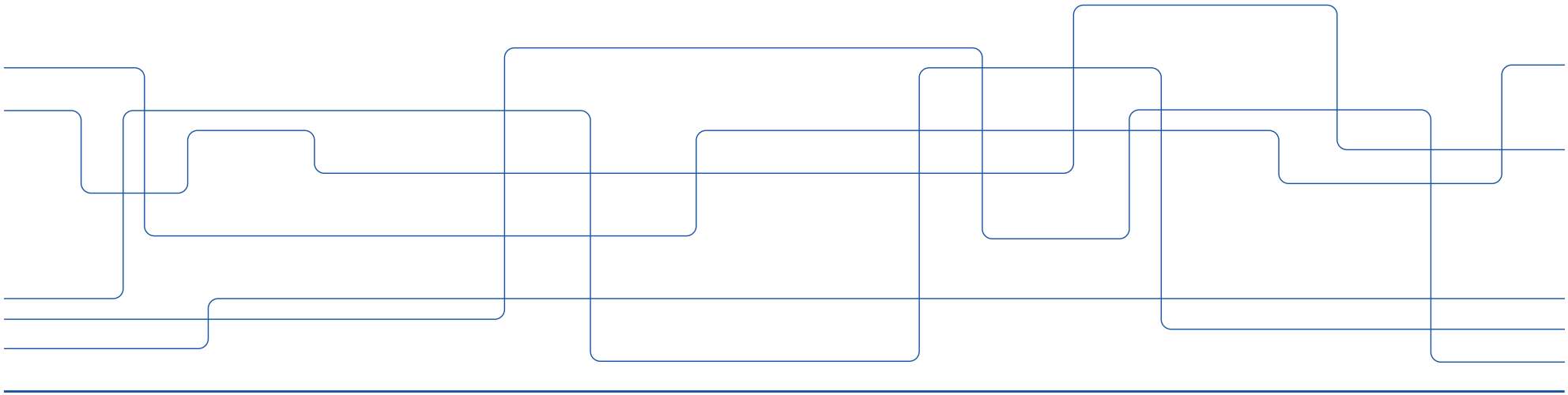
KTH ROYAL INSTITUTE
OF TECHNOLOGY

Introduction to computer vision

Mårten Björkman



General course information

This year the course will run in hybrid mode, with labs online, but lectures and exam on campus.

- 7.5 hp course (labs 4.0 hp, exam 3.5 hp)
- Course Web in Canvas under course code DD2423
- 2-3 lectures a week
- 16 lectures in total (3 exercise sessions)
- TAs: Wenjie, Marcel, Nona, Zehang, Alberta, Alfredo, Ci, Alice, Fanxuan, Giovanni, Rafael, and Tawsiful.
- If you have questions: preferably use Canvas.

General course information

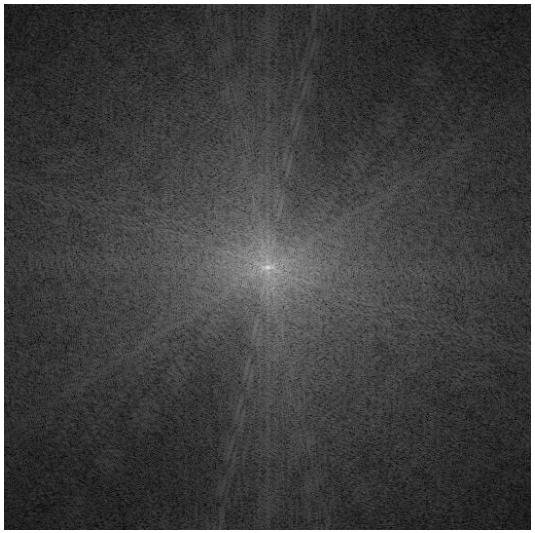
- The course is a broad introduction to computer vision, including image processing and image analysis.
- The goal is for students to ...
  - learn about basic concepts, terminology, models, and methods in computer vision,
  - become acquainted with common methods in computer vision by developing and evaluating them in practice.
- The focus is on ...
  - general problems in computer vision (segmentation, recognition, feature detection, stereo matching, etc),
  - the theoretical basis behind those problems, with
  - examples of new, often deep learning-based, and traditional methods to solve them.
- Note: The course is NOT a course in deep learning.

Assessment

- There are 3 labs (LAB1), 1 online quiz, and 1 written exam (TEN1)
  - Online quiz for grade E, and written exam for grades A-D
- Grading: A-F
  - Final grade: average of exam and labs, rounded towards exam
  - Labs grade: average of labs, rounded towards nearest grade
- Labs are done in Matlab (or Python), possibly on your own laptop.
- There are scheduled times for labs:
  - Help: ask for help at queue.csc.kth.se
  - Presentation: book a slot in Canvas - no help!
  - Primarily in Zoom, or in person if requested
- Doing labs before the deadline - up to 3 pts on the exam

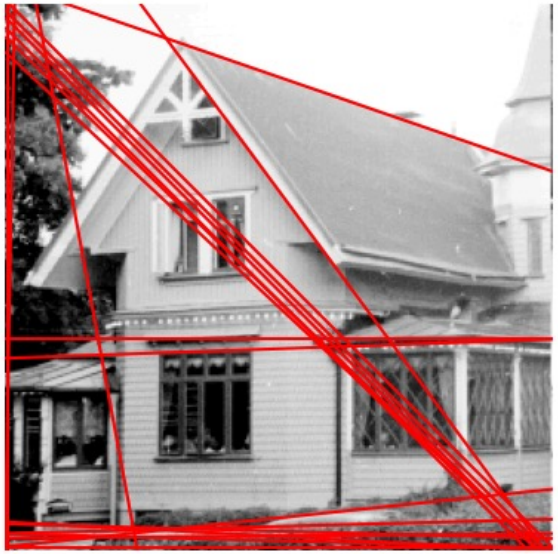
Lab 1: Filtering operations

- Learn about image filtering and the effect of image noise.
- Frequency based representations using the Fourier Transform.



Lab2: Edge detection & Hough transform

- Learn about how to detect edges and lines in images.
- Get your hands dirty by building something from scratch.



Lab3: Image segmentation

- Learn about how to divide an image into image regions.
- Study different methods and understand their challenges.

Matlab or Python? – That is the question!

- Matlab Pros:
  - Everything is in one environment.
  - Little fuss with image formats.
- Matlab Cons:
  - Not the natural environment for machine learning.
  - More restricted to the scientific community.
- Python Pros:
  - Dominating language for machine learning.
  - Allows for a lot of flexibility.
- Python Cons:
  - Need for additional packages: `NumPy, SciPy, Matplotlib.`
  - Different packages represent images in different ways.

Lab presentations

- What to do for each lab:
  - Book a slot for presentation in Canvas in the Calendar.
  - Go through the lab instructions!
  - Implement the required functions and run experiments.
  - Answer the questions in the attached answer sheet
  - Upload the following to Canvas in a zip file:
    1. *All your code from the lab*
    2. *A Matlab/Python script that steps through the lab*
    3. *Filled in answer sheet*
  - Present your lab online using Zoom
- Start to work on labs as soon as possible!
- Don’t over-do the answer sheets! Consider it as notes for yourself.
- Think what you learned from the lab, not what you did!

Lab grading

- All labs can be done in pairs, but examined individually.
- A cumulative definition of grades:

  E - Lab completed, but many written answers not correct.

  D - Some written questions have not been answered correctly.

  C - Minor difficulties in presenting lab results and responding to oral questions posed by TAs.

  B - No difficulties in presenting lab results and responding to oral questions posed by TAs.

  A - Is able to reason about questions beyond the scope of the lab.

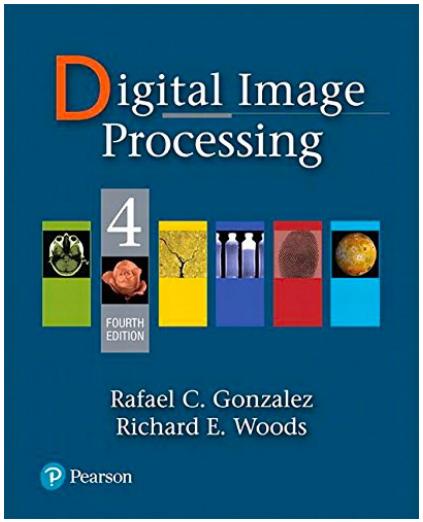
- More detailed formal definition on the Canvas page.
- Good idea: Present to each others for practice!

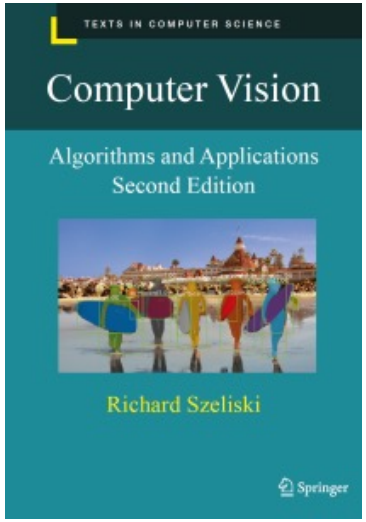
Additional ungraded quizzes for feedback

- Every week quizzes will be posted on Canvas
  - Should not take more than 10–15 minutes to complete
  - Quizzes are recommended, but not compulsory
- Quizzes provide feedback:
  - For you to test your degree of understanding
  - For me to know what requires rehearsal
- Recommendation:
  - After each week, do the corresponding quiz
  - Before attending the exam, redo the quizzes

Recommended books

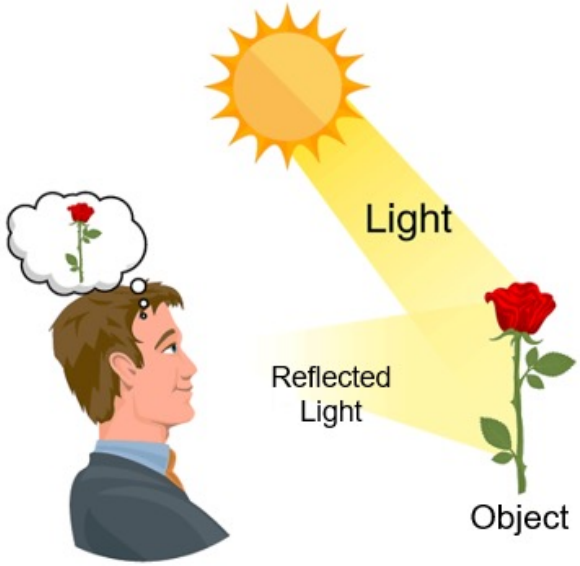
- R. Gonzalez and R. Woods: “Digital Image Processing, 4e edition”, Pearson, 2018.
- R. Szeliski: “Computer Vision: Algorithms and Applications”, Springer, 2022 (available for free: http://szeliski.org/Book)





- Note: the books can give you a second opinion on topics, but assessment is based only on lecture and lab notes.

What does it mean to see?



- We say that we ‘see an object’, but we really see light reflected on the object.

  Reflected light = Incoming light + Object properties

- Vision is an active process for deriving efficient symbolic representations of the world from the light reflected from it.

Vision as an active process!

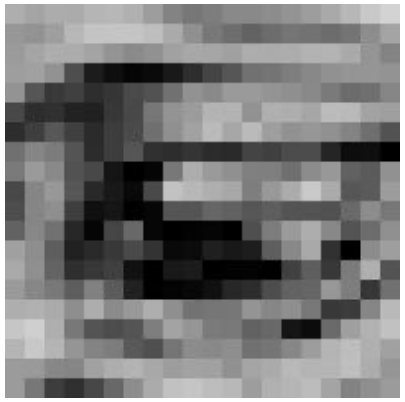
Active

- In nature seeing is always (?) associated with acting.
- Acting can simplify seeing, e.g. move your head around an object.
- A computer vision system may control its sensory parameters, e.g. viewing direction, focus and zoom.

Process

- No “final solution”. Perception is a result of continuous hypothesis generation and verification.
- Vision is not performed in isolation, it is related to task and behaviours.

But what does a computer see?



174	164	172	171	181	183	177	157	131	119	125	137	137	141	147	157	169	184	200	211
151	147	159	183	196	197	197	190	171	144	121	110	99	109	117	114	109	117	142	164
155	165	174	172	164	148	138	145	161	168	166	161	173	172	173	179	180	170	149	132
158	161	146	78	28	15	6	10	25	46	69	86	113	117	126	136	145	149	149	147
147	86	57	64	46	54	74	82	113	146	138	154	154	148	144	137	122	108	112	134
72	46	49	35	33	56	73	119	148	156	180	179	146	176	174	166	187	186	163	159
46	63	88	64	43	61	75	112	144	162	187	167	186	158	146	151	145	132	112	83
112	134	124	84	44	61	85	48	45	53	44	42	68	64	76	76	37	13	15	0
122	154	138	82	72	21	0	5	120	173	154	144	128	127	112	140	152	97	91	145
73	134	153	87	41	25	10	74	194	184	175	168	126	153	172	201	171	87	91	156
78	134	143	63	16	22	1	6	85	85	103	97	87	94	84	84	82	106	141	120
98	135	114	39	5	70	105	6	0	3	5	0	79	125	154	173	136	84	108	144
160	142	94	33	42	57	74	51	9	12	32	24	29	95	137	169	49	0	140	158
169	144	100	47	55	64	18	0	25	28	0	3	4	0	65	95	56	124	176	84
154	136	122	94	129	92	34	1	4	8	17	39	45	78	99	102	152	126	55	113
174	177	171	142	173	139	153	174	143	115	118	119	148	134	143	120	56	60	134	185
198	207	175	120	84	77	85	126	163	150	123	121	98	105	21	12	106	159	178	193
180	200	181	149	132	125	90	80	117	140	137	142	113	118	135	157	180	190	176	161
155	154	136	134	161	174	180	168	157	162	161	145	165	171	185	180	181	191	173	149
155	119	61	48	71	104	141	171	193	197	189	184	157	177	181	188	164	131	143	155

- A computer just sees a bunch of numbers, organised in a two-dimensional array.
- For colour images you have, you have three values per pixel (red, green, blue).
- For video sequences you have an additional dimension corresponding to time.

Can computers match human vision?

- Computers can be better at “easy” things (when things look the way the usually do)
- Humans are much better at “hard” things (when things are highly unusual)

But human vision has limitations - illusions

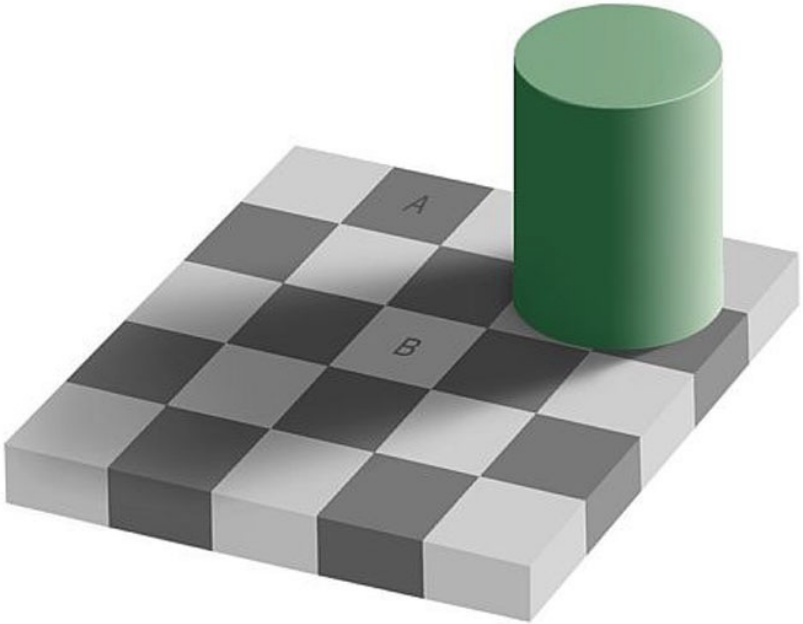
- If we look locally at an impossible object, it makes perfect sense, but globally it does not.
- We take multiple snapshots and integrate to a whole, i.e. vision is an *active process*.

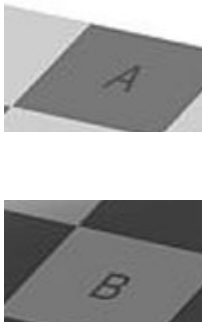
Depth illusion – size constancy



- Projective size = Object size + Viewing distance, but we tend to invert the problem and directly see the Object size.
- Context provides information on Viewing distance, which allows us to be fooled.

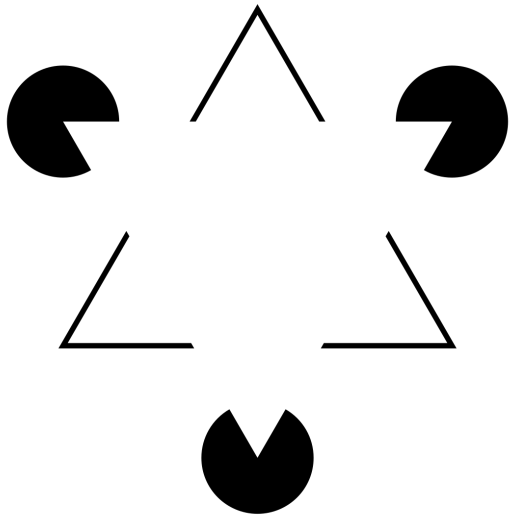
Colour illusion – colour constancy

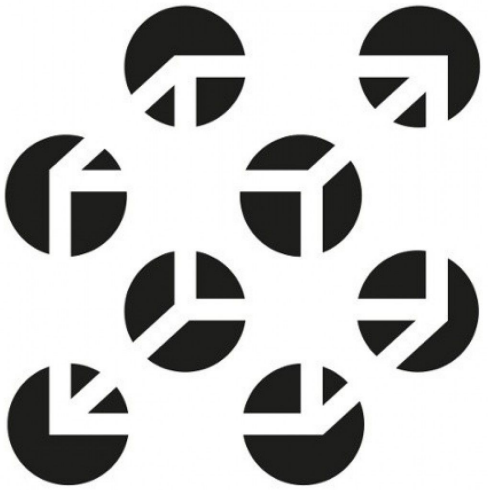




- Image colour = Object colour + Illumination, but (again) we tend to invert the problem and directly see the Object colour.
- Context provides information on Illumination, which (again) allows us to be fooled.

Subjective contours





- Using prior experience, we tend to ‘fill in the blanks’ with information that is not there.
- In many cases our interpretation is correct, but in other cases it is not.

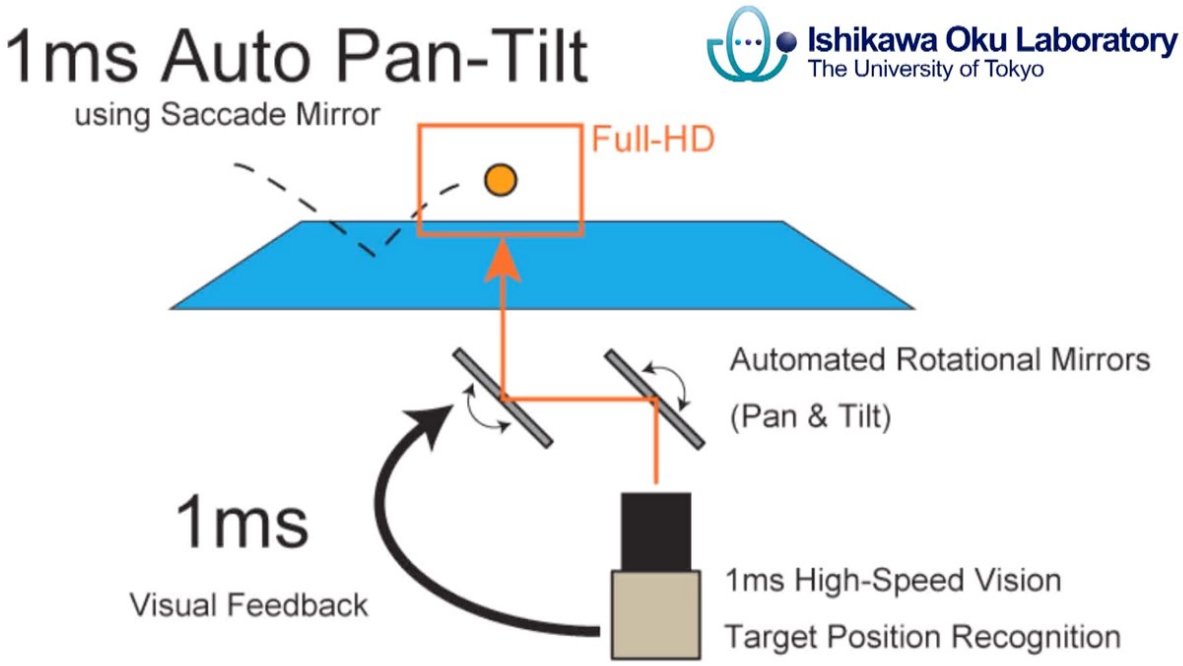
The importance of discontinuities

- We see discontinuities (edges) more than what is between the discontinuities.
- It is logical, since the edges must come from discontinuities in the real world, e.g. occlusions, surface orientations, illumination changes, surface texture.

Lessons learned from human vision

- Prior experience and context are very important for interpretation.
- Finding and interpreting edges should in many cases be enough.
- Our own vision system leads us to believe that it is easier than it really is.

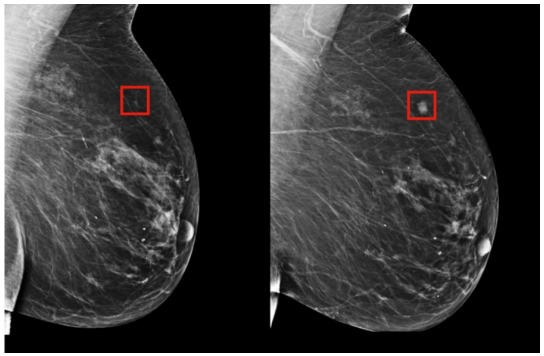
A 1000 Hz video camera



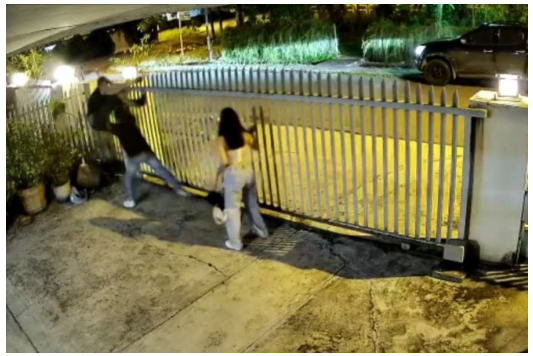
- Most video cameras run in 25-60 Hz, since we cannot observe faster changes anyway.
- But for machine vision, we do not need such a limitation. Why not 1000 Hz?

Why is computer vision relevant?

Safety



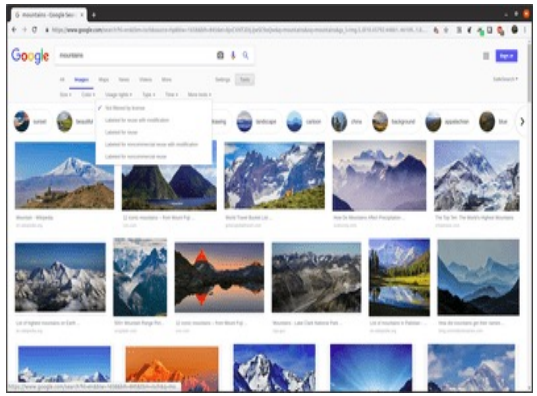
Health



Surveillence

Comfort

Entertainment



Access

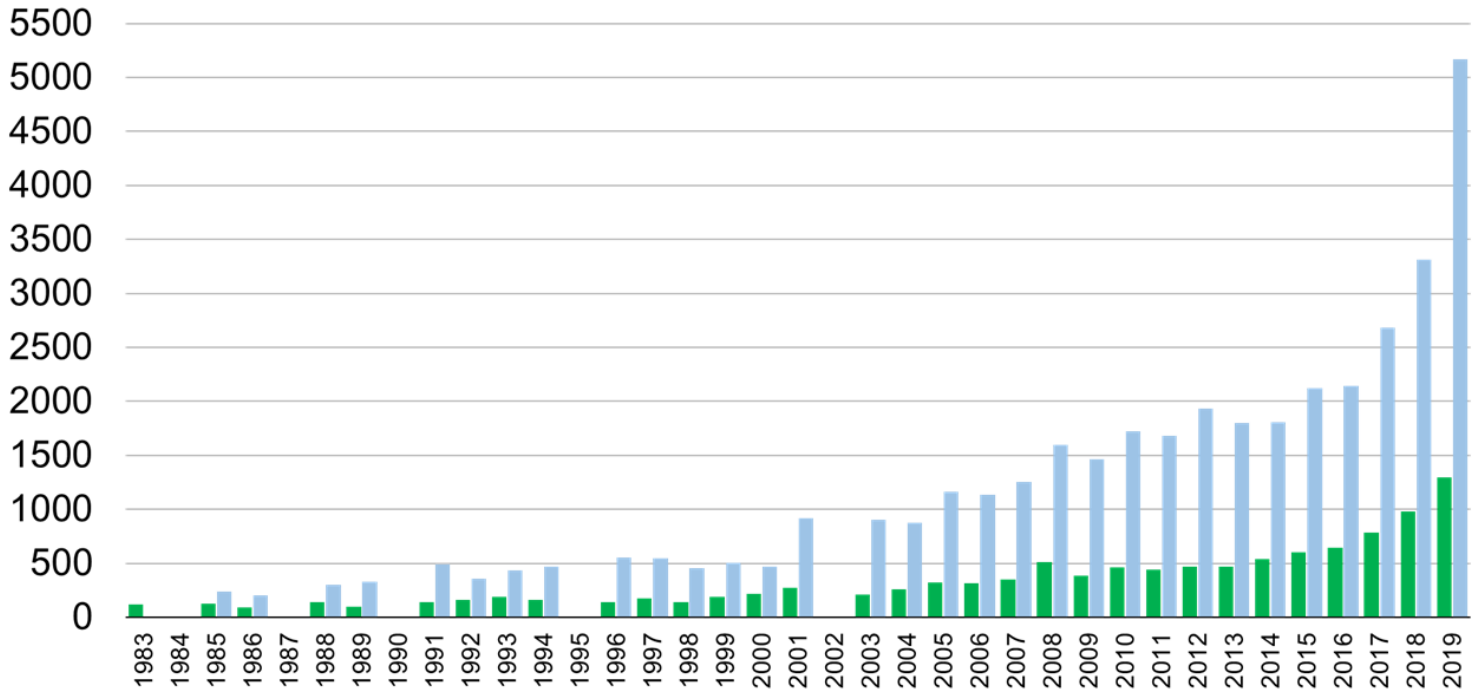
Example: Robotic tasks using computer vision

From experiments done at RPL in 2004, equally relevant today.

Why is computer vision interesting?

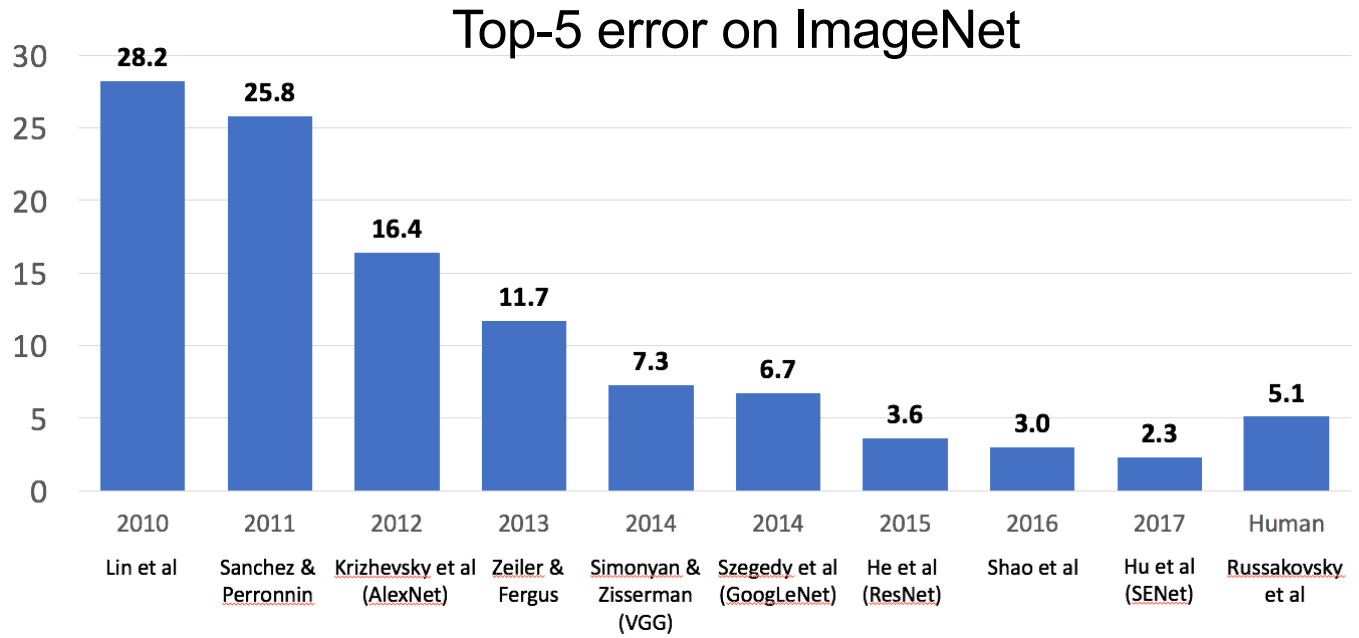
- Intellectually interesting
  - We are too good at it to appreciate how difficult it really is.
  - Going from 2D to 3D is an under-determined inverse problem.
- Psychology
  - After all, about 50% of cerebral cortex is for vision.
  - Vision is (to a large extent) how we experience the world.
- Engineering
  - It is fun to build ”intelligent” machines.
  - Computer vision opens up for multi-disciplinary work.
  - We have images everywhere and need ways to organize them.

The growth of computer vision



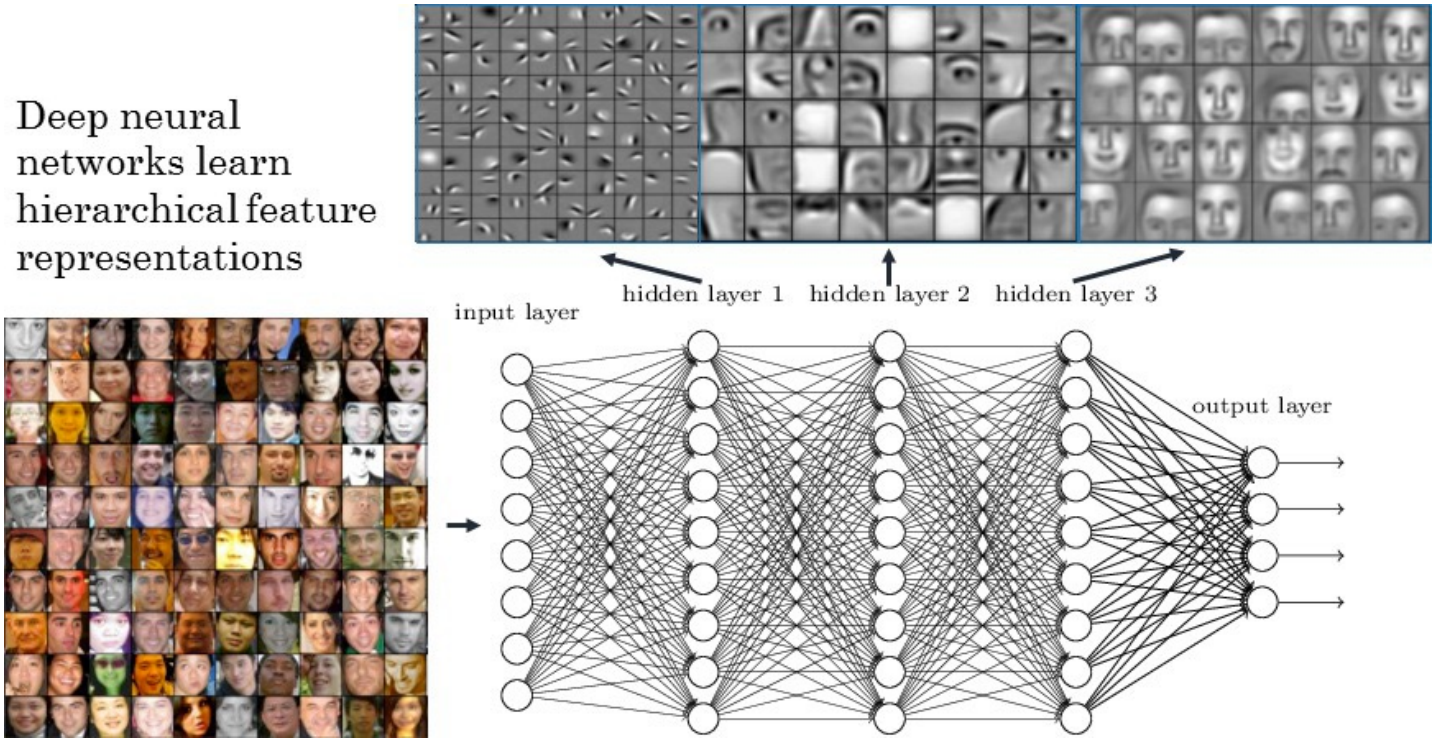
Number of submitted (blue) and accepted (green) papers at CVPR

Deep networks for image classification



- Neural networks were long forgotten in computer vision.
- But... deep neural networks have become state-of-the-art.
- The best networks today have an error rate of about 1%.

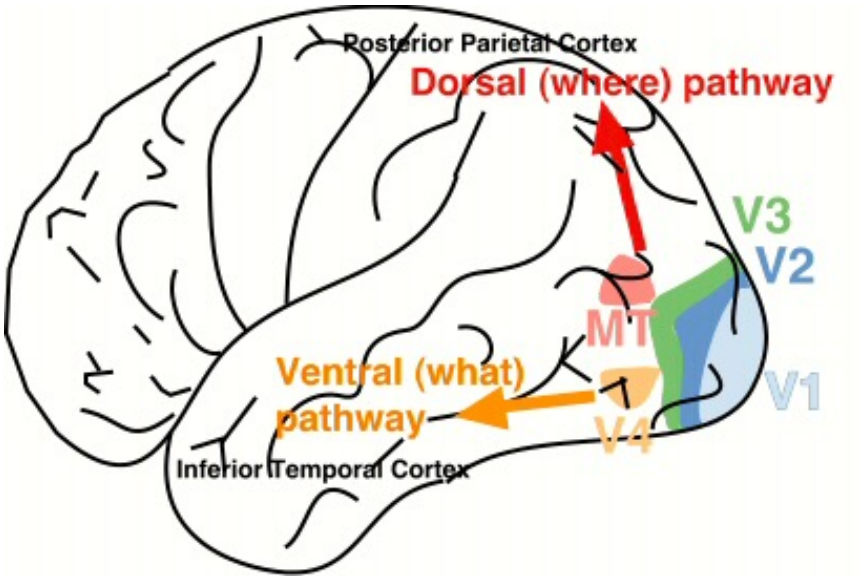
Fully Connected Neural Network (FCN)



- Neural networks typically consists on layers of neurons.
- Possibly images on inputs, some hypotheses on outputs.

What about deep learning?

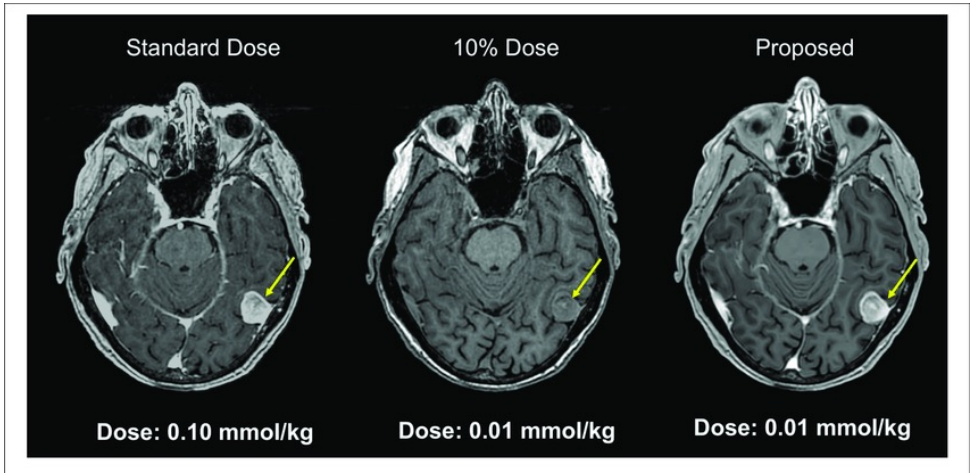
Visual context with *what* and *where* pathways



Deep learning can

- answer *what*-questions – but *where*-questions are also important.
- benefit from lots of data – but what if you don't have much data?

Image processing



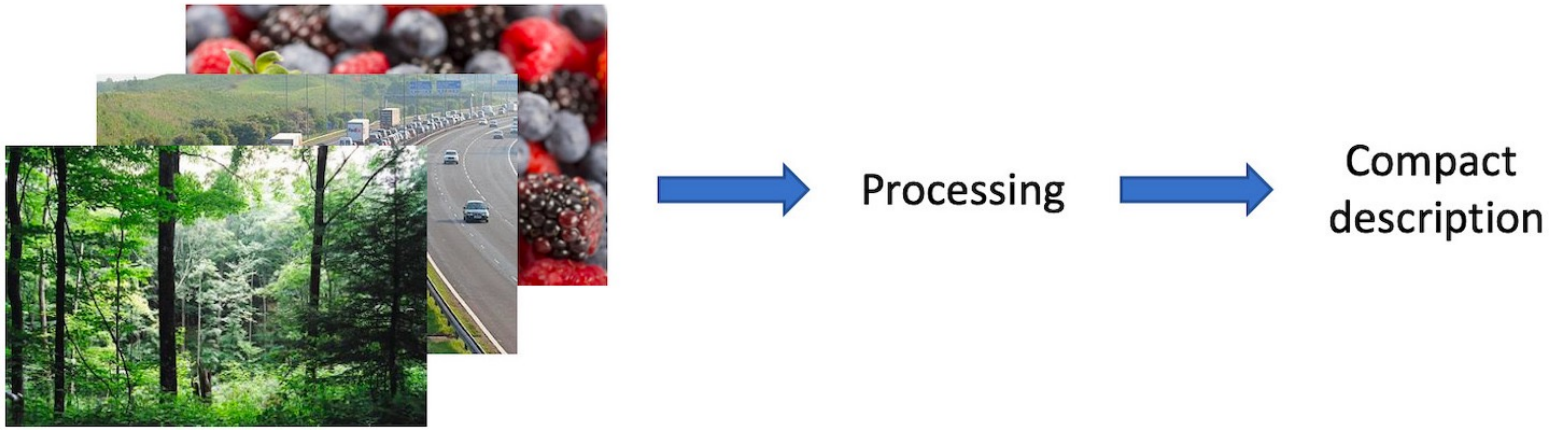
Purpose: Enhance important features & suppress disturbances (noise).

- Examples: Poor image data in medicine, astronomy, surveillance.

Subjects treated in this course:

- Image sampling, digital geometry
- Linear filter theory, the Fourier Transform
- Enhancement: grey-scale transform, image filtering

Image analysis



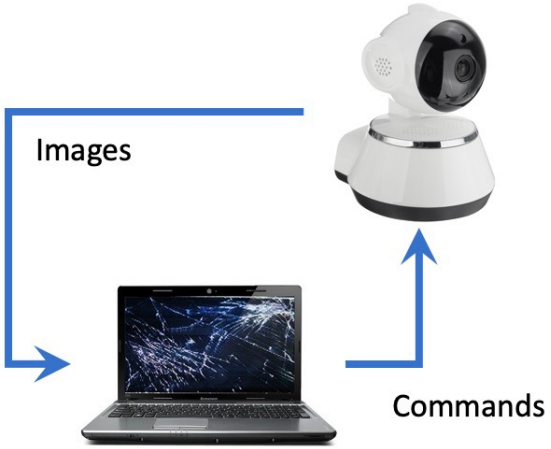
Purpose: Generate a useful compact description of the image.

- Example: Matching representations of objects to image data.

Subjects studied in this course:

- Image and shape representation
- Feature detection and matching
- Object and image segmentation

Computer vision



Purpose: Achieve an understanding of the world, possibly under active control of the image acquisition process.

Subjects studied in this course:

- Recognition and classification
- Stereo geometry
- Motion and optical flow

- However, whole field often called computer vision (incl. image analysis)

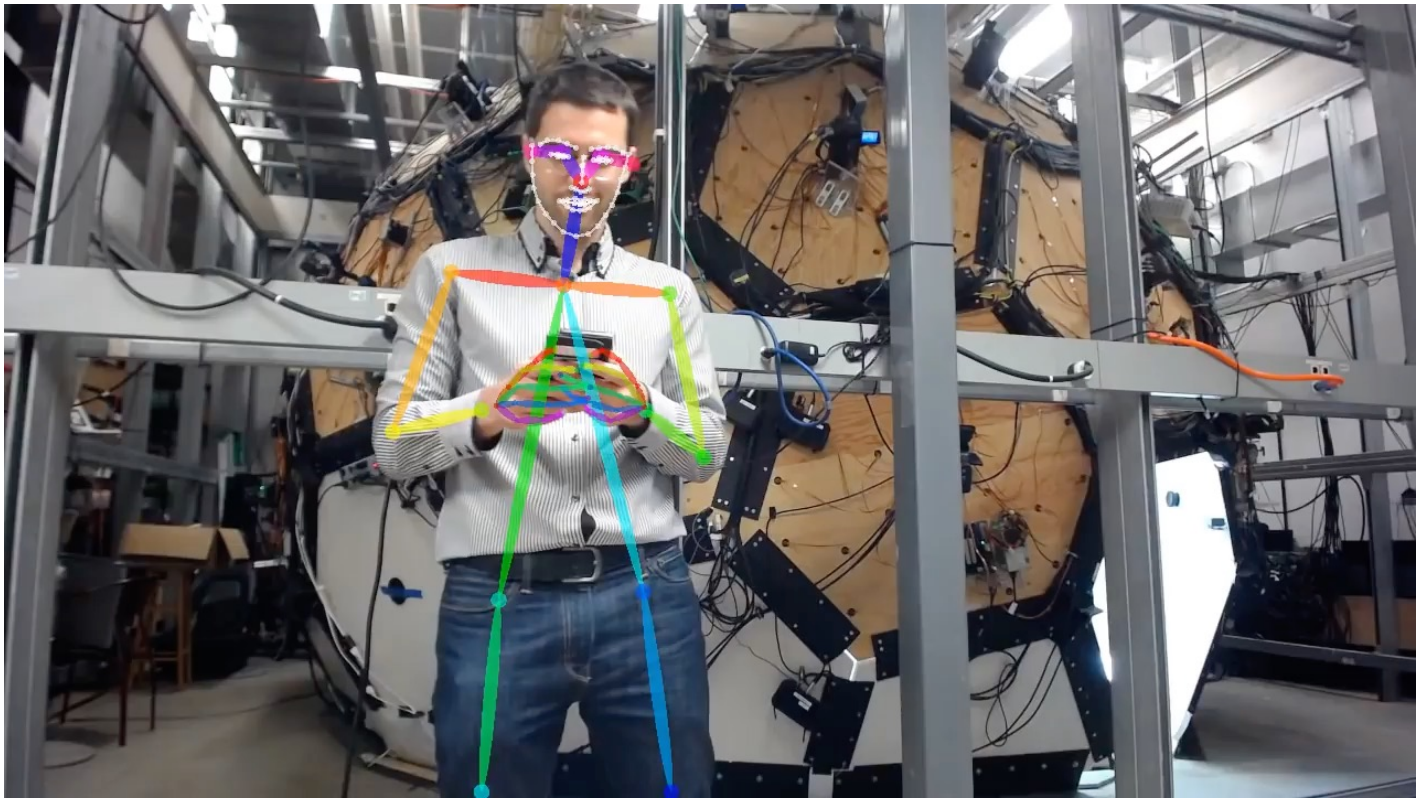
Example: Semantic segmentation

If you know what a scene usually consist of, you can exploit that for image segmentation.

Example: Google self-driving cars

This is what a Google self-driving car is seeing. Hopefully, it has been improved since then.

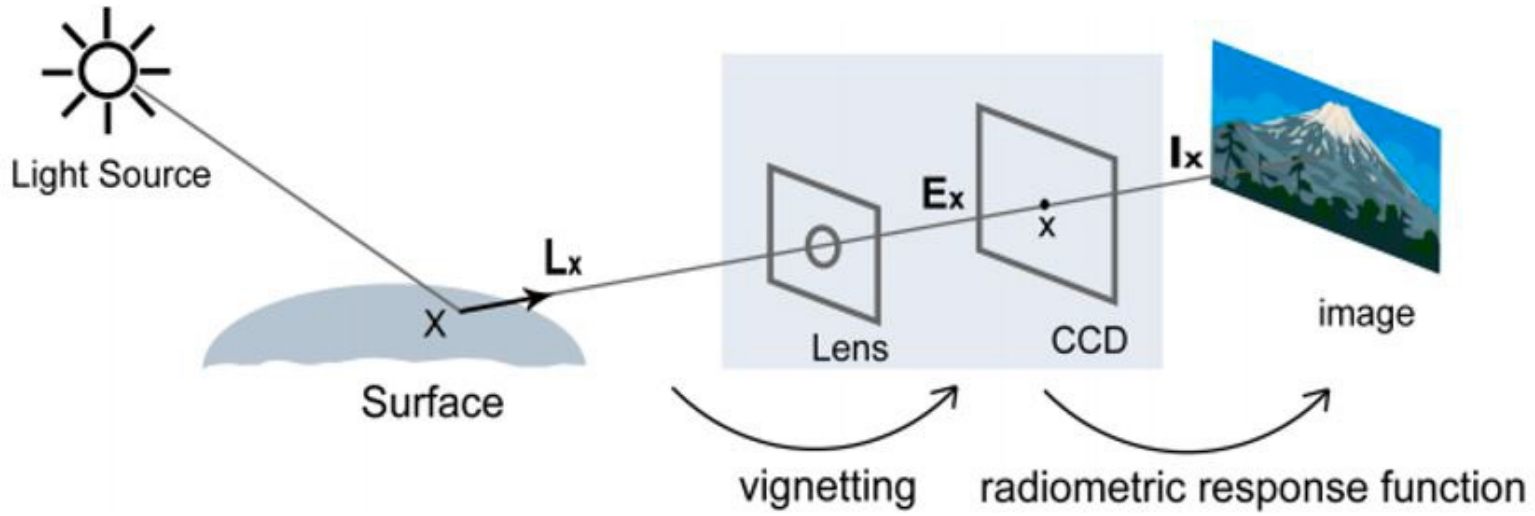
Example: Multi-person tracking (CMU)



If you know you are tracking human bodies, the known shape of the body can be exploited.

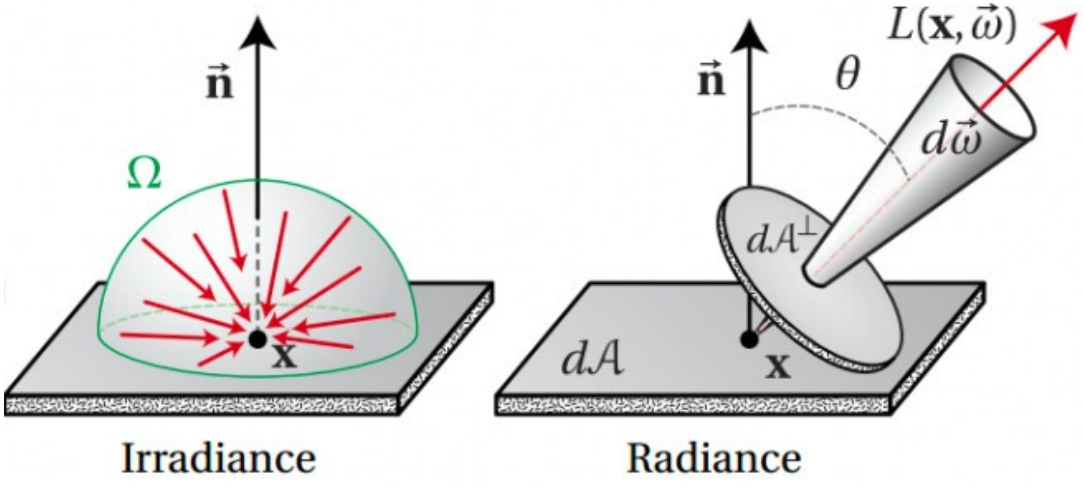
Image formation

Image formation is a physical process that captures scene illumination through a lens system and relates the measured energy to a signal.



Radiance and Irradiance

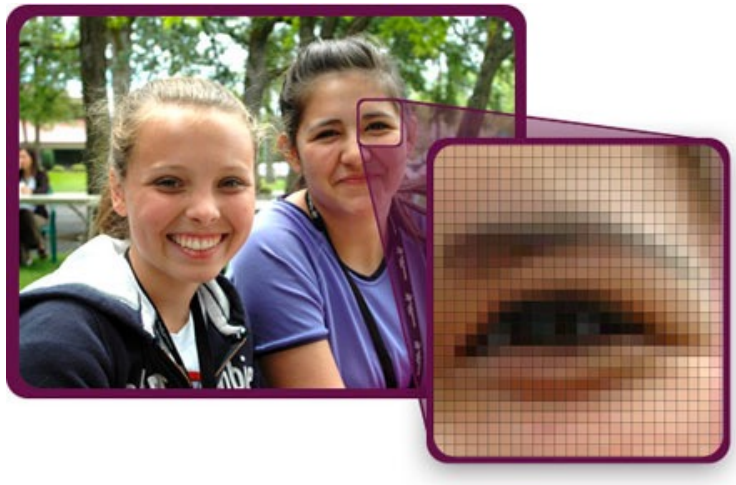
- Irradiance E: Amount of light falling on a surface, in power per unit area (watts per square meter).
- Radiance L: Amount of light radiated from a surface, in power per unit area per unit solid angle. Informally “Brightness”.



- But… Image irradiance E is proportional to scene radiance

Image acquisition – exposure

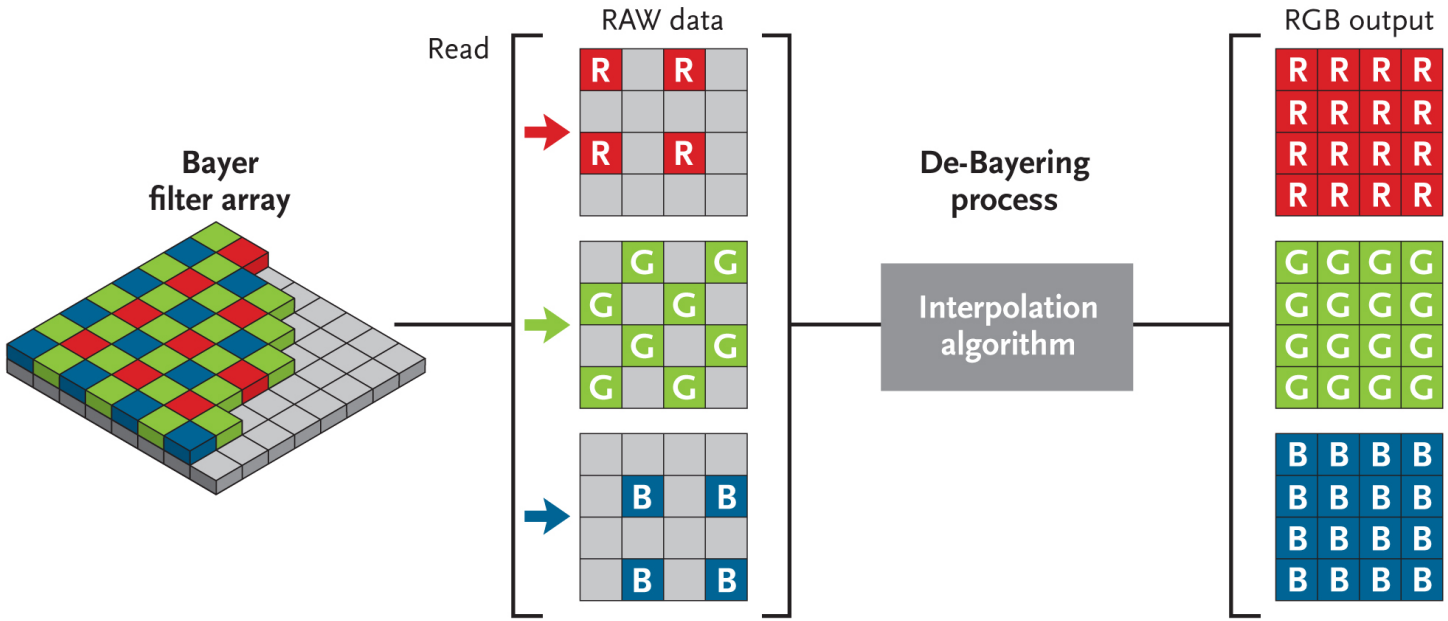
- Light enters the camera and reaches the chip inside that has an array sensors.
- Over a period of (expose) time the amount of light is measured at each sensor.
- The amount of light is converted into an intensity value, one value per sensor.



*Intensity = Image irradiance * sensor size * exposure time*

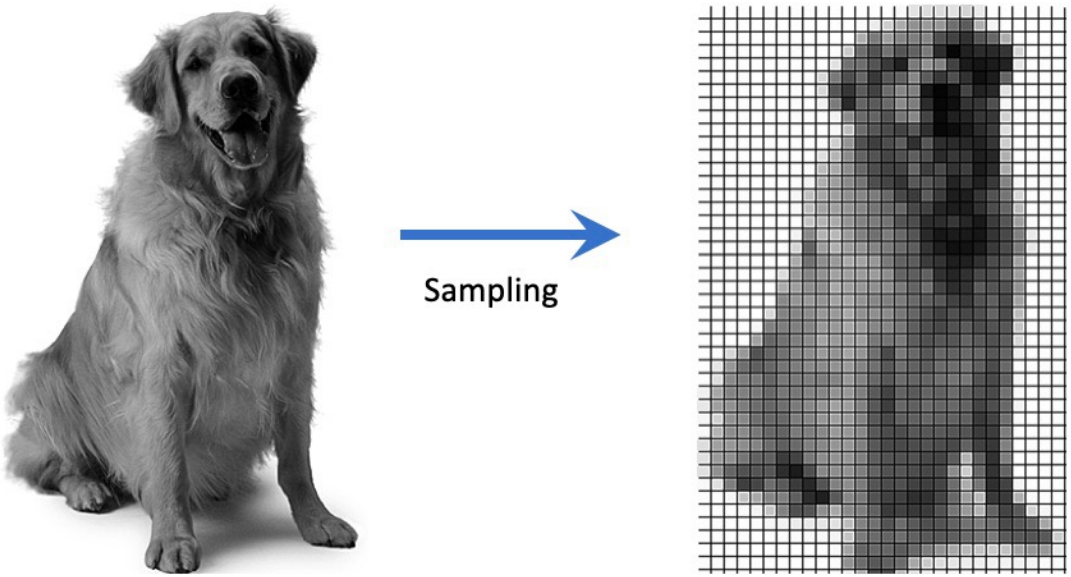
Image acquisition – colours

- A Bayer filter is placed on-top of sensors to filter out light of different colours.
- Measurements are then interpolated to get red, green and blue for each pixel.

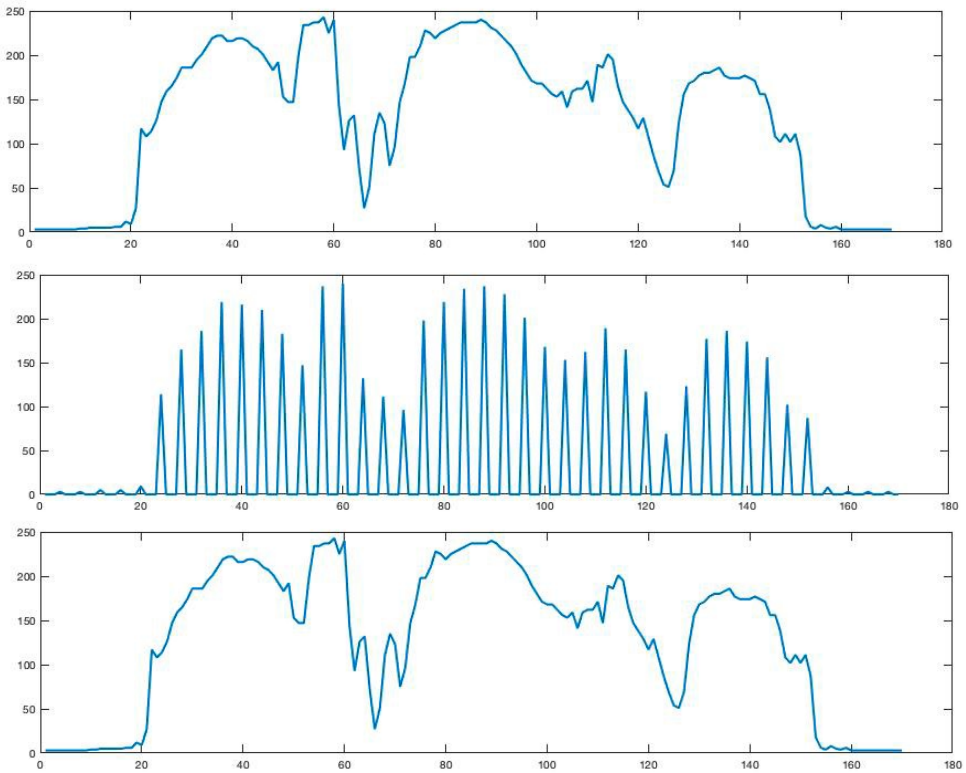


Sampling and quantization

- Sample the continuous image signal at a finite set of points and quantize the registered values into a finite number of levels, usually from 0 to 255.
- Sampling distances $\Delta x, \Delta y$ and Δt determine how rapid spatial and temporal variations can be captured.



Sampling and quantization



If sampling rate is high enough, the original image can (at least in theory) be perfectly reconstructed.

Different kinds of image errors

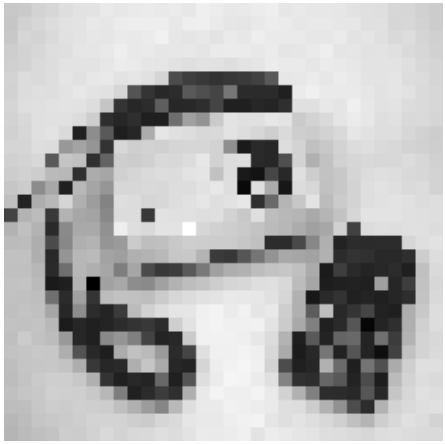
- Sampling errors: When the resolution is not high enough to capture the variations in the image.
- Quantization errors: The difference between the real intensity value and the assigned discrete value.
- Image noise: What sensors really count are photons. If exposure time is too short, too few photons are counted and intensity values become inaccurate.
- Saturation: When the physical value moves outside the allocated range [0, 255], then it is represented by the end of range value.

Different image resolutions

256 x 256

128 x 128

64 x 64



32 x 32

Sampling due to limited spatial resolution. Often the resolution is so high that we don’t view it as a problem, but for small objects it can still be an issue.

Different number of grey levels

1 bit per pixel

2 bits per pixel

4 bits per pixel

8 bits per pixel

Quantization due to limited intensity resolution. In most cases only 6 bits (64 grey-levels) is enough to make an image visually pleasing.

Good questions to ask

- What is computer vision good for?
- What can human vision tell us about how to do computer vision?
- Why is vision an active process?
- Why are discontinuities (edges) important for vision?
- For what kind of problems is deep learning useful in computer vision?
- What is image processing, image analysis and computer vision?
- What errors do we have in the acquisition process?
- What does sampling and quantization mean?

Recommended readings

- Gonzalez and Woods: Chapters 1.4
- Szeliski: Chapters 1.1 - 1.3, 2.3
- Introduction to labs (on web page)