INTRO TO COMPUTER VISION

From hand-crafted features to Convolutional Neural Networks

Marco Zullich, PhD student @ DIA, UniTS
Alessio Ansuini, adjunct professor @ DMG, UniTS
April 12, 2021

WHAT DO YOU SEE IN THIS IMAGE?

AND THIS?

Queen Elizabeth II

Joy

Shame

Victory

Geoff Hurst

Rimet Cup

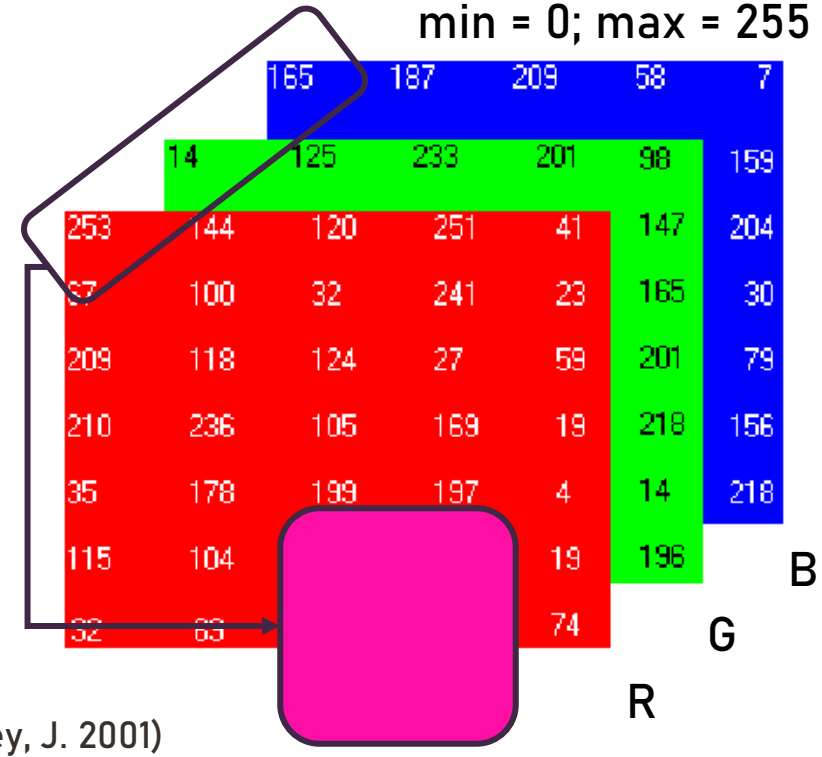
WHAT DOES A COMPUTER SEE?

234	235	236	236	237	237	237	238	238	239	238	239	239	239	239	238	239	239	239	239	238	238	238	237	237	237	236	235	235	234
234	235	236	237	237	237	237	238	238	239	239	239	239	239	239	239	239	239	239	239	239	238	238	237	237	237	237	236	235	234	234
234	235	236	236	237	237	237	238	238	239	239	239	239	239	239	239	239	239	239	239	239	238	238	237	237	237	236	236	235	234	234
234	235	236	236	237	236	237	238	238	239	239	197	85	217	240	214	226	239	239	239	238	237	237	237	237	236	236	235	234	233
233	234	235	236	237	237	237	236	230	152	62	38	89	38	46	114	133	209	222	238	238	237	237	237	237	236	235	235	234	233
233	234	235	236	236	237	236	219	127	16	17	13	7	6	18	35	87	216	238	238	237	237	237	237	236	235	235	234	233	233
233	234	234	235	236	237	235	210	16	2	13	20	16	19	2	22	38	48	204	238	237	238	237	236	236	235	234	233	233	232
233	233	234	234	235	236	180	16	48	185	192	183	182	163	61	7	2	32	171	208	228	234	233	233	233	231	231	228	228	227
233	233	233	235	235	230	16	46	174	192	192	192	191	185	179	170	61	29	187	227	222	231	231	231	230	228	229	227	226	226
229	229	229	229	230	221	14	63	173	176	177	182	180	161	172	171	129	31	217	223	223	230	230	230	230	229	228	227	225	225
227	227	228	228	228	33	31	107	167	185	182	181	185	179	189	173	150	21	225	223	228	227	227	225	225	223	223	220	218	219
220	222	226	225	210	12	10	172	58	26	19	73	147	164	143	155	161	16	207	221	222	222	221	219	221	220	219	219	217	216
212	213	216	216	149	2	17	170	100	40	62	53	163	126	4	6	81	12	219	215	222	221	220	220	219	218	216	215	216	213
208	211	211	209	82	4	91	190	162	133	92	116	168	139	13	58	80	34	2	210	214	217	219	219	217	216	217	215	213	212
207	207	206	207	85	90	152	174	184	178	169	161	174	139	79	63	146	8	167	195	204	212	215	214	212	211	209	206	202	203
204	204	205	208	157	113	105	170	173	171	124	165	165	157	163	187	185	9	199	181	201	204	205	204	205	203	203	201	199	199
201	202	203	201	153	162	141	150	100	51	123	75	75	6	114	112	146	30	201	198	195	200	201	202	202	203	201	200	199	199
198	199	199	176	17	110	144	124	93	99	134	107	99	105	118	32	67	149	128	194	202	202	202	202	202	200	200	199	196	195
195	198	198	197	108	0	120	105	130	142	144	150	142	115	75	105	84	198	192	194	201	201	200	200	200	199	198	197	196	194
9	10	8	4	203	7	29	77	87	154	143	106	76	107	126	101	78	189	197	199	199	199	199	199	198	197	195	194	194	191
9	7	3	224	210	0	148	51	89	113	167	154	128	151	132	33	159	192	195	195	196	196	196	196	195	194	194	191	191	190
12	121	211	202	192	2	153	28	80	54	112	82	112	107	63	35	178	190	192	192	194	193	193	194	193	191	193	191	190	190
25	185	128	191	197	88	138	104	25	5	28	13	17	15	5	47	27	117	110	185	187	185	186	186	188	190	190	189	187	186
55	45	138	188	180	207	110	114	63	28	29	12	8	18	8	12	69	34	86	98	101	163	178	178	177	178	175	177	179	180
59	39	90	159	152	147	48	105	75	63	58	10	26	73	11	14	75	18	60	78	83	96	121	171	172	173	172	170	169	168
16	23	24	125	174	147	129	64	91	74	49	49	63	10	14	18	21	13	20	44	62	82	99	84	148	173	171	168	167	165
25	18	19	164	168	162	185	131	65	81	57	30	7	19	24	17	20	65	12	22	27	44	65	95	77	96	169	167	167	163
28	21	16	16	141	136	176	166	182	63	65	39	23	23	4	0	16	27	27	18	26	27	27	51	61	65	10	23	167	164
17	22	16	15	189	191	107	103	27	72	67	158	17	3	2	35	27	52	35	28	21	33	34	26	40	16	13	9	9	165
11	20	24	18	29	187	190	156	118	60	8	7	28	25	26	40	10	4	4	11	22	22	27	26	17	7	10	10	14	21

COLOR IMAGES

CMYK · CIELAB · HSV · RGB and many more…

- There exists a great variety of encodings of color images
- All these encodings work as composition of multiple color channels

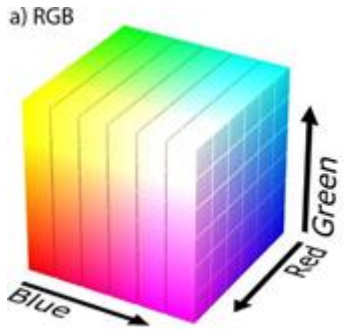


Credits: Application of Digital Image Processing to Marker-free Analysis of Human Gait (Courtney, J. 2001)

COLOR SPACES (ENCODINGS)

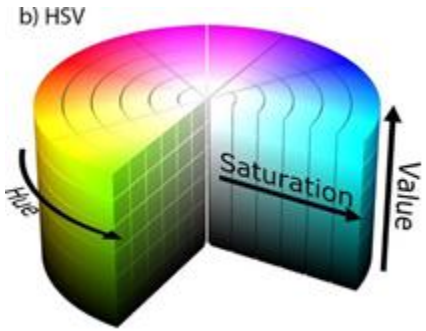
Apart from RGB, there exist a myriad of other color spaces, each with its own pros and cons.

RGB



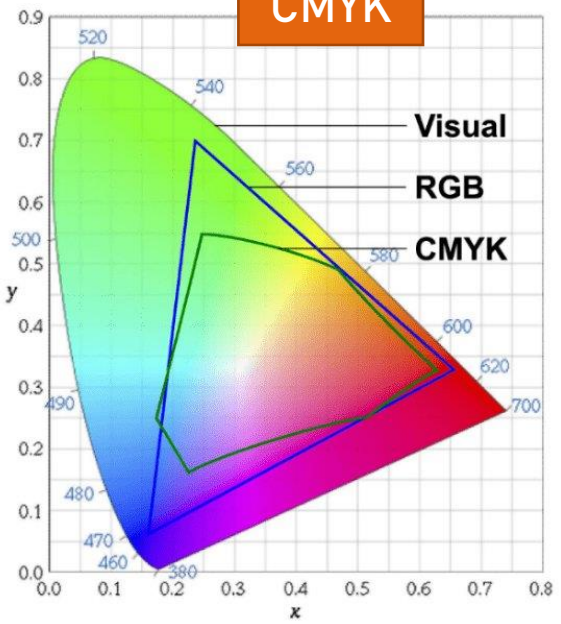
Tailored to project on screen

HSV



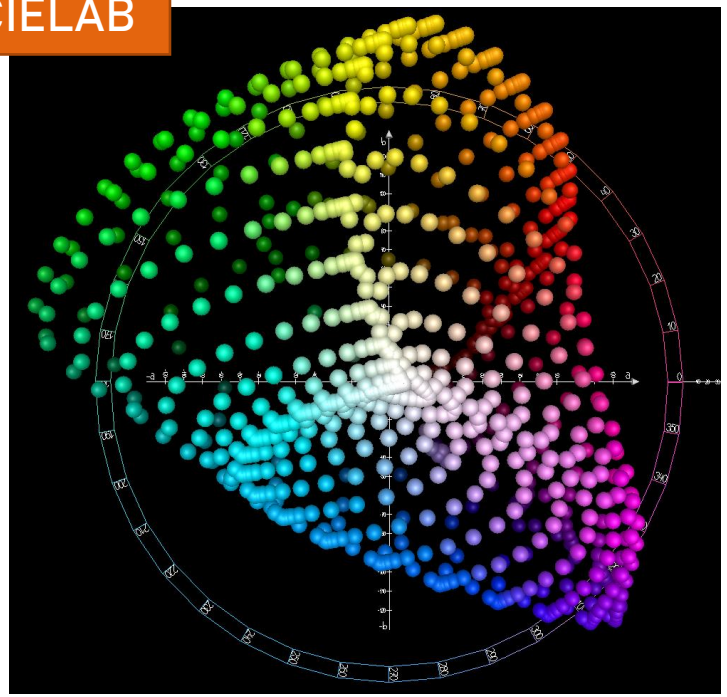
More intuitive for humans than RGB

CMYK



Can you guess what it's good for? (printing)

CIELAB



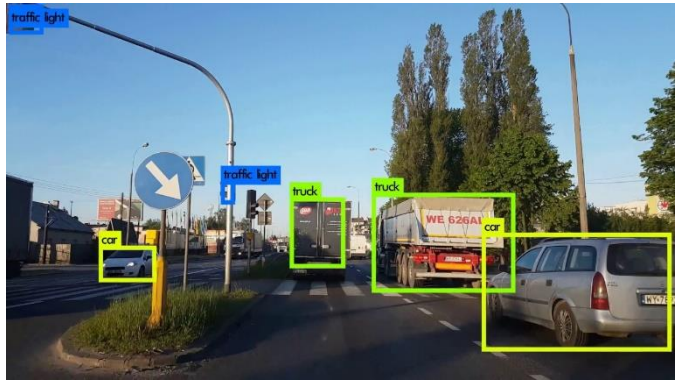
Perceptual uniformity

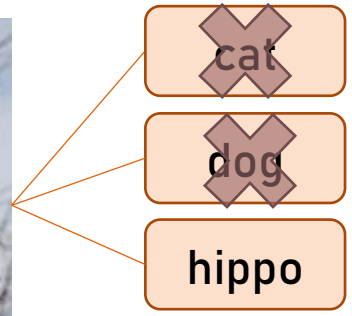
...

Image credits
1) and 2) https://medium.com/neurosapiens/segmentation-and-classification-with-hsv-8f2406c62b39
3) https://aldertech.com/color-101-color-spaces/
4) https://commons.wikimedia.org/wiki/File:CIELAB_color_space_top_view.png

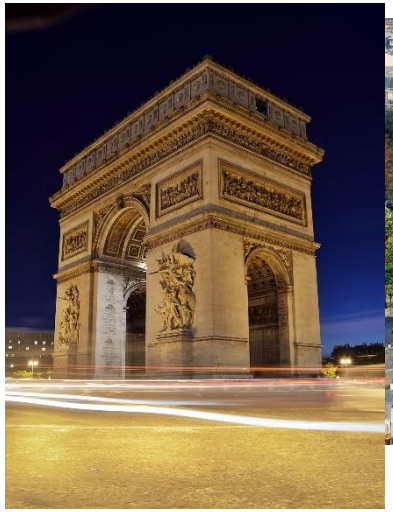
THE MULTIPLE FACETS OF RECOGNITION

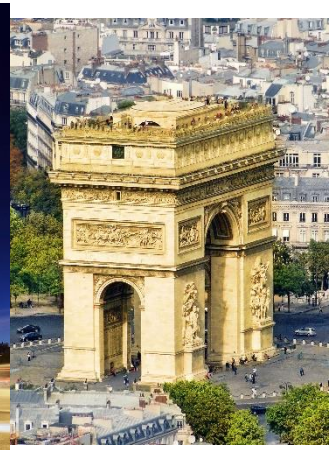
Object Detection
Find specific categories of objects in an image





Instance Recognition
Recognize an instance of an object in a novel environment or from a novel viewpoint





Category-level Recognition
Categorize an image as belonging to one specific category

Credits: https://reckoning.dev/yolo-v4/; https://www.pexels.com/photo/cars-france-landmark-lights-2344/; https://commons.wikimedia.org/wiki/File:Arc_de_Triomphe,_Paris_3_October_2010.jpg; https://upload.wikimedia.org/wikipedia/commons/thumb/4/4d/Cat_November_2010-1a.jpg/140px-Cat_November_2010-1a.jpg

RECOGNITION: THE OLD WAY

- In «classical» CV, researchers usually acted by means of **local operators**
- LOCAL OPERATOR
  - Image $I \in \mathbb{R}^{h \times w}$; $I(i,j) =$ intensity of pixel i,j
  - Operator $f: \mathbb{R}^{h \times w} \to \mathbb{R}^{h' \times w'}$ takes in an input image and outputs another image, **possibly of different size**
  - The operator is local if, $f(I)(i,j)$ depends on the pixel values of a **neighborhood of i,j**

EASY EXAMPLE: MEAN FILTERING

I

35	142	125	10
37	100	38	154
220	200	122	0
0	22	44	123

$J = f(I)$

113	99
87	89

Target: apply the mean of the image for of each possible contiguous neighborhood of size 3×3

35	142	125	10
37	100	38	154
220	200	122	0
0	22	44	123

$\cdot \frac{1}{9}$

K

1	1	1
1	1	1
1	1	1

kernel

$=$

113	99
87	89

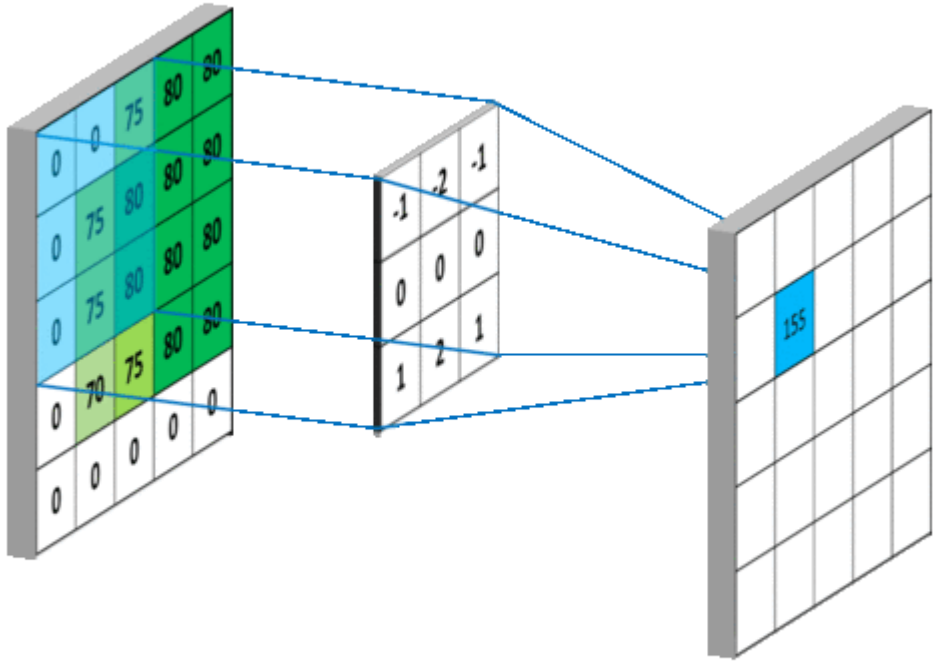
Cross-correlation

$J = I \otimes K$

2D CROSS-CORRELATION

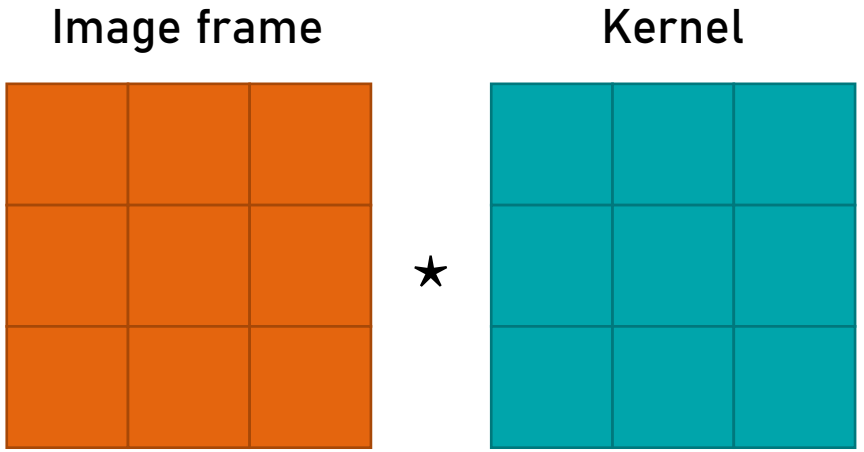
- We're given an image I of size $h \times w$
- And a kernel K of size $k \times l$, both odd
  - The center of the kernel is $c = \left(\left\lceil \frac{k}{2} \right\rceil, \left\lceil \frac{l}{2} \right\rceil\right)^T$

$$I \otimes K(i,j) = \sum_{m=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{n=\lfloor l/2 \rfloor}^{\lfloor l/2 \rfloor} I(i+m, j+n) \cdot K(c_x + m, c_y + n)$$



Credits: https://mlnotebook.github.io/img/CNN/convSobel.gif

2D CONVOLUTION



$$I \star K(i,j) = \sum_{m=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{n=\lfloor l/2 \rfloor}^{\lfloor l/2 \rfloor} I(i+m, j+n) \cdot K(c_x - m, c_y - n) = \sum_{m=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{n=\lfloor l/2 \rfloor}^{\lfloor l/2 \rfloor} I(i-m, j-n) \cdot K(c_x + m, c_y + n)$$

Note: for symmetric kernels ($K(c_x + i, c_y + j) = K(c_x - i, c_y - j)$), convolution ≡ cross-correlation

BORDER EFFECTS (PADDING)

$I \in (\mathbb{N} \cap [0,255])^{4\times4}$

35	142	125	10
37	100	38	154
220	200	122	0
0	22	44	123

$\otimes$

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

$=$

$J = f(I) \in (\mathbb{N} \cap [0,255])^{2\times2}$

113	99
87	89

$\bar{I} \in (\mathbb{N} \cap [0,255])^{6\times6}$

?	?	?	?	?	?
?	35	142	125	10	?
?	37	100	38	154	?
?	220	200	122	0	?
?	0	22	44	123	?
?	?	?	?	?	?

$\otimes$

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

$=$

?	?	?	?
?	113	99	?
?	87	89	?
?	?	?	?

$J = f(\bar{I}) \in (\mathbb{N} \cap [0,255])^{4\times4}$

Question

Is there a way to keep the resulting $f(I)$ of the same dimension of I?

Answer

Padding.

TYPES OF PADDING

mirroring

zero-padding
(constant padding)

warping

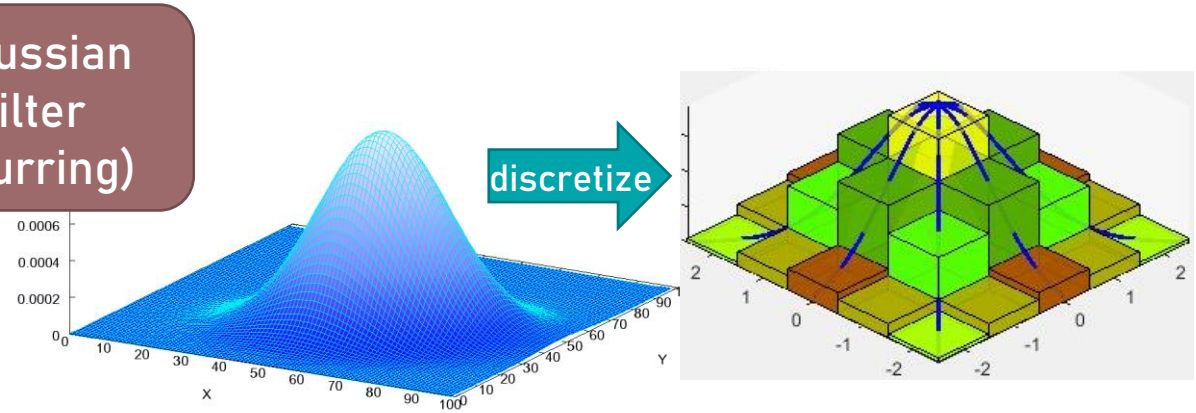
clamping

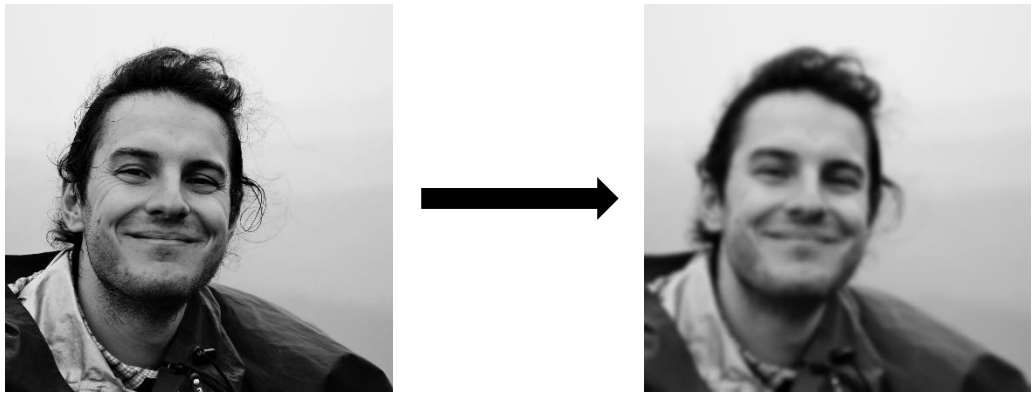
USAGE OF CORRELATION/CONVOLUTION

- Correlation/Convolution have historically been used to recognize some specific low-level features and to execute given tasks

Sobel filter (oriented edges)

Gaussian filter (blurring)





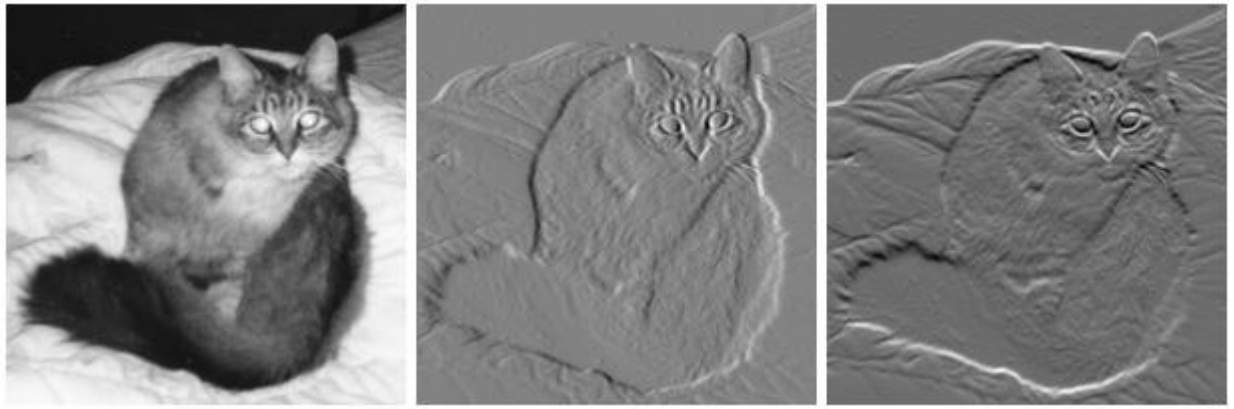
Credits: reelaboration of https://en.wikipedia.org/wiki/Multivariate_normal_distribution#/media/File:Multivariate_Gaussian.png and https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/e573acb5-4a80-11e4-9553-005056977bd0/fecbef1b-a5bd-4937-b4df-c166a7442b2b/images/1597241565.jpg

IMAGE GRADIENT

«Directional change in the intensity of color of an image» (wiki)

$$\nabla \mathrm{I} = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right)^T$$

Can be approximated in various ways from the original image, e.g. using the Sobel filter from the previous slide.



By Njw000 - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=10588443

SHAPE RECOGNITION

One of the most successful algorithm for shape recognition used in industrial settings (e.g. HALCON) is the **shape-based matching** applied to image gradients.

Template
T

Scene
I

"(Normalized) cross-correlation" between image and template gradients

$$\nabla \mathbf{I} = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right)^T \in \mathbb{R}^{2\times h\times w} \quad \text{reshape} \quad \mathbf{G_I} \in \mathbb{R}^{h\times w x 2}$$

$$s = \frac{1}{n}\sum_{x,y\in R} \frac{\langle \boldsymbol{G_I}(x,y), \boldsymbol{G_T}(x,y)\rangle}{\|\boldsymbol{G_I}(x,y)\| \cdot \|\boldsymbol{G_T}(x,y)\|}$$

$s \to 1 \Rightarrow$ match
$s \to 0 \Rightarrow$ no match

Region of interest

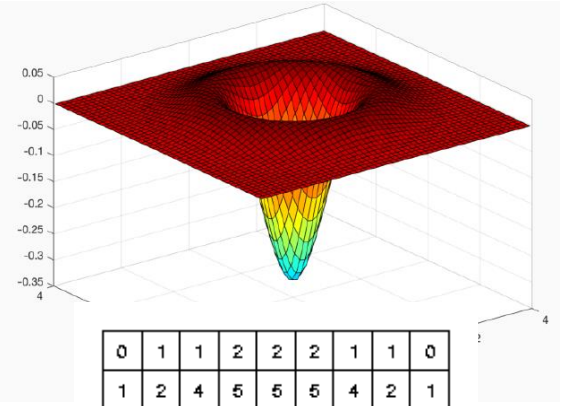
Images from Steger, Carsten. "Similarity measures for occlusion, clutter, and illumination invariant object recognition." *Joint Pattern Recognition Symposium*. Springer, Berlin, Heidelberg, 2001.

KEYPOINT DETECTION USING SIFT - BLOB DETECTION

SIFT (Scale Invariant Feature Transform) is built upon the concept of detection of *blobs* at different scales.

Blobs can be detected by means of the second derivatives $\frac{\partial^2 I}{\partial x^2}, \frac{\partial^2 I}{\partial y^2}$ and then obtaining the Laplacian:

$$L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}$$



0	1	1	2	2	2	1	1	0
1	2	4	5	5	5	4	2	1
1	4	5	3	0	3	5	4	1
2	5	3	-12	-24	-12	3	5	2
2	5	0	-24	-40	-24	0	5	2
2	5	3	-12	-24	-12	3	5	2
1	4	5	3	0	3	5	4	1
1	2	4	5	5	5	4	2	1
0	1	1	2	2	2	1	1	0

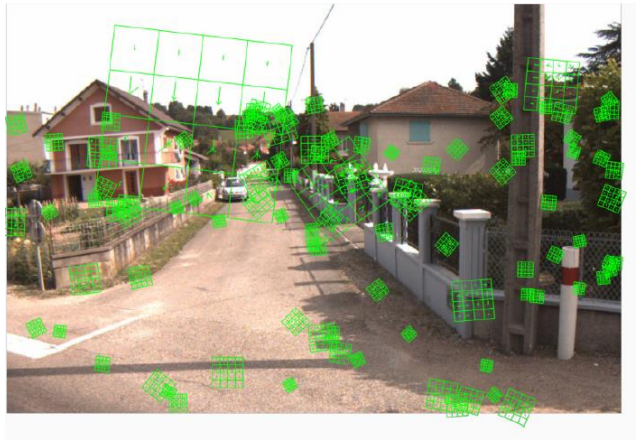
And then convolving it with a 2D homoschedastic Gaussian to obtain the Laplacian of Gaussian (LoG) operator.

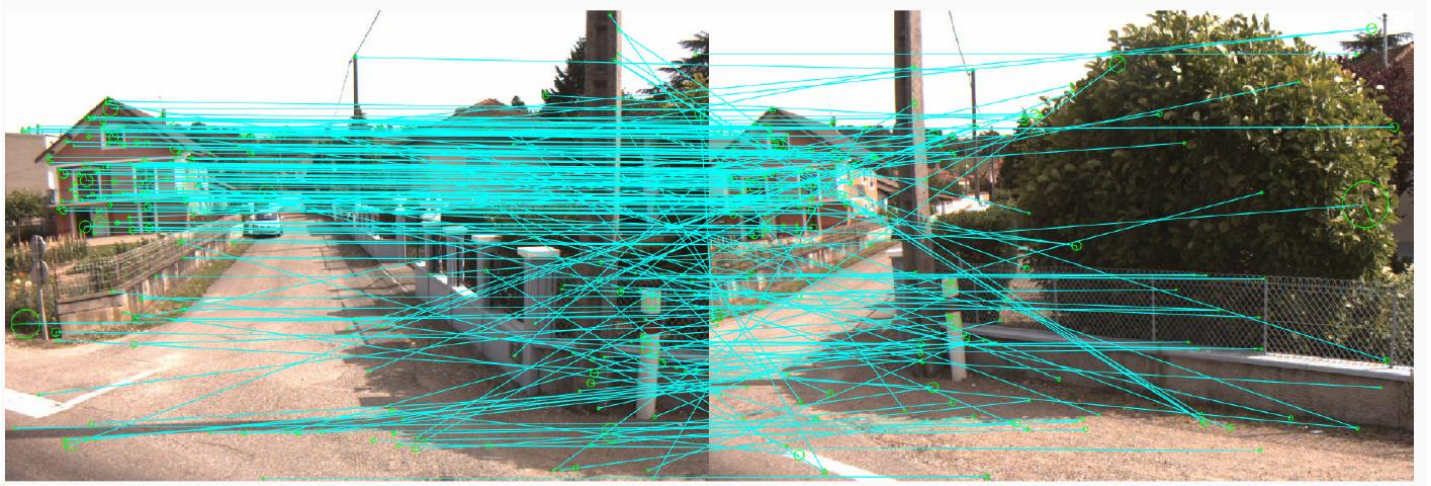
The LoG can be directly obtained by convolving the original image with a specific filter (example with $\sigma = 1.4$)

Last image from https://homepages.inf.ed.ac.uk/rbf/HIPR2/log.htm

SIFT - KEYPOINT DETECTION

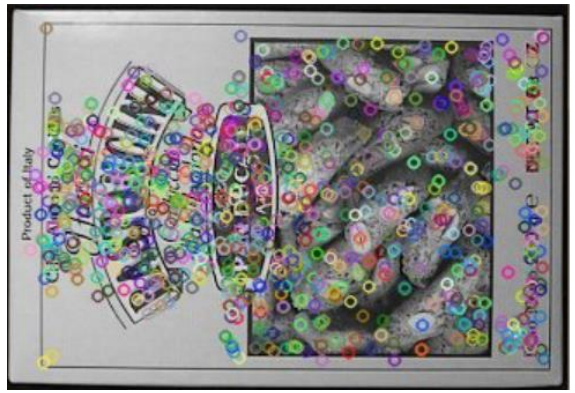
- The construction of SIFT is much more complicated than that and involves several passages to gain both robustness and efficiency
- The process concludes with the definition of a descriptor $d_k \in \mathbb{R}^{128}$ for each keypoint k
- The descriptor defines the image gradient in the neighborhood of the keypoint



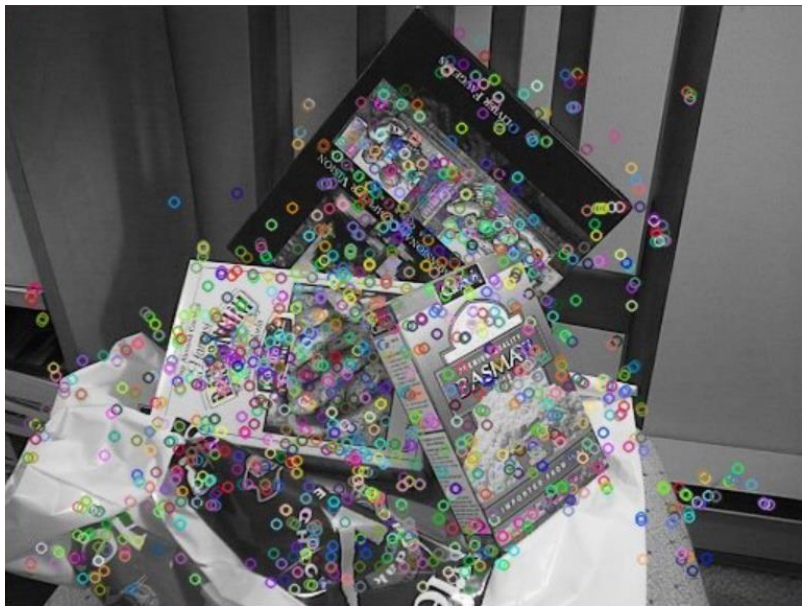


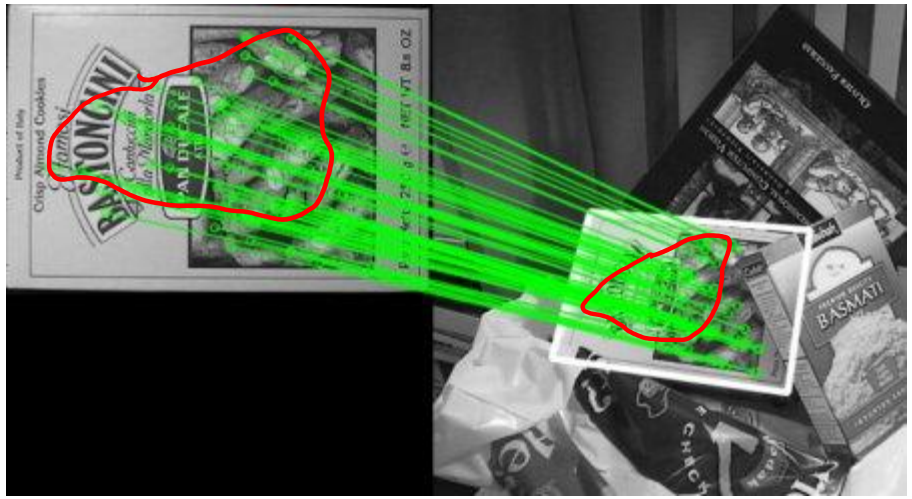
USING KEYPOINTS FOR RECOGNITION

Identify and describe keypoints in template



Identify and describe keypoints in scene





Match keypoints

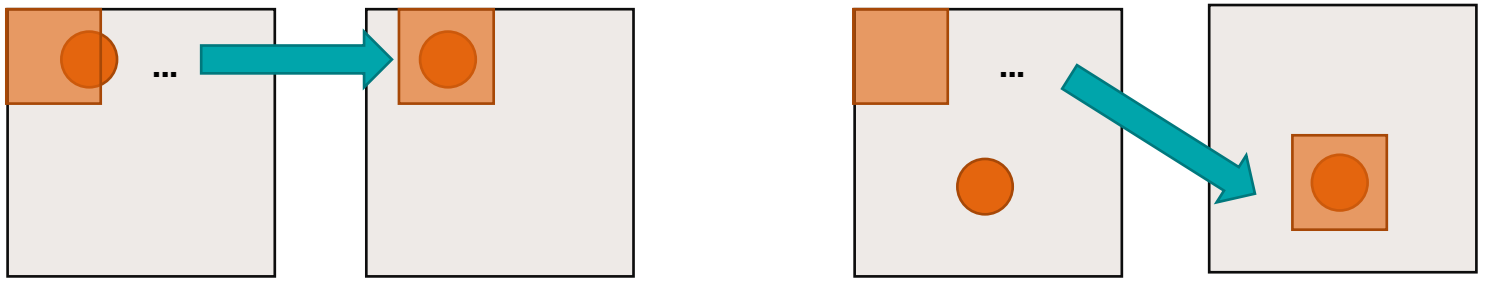
Build homography with subset of matching keypoints

Stop when reconstruction is "sufficiently good" from a geometrical viewpoint

RanSaC

MORE ON CONVOLUTION/CORRELATION

Convolution is translation-invariant: given its “moving window” approach, the response to a given object is the same if it is translated.



“Convolution was designed by hand as a heuristic solution to the problem of capturing translation-invariant features at different levels of hierarchy”

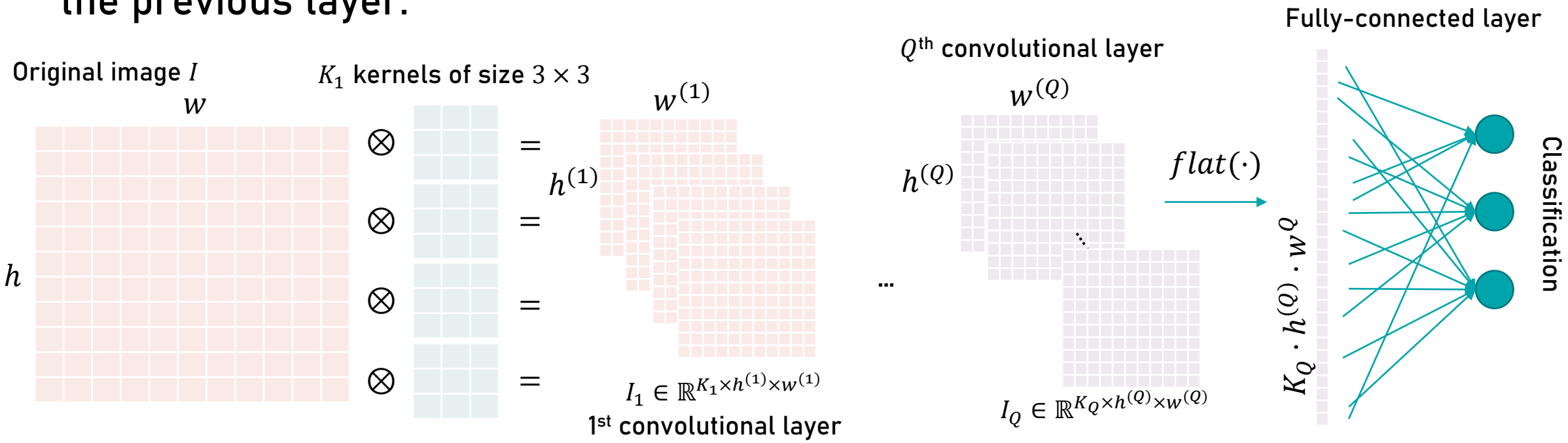
Stanley, Kenneth O., et al. "Designing neural networks through neuroevolution."

THE NEXT STEP - DEEP LEARNING

- Nowadays, we have the tools to do more than that
- Artificial Neural Networks (ANNs) have achieved a huge success because of their ability to iteratively **learn** their weights in order to **adapt** to a given task
- There's no need anymore for *hand-crafted features*
- **Let the machine learn the weights of a convolutional kernel rather than designing them heuristically**

CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) are essentially a cascade of sequential cross-correlations* grouped in layers.
Each layer learns higher level features starting from the features learned by the previous layer.



* we have seen before that convolution and correlation are similar in concept. Correlation is slightly more efficient.

WHERE CAN I LEARN MORE?

- CNNs (a bit more rigorously)

  Next lectures…

- Classical Computer Vision, specific models for instance recognition, object detection…

  "Computer Vision and Pattern Recognition", 3rd semester DSSC course by prof. Pellegrino

- References?

  See next slide…

REFERENCES

- Classical Computer Vision:
  - Szeliski, R. (2010). *Computer vision: algorithms and applications, 1st edition*.
  - Forsyth, D. and Ponce, J. (2012). *Computer vision: a modern approach, 2nd edition*.
- Shape recognition
  - Steger, C. (2001). Similarity measures for occlusion, clutter, and illumination invariant object recognition.
- Convolutional Neural Networks
  - Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning*.
  - Nielsen, M. (2019) *Neural Networks and Deep Learning*.

THANKS FOR THE ATTENTION!

marco.zullich@phd.units.it

contact me on Teams as well!