


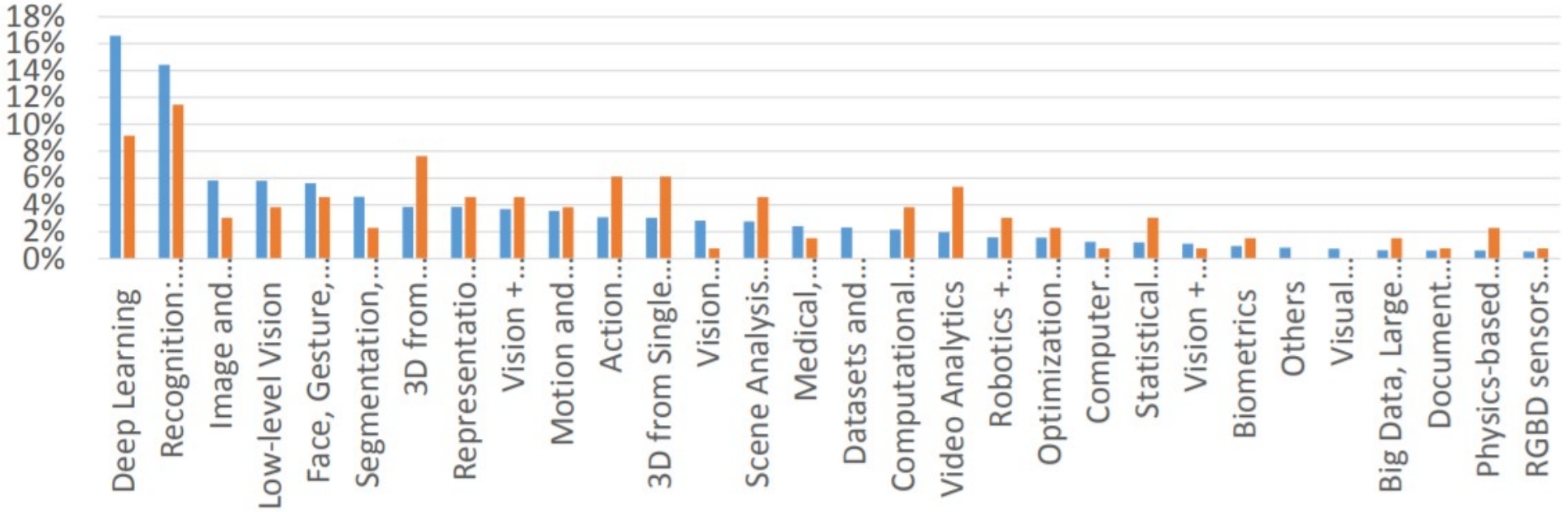
Figure 3.2.2

AI JOB POSTINGS (% of ALL JOB POSTINGS) in the UNITED STATES by SKILL CLUSTER, 2013-20
Source: Burning Glass, 2020 | Chart: 2021 AI Index Report

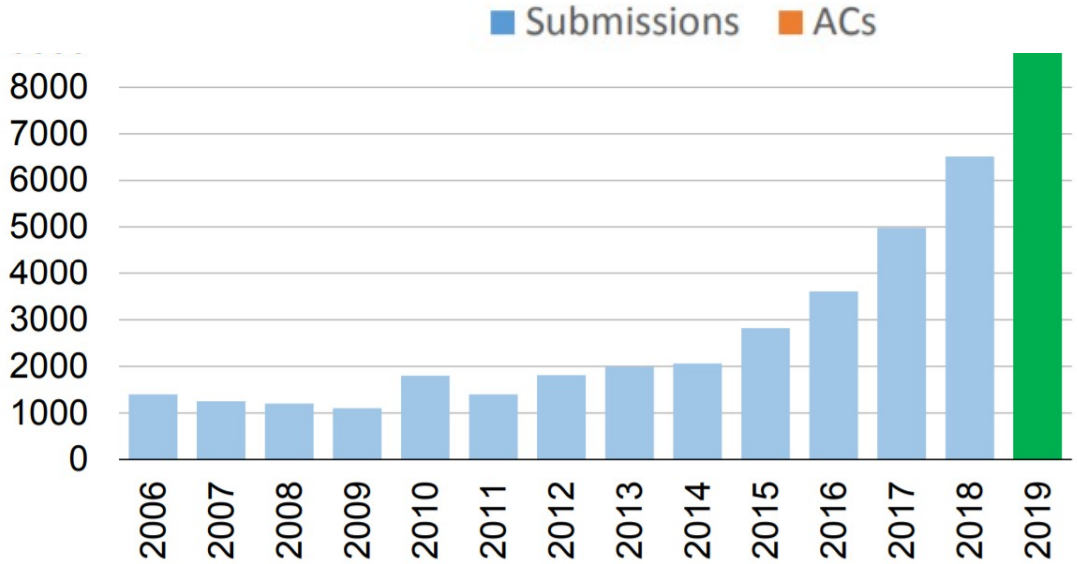


https://aiindex.stanford.edu/report/

CVPR 2019



Topics distribution



Number of attendees

Start-ups developing Computer Vision

Ceres Imaging combines high-resolution aerial imagery with advanced analytics to help farms, orchards, and vineyards optimize irrigation strategy and performance

Pixee Medical has created the first computer-assisted orthopaedic surgery solution using augmented reality to support total knee arthroplasty.

OrCAM: Wearable device, for people who are blind or visually impaired, that reads text, recognizes faces, identifies products & more

" Out-of-stocks" detection - FocalSystems

Valuable property attributes combined with a simple API allow investors and insurers to better select properties, evaluate risk, and streamline underwriting processes.

https://www.forbes.com/sites/robtoews/2021/02/28/a-wave-of-billion-dollar-computer-vision-startups-is-coming/?sh=21ce139b3f4c

Course overview

Part 1: Fundamentals. Artificial Neural Networks

Multi layer perceptron

Optimization. Loss functions

Neural Networks

Part 2: Convolutional neural networks and visual transformers

Convolutions. Convolutional layers. Pooling layers.

Recurrent neural networks

Visual transformers

Part 3: Applications and case studies

Object detection

Segmentation. Instance segmentation

Style transfer. Image generation

Visualizing and understating

Laboratories

1. Introduction to python, numpy. Getting to know each other.
2. Implement a softmax classifier (from scratch). Evaluating a model.
3. Implement a neural network in keras. Hyperparameter tunning.
4. Convolutional neural networks, transfer learning, fine tuning.
5. Segmentation, some advanced keras features.
6. Recurrent neural networks, visual transformer.
7. Neural network visualization techniques.

Projects

- Work in teams: 2-4 students
- From week 6, individual meetings with each team.
- Deadlines
  - Week 3: project proposal
  - Week 6: state of the art, find datasets to train and test on, (maybe) run some of the state of the art works and analyse their results
  - **Week 9: bring your own contribution!**
  - **Week 12: improve your first solution, more experiments, documentation**
  - Week 14 (lecture): have our own CV conference. 3 min teaser presentations, vote for the best projects
- **Documentation**: https://www.overleaf.com/latex/templates/cvpr-2018-template/qgmrfntfbqns
- **Useful tips (Zachary Lipton):** https://www.approximatelycorrect.com/2018/01/29/heuristics-technical-scientific-writing-machine-learning-perspective
  /

Course logistics

- Online lectures:
  - Microsoft Teams
  - Build a virtual community among the students
  - Channels:
    - Funny stuff
    - Research articles
    - Lab questions
    - General announcements

Grading

- Written examination: 40%
- Project and lab: 60 %

The final grade (average between written exam and project) should be at least 5

Recommended resources

- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
  - http://www.deeplearningbook.org/
- Langr, Jakub, and Vladimir Bok. GANs in Action. (2018)
- deeplearning.ai
  - https://www.deeplearning.ai/

I hope you’ll enjoy this course ☺