Perception – Computer Vision II –

CSC398 Autonomous Robots

Ubbo Visser

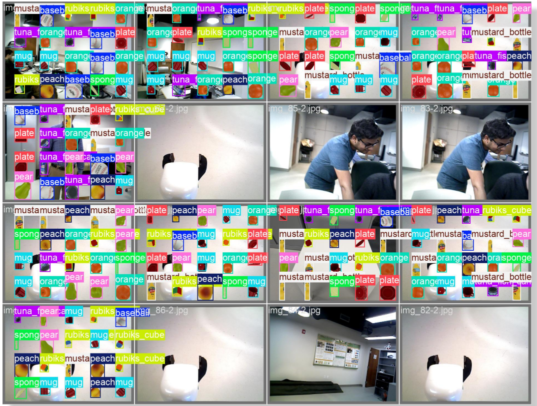
Department of Computer Science
University of Miami

November 19, 2024

UNIVERSITY
OF MIAMI

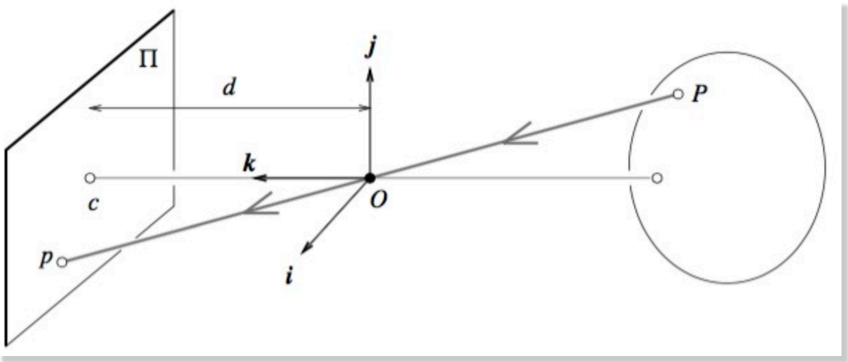
Outline

1. Representations in Computer Vision
2. Visualizing & Understanding of CNNs
3. Filtering
4. Detectors



From 3D world to 2D images

- So far we have focused on mapping 3D objects onto 2D images and on leveraging such mapping for scene reconstruction
- Next step: how to represent images and infer visual content?



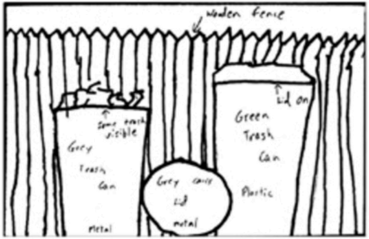
Today's lecture

- **Aim:**
  - Learn fundamental tools in image processing for filtering and detecting similarities
  - Learn how to detect and describe key features in images
- **Readings:**
  - Siegwart, Nourbakhsh, Scaramuzza. Introduction to Autonomous Mobile Robots. Sections 4.3 – 4.5.4.

Representations in Computer Vision

Observed image

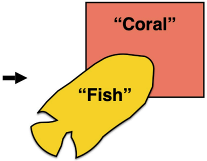
Drawn from memory



[Bartlett, 1932]
[Intraub & Richardson, 1989]

X

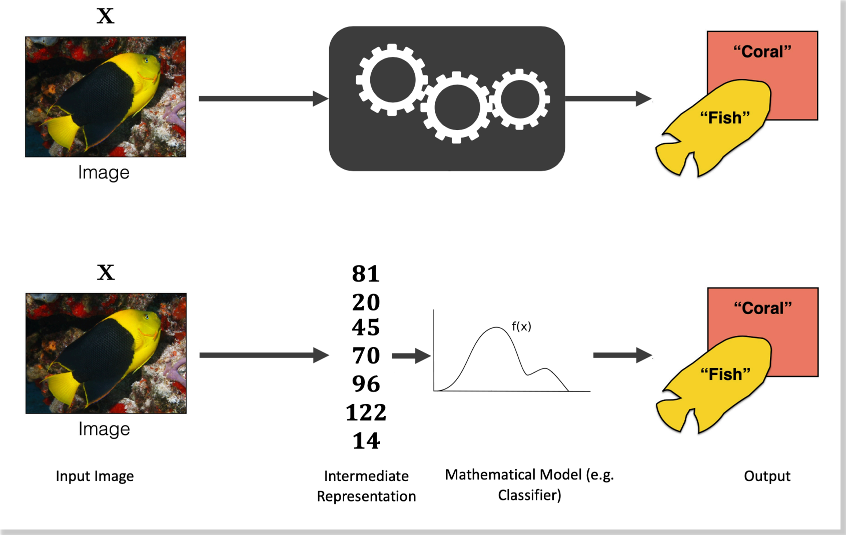
Image



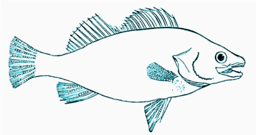
Compact mental representation

Example from Advances in Computer Vision – MIT – 6.869/6.819

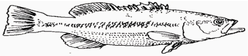
Typical CV Pipeline



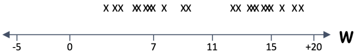
Example



~12 lbs

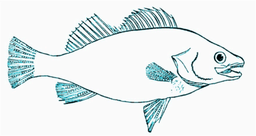


~8 lbs

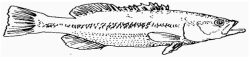


Example from CS331B: Representation Learning in Computer Vision

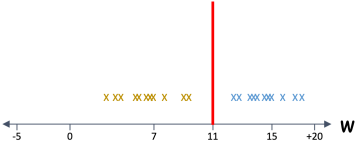
Example



~12 lbs

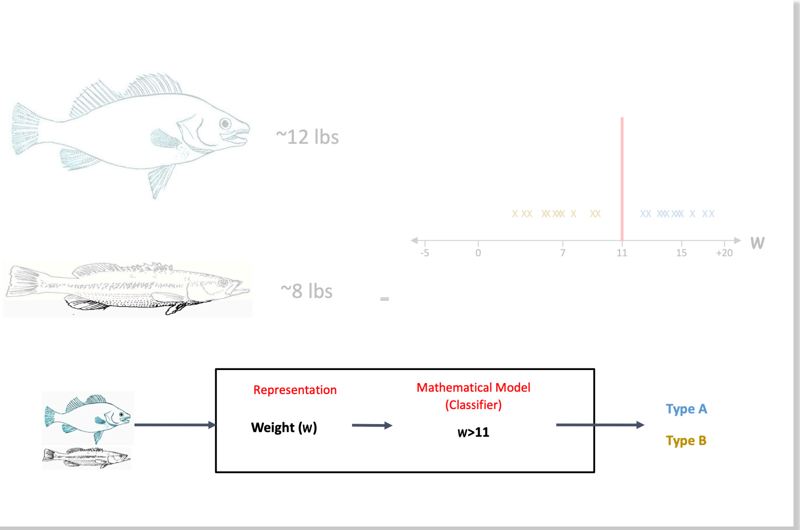


~8 lbs

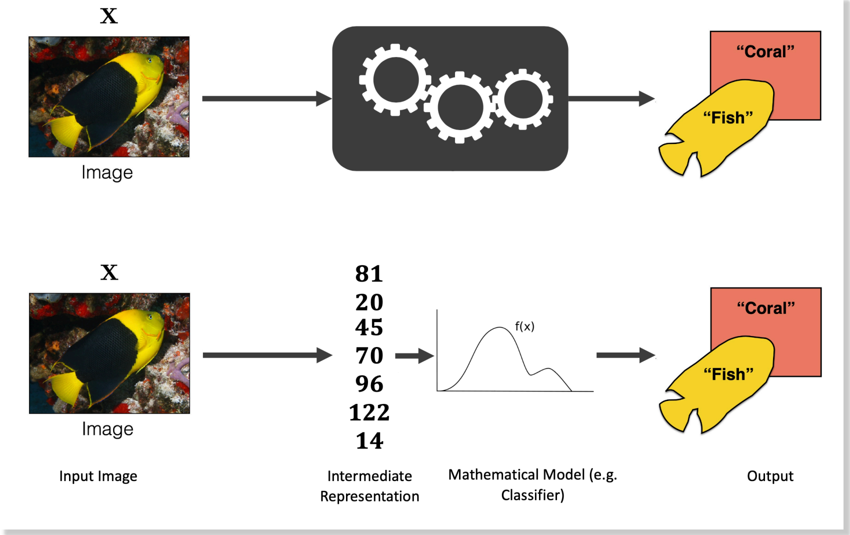


Example from CS331B: Representation Learning in Computer Vision

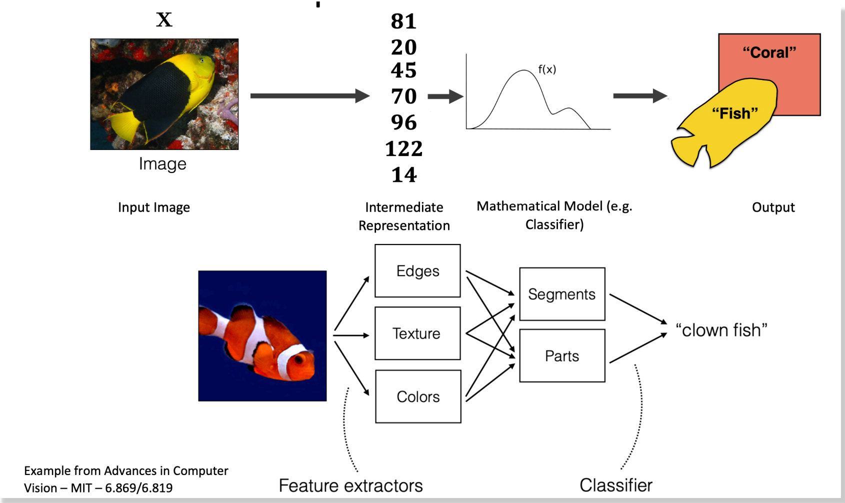
Example


~12 lbs
~8 lbs
-5
0
7
11
15
+20
W
Representation
Weight (w)
Mathematical Model
(Classifier)
w>11
Type A
Type B

Typical CV Pipeline



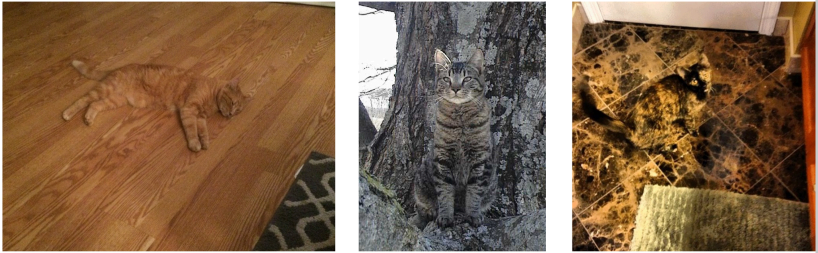
Traditional CV Pipeline



Represent these cats with a cat detector!

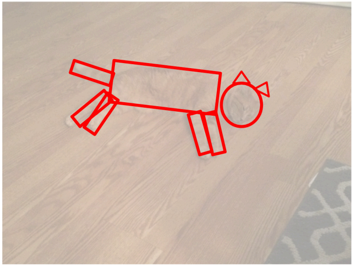
Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector (II)



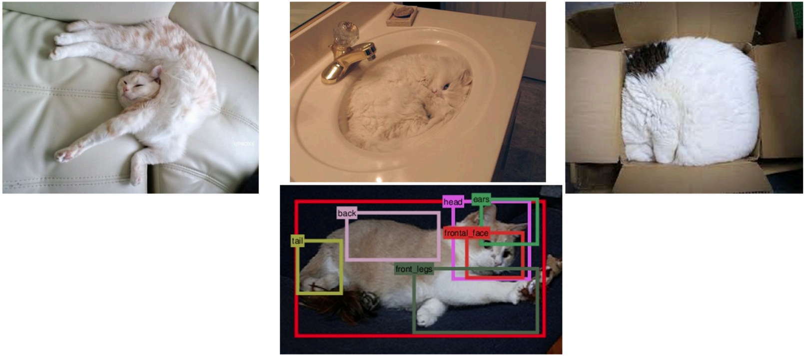
Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector (III)



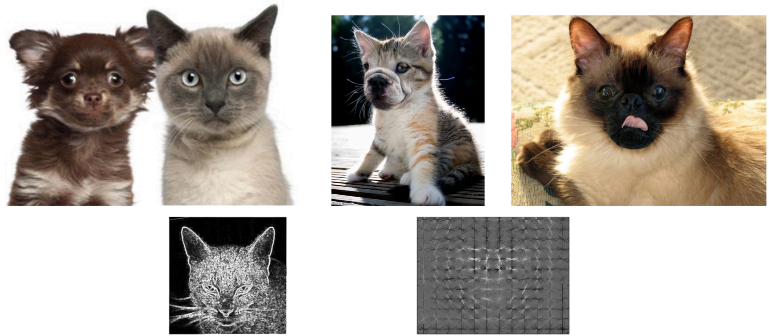
Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector (IV)



Example from CS331B: Representation Learning in Computer Vision

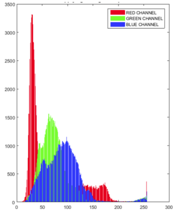
Represent these cats with a cat detector (V)

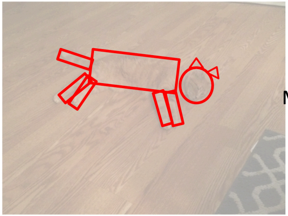


Example from CS331B: Representation Learning in Computer Vision

Summary of Traditional Components

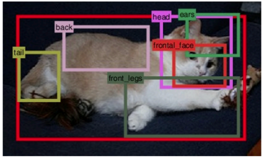
Color Histograms

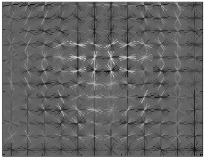




Model based Shapes

Deformable Part based Models (DPM)



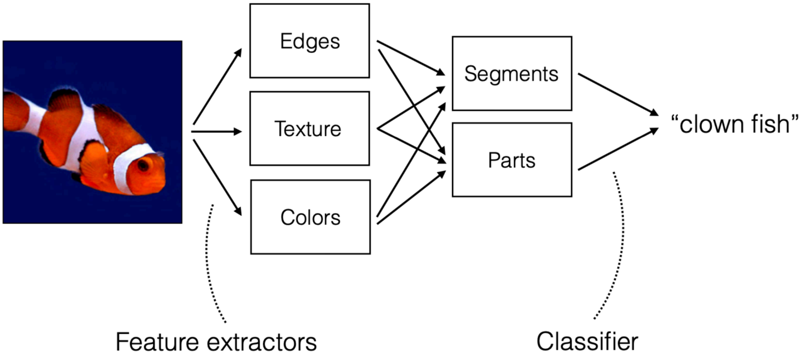


Histogram of Gradients (HOG)

Felzenszwalb et al. 2010.
Dalal and Triggs, 2005.
Beis and Lowe, 1997.

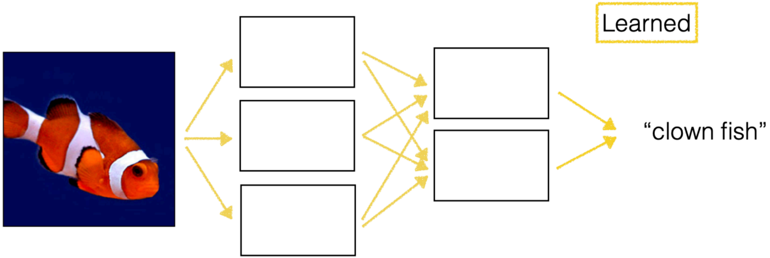
Example from CS331B: Representation Learning in Computer Vision

Traditional CV Pipeline



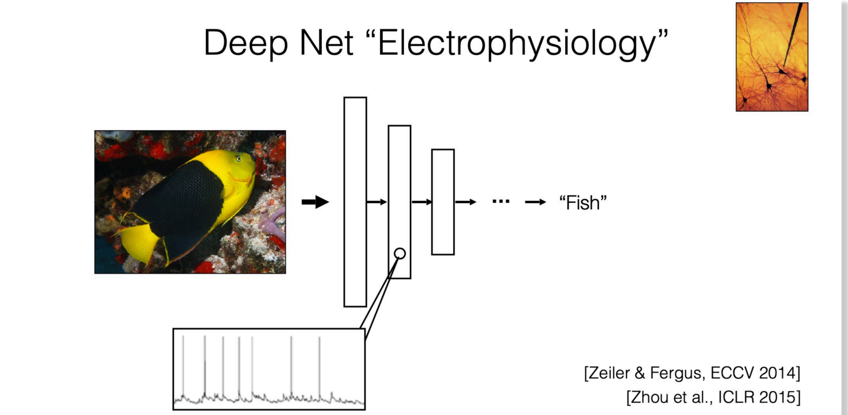
Example from Advances in Computer Vision – MIT – 6.869/6.819

Traditional CV Pipeline



Example from Advances in Computer Vision – MIT – 6.869/6.819

How do you interpret what the network has learned?



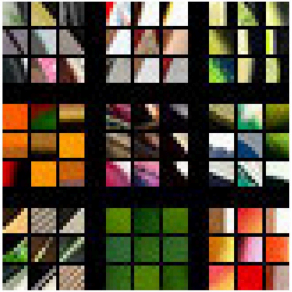
Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

Gabor-like filters learned by **layer 1**

Image patches that activate each of the **layer 1** filters most strongly

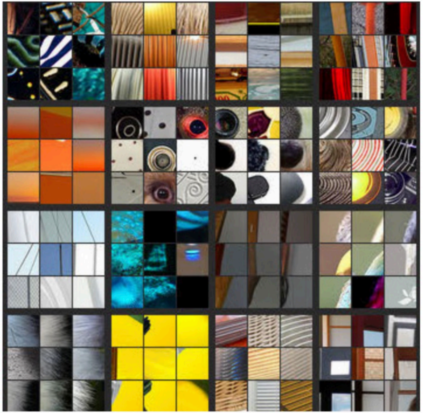


Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

Image patches that activate each of the **layer 2** neurons most strongly



Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

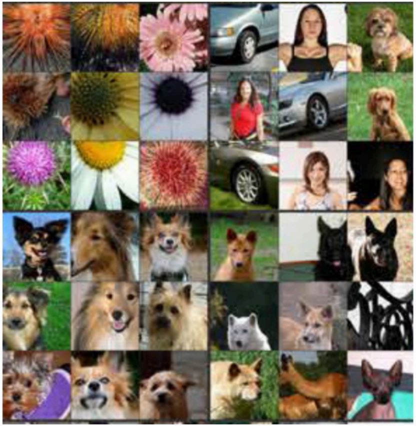
Image patches that activate each of the **layer 4** neurons most strongly

Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

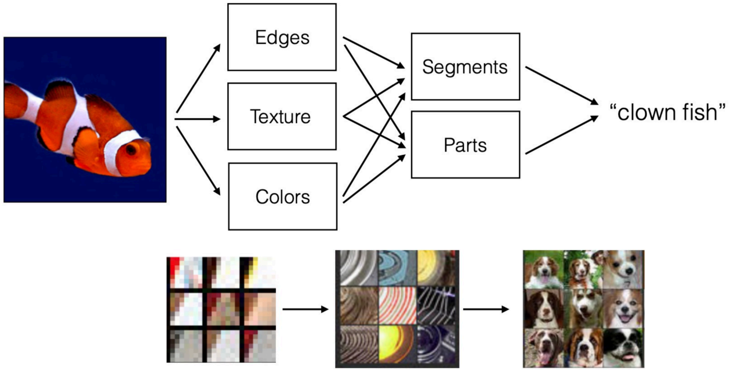
Image patches that activate each of the **layer 5** neurons most strongly



Example from Advances in Computer Vision – MIT – 6.869/6.819

Visualizing and Understanding CNNs

CNNs *learned* the classical visual recognition pipeline!



Example from Advances in Computer Vision – MIT – 6.869/6.819

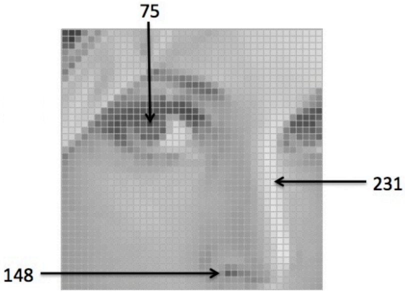
How to represent images?

Typical image processing pipeline

1. Signal treatment / filtering
2. Feature detection (e.g., DoG)
3. Feature description (e.g., SIFT)
4. Higher-level processing

Image filtering

- **Filtering:** process of accepting / rejecting certain frequency components
- Starting point is to view images as functions $I : [a, b] \times [c, d] \rightarrow [0, L]$, where $I(x, y)$ represents intensity at position (x, y)
- A color image would give rise to a vector function with 3 components



Represented as a matrix

j

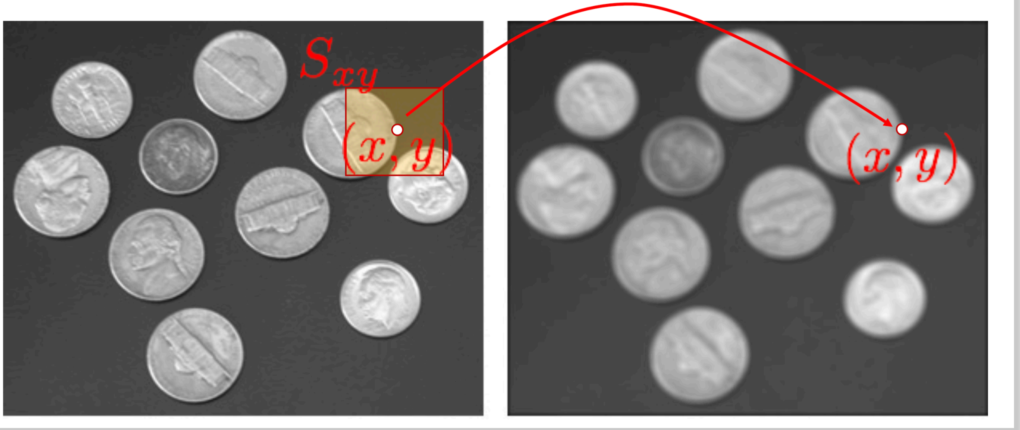
i

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Spatial filters

A spatial filter consists of

- A neighborhood S_{xy} of pixels around the point (x, y) under examination
- A predefined operation F that is performed on the image pixels within S_{xy}



Linear spatial filters

- Filters can be linear or non-linear
- We will focus on linear spatial filters

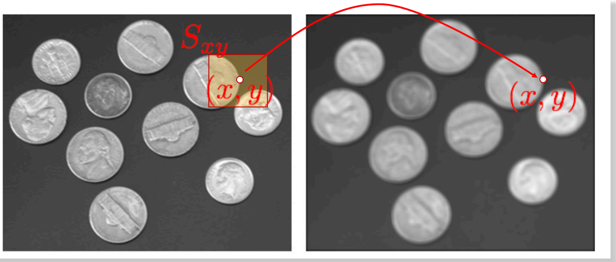
$$\underbrace{I'(x,y)}_{\textit{Filtered image}} = F \circ I = \sum_{i=-n}^{n} \sum_{j=-m}^{m} \underbrace{F(i,j)}_{\textit{Filter mask}} \underbrace{I(x+i, y+j)}_{\textit{Original image}}$$

- Filter F (of size $(2N+1)x(2M+1)$) is usually called a mask, kernel, or window
- Dealing with boundaries: e.g., pad, crop, extend, or wrap

Filter example #1: moving average

- The moving average filter returns the average of the pixels in the mask
- Achieves a smoothing effect (removes sharp features)
- E.g., for a *normalized* 3x3 mask

$$F = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$



Generated with a 5x5 mask

Filter example #2: Gaussian smoothing

- Gaussian function

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2 + y^2}{2\sigma^2})$$

- To obtain the mask, sample the function about its center
- E.g., for a normalized 3x3 mask with $\sigma = 0.85$

$$G = \frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

Convolution

- Still a linear filter, defined as

$$I'(x,y) = F * I = \sum_{i=-n}^{n} \sum_{j=-m}^{m} F(i,j) I(x-i, y-j)$$

- Same as correlation, but with negative signs for the filter indices
- Correlation and convolution are identical when the filter is symmetric
- Convolution enjoys the associativity property

$$F * (G * I) = (F * G) * I$$

- Example: to smooth an image & take its derivative = create a combined filter by convolving a derivative filter with a Gaussian filter & convolving the resulting combined filter directly with the image to achieve smoothing and differentiation in one step

Separability of masks

- A mask is separable if it can be broken down into the convolution of two kernels
$$F = F_1 * F_2$$
- If a mask is separable into "smaller" masks, then it is often cheaper to apply F_1 followed by F_2, rather than F directly
- Special case: mask representable as outer product of two vectors (equivalent to two-dimensional convolution of those two vectors)
- If mask is $M \times M$, and image has size $w \times h$, then complexity is
  - $O(M^2wh)$ with no separability
  - $O(2Mwh)$ with separability into outer product of two vectors

Example of separable masks

- Moving average

$$F = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \cdot \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

- Gaussian smoothing

$$\begin{aligned} G_\sigma(x, y) &= \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2 + y^2}{2\sigma^2}) \\ &= \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2}{2\sigma^2})\frac{1}{2\pi\sigma^2}\exp(-\frac{y^2}{2\sigma^2}) \\ &= g_\sigma(x) \cdot g_\sigma(y) \end{aligned}$$

Differentiation

Used to detect gradients and edges in the x and y-directions of an image

- Derivative of discrete function (centered difference)

$$\frac{\delta I}{\delta x} = I(x+1, y) - I(x-1, y) \qquad [1\,0-1]$$

$$\frac{\delta I}{\delta y} = I(x, y+1) - I(x, y-1) \qquad F_x = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

- Derivative as a convolution operation; e.g., Sobel masks:

$$S_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \qquad S_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{bmatrix}$$

Note: masks are **mirrored** in convolution

Along x direction

Along y direction

Similarity measures

- Filtering can also be used to determine similarity across images (e.g., to detect correspondences)

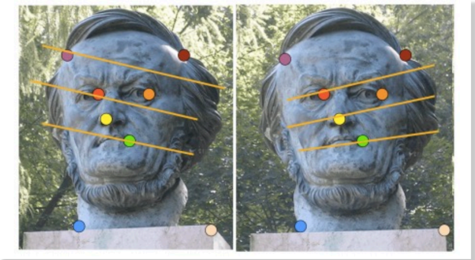
$$SAD = \sum_{i=-n}^{n} \sum_{j=-m}^{m} |I_1(x+i, y+j) - I_2(x'+i, y"+j)| \qquad \sum \textit{absolute differences}$$

$$SAD = \sum_{i=-n}^{n} \sum_{j=-m}^{m} [I_1(x+i, y+j) - I_2(x'+i, y"+j)]^2 \qquad \sum \textit{squared differences}$$

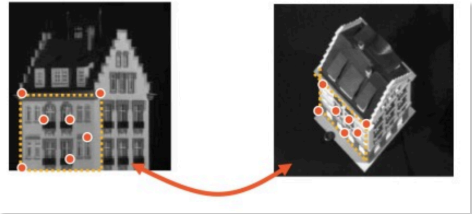
Detectors

- **Goal:** detect **local features**, i.e., image patterns that differ from immediate neighborhood in terms of intensity, color, or texture
- We will focus on
  - Edge detectors
  - Corner detectors

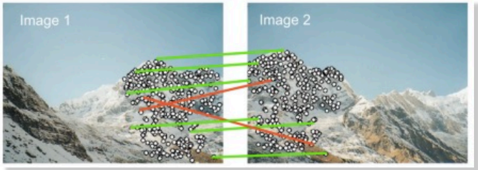
Use of detectors/descriptors: examples



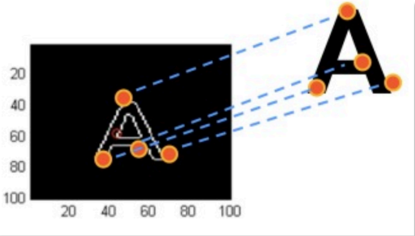
Stereo reconstruction



Estimating homographic transformations



Panorama stiching



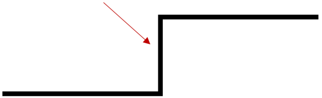
Object detection

Edge detectors

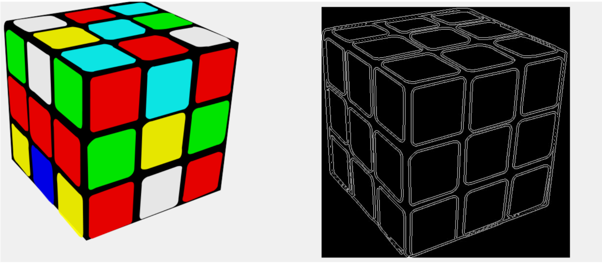
- **Edge:** region in an image where there is a significant change in intensity values along one direction, and negligible change along the orthogonal direction

In 1D

Magnitude of 1st order derivative is large, 2nd order derivative is equal to zero



In 2D



Criteria for "good" edge detection

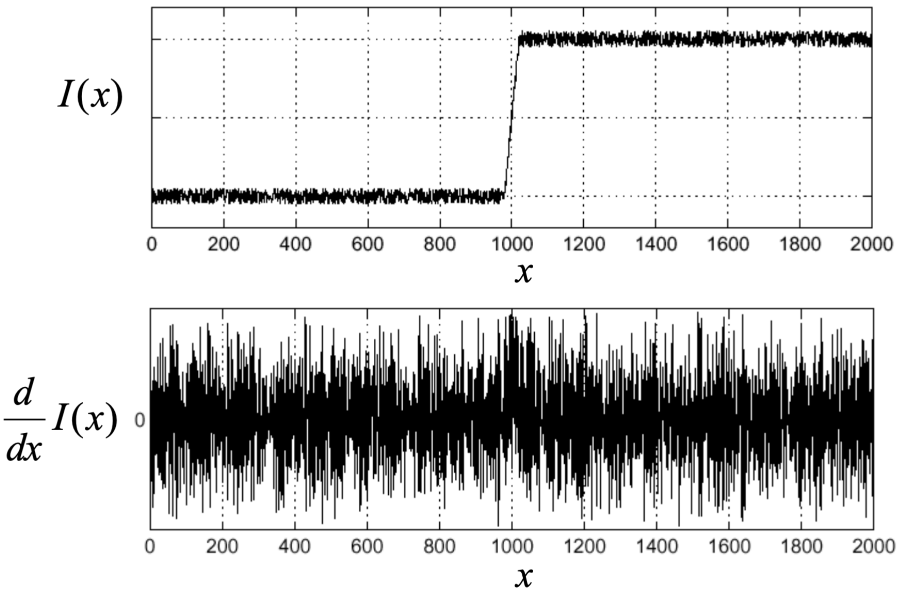
- **Accuracy:** minimize false positives and negatives
- **Localization:** edges must be detected as close as possible to the true edges
- **Single response:** detect one edge per real edge in the image

Strategy to design an edge detector

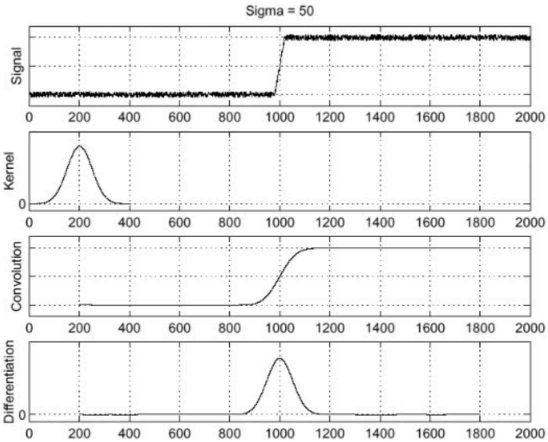
Two steps:

- **Smoothing:** smooth the image to reduce noise prior to differentiation (step 2)
- **Differentiation:** take derivatives along x and y directions to find locations with high gradients

1D case: differentiation without smoothing



1D case: differentiation with smoothing



$I(x)$

$g_\sigma(x)$

$s(x) = g_\sigma(x) * I(x)$

$s'(x) = \frac{d}{dx} * s(x)$

Edges occur at maxima or minima of $s'(x)$

A better implementation

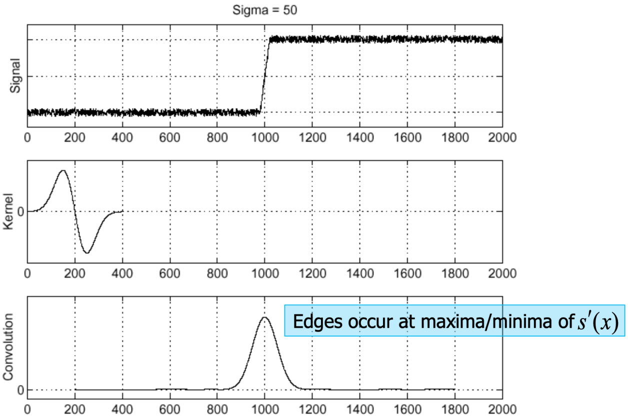
- Convolution theorem:

$$s'(x) = \frac{\delta}{\delta x} * (g_\sigma(x) * I(x)) = \underbrace{(\frac{\delta}{\delta x} * g_\sigma(x))}_{g'_\sigma(x)} * I(x)$$

$I(x)$

$g'_\sigma(x) = \frac{d}{dx} g_\sigma(x)$

$s'(x) = g'_\sigma(x) * I(x)$



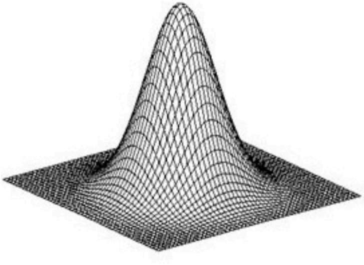
Edge detection in 2D

1. Find the gradient of smoothed image in both directions

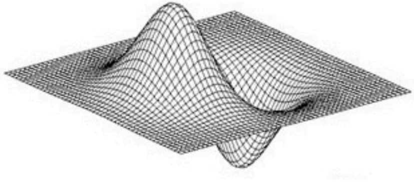
$$\nabla S := \begin{bmatrix} \frac{\delta}{\delta x} * (G_\sigma * I) \\ \frac{\delta}{\delta y} * (G_\sigma * I) \end{bmatrix} = \begin{bmatrix} (\frac{\delta}{\delta x} * G_\sigma) * I) \\ (\frac{\delta}{\delta y} * (G_\sigma) * I) \end{bmatrix} = \begin{bmatrix} (G_{\sigma,x}) * I) \\ (G_{\sigma,y}) * I) \end{bmatrix} := \begin{bmatrix} S_x \\ S_y \end{bmatrix}$$

2. Compute the magnitude $|\nabla S| = \sqrt{S_x^2 + S_y^2}$ and discard pixels below a certain threshold
3. Non-maximum suppression: identify local maxima of $|\nabla S|$

Derivative of Gaussian filter

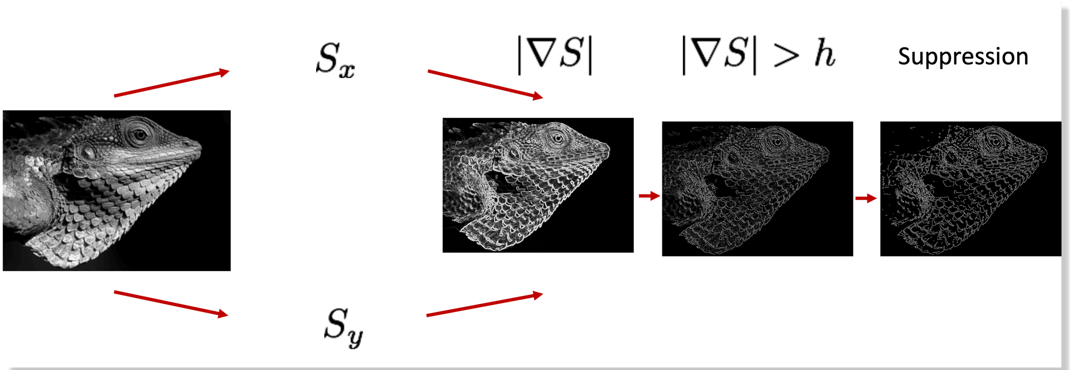


$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$



$$\frac{\partial G_\sigma(x, y)}{\partial x}$$

Canny edge detector



Corner detectors

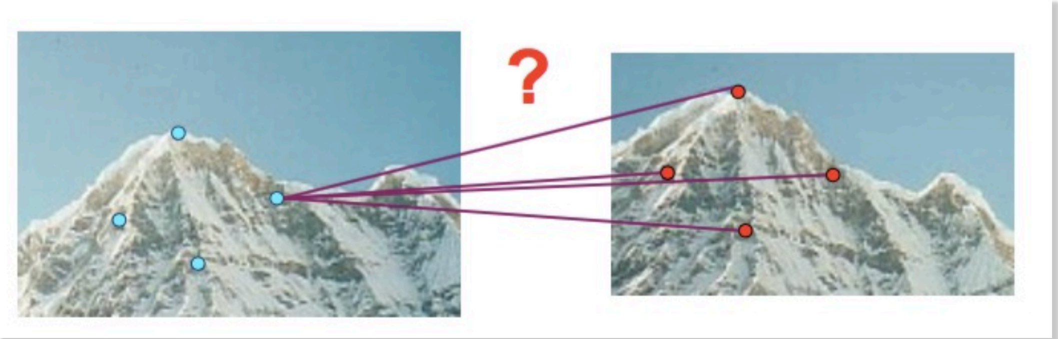
Key criteria for “good” corner detectors

- **Repeatability:** same feature can be found in multiple images despite geometric and photometric transformations
- **Distinctiveness:** information carried by the patch surrounding the feature should be as distinctive as possible

Repeatability

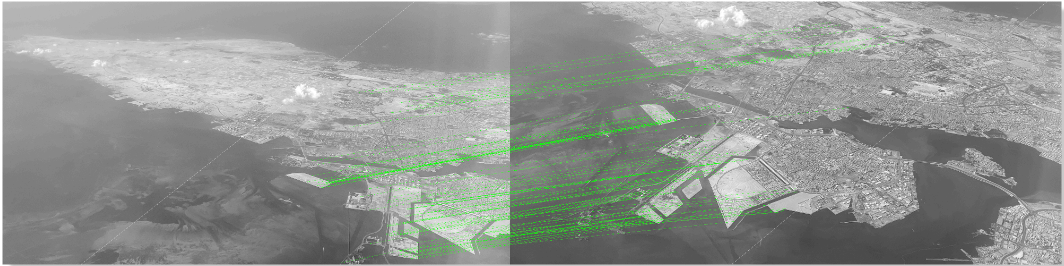
Without repeatability, matching is impossible

Distinctiveness



Without distinctiveness, it is not possible to establish reliable correspondences; distinctiveness is key for having a useful descriptor

Panorama Stiching

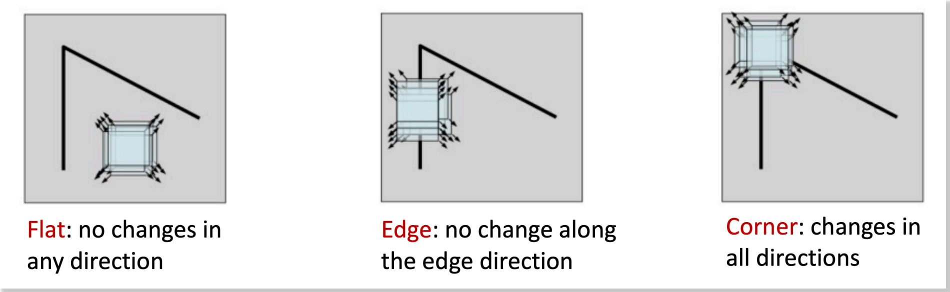


SIFT for feature extraction, flann-based matcher instead of brute-force, maxlines=100

Corner detectors

Key criteria for "good" corner detectors

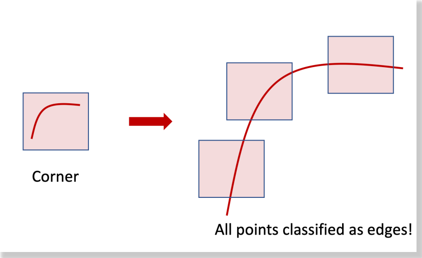
- **Corner:** intersection of two or more edges
- Geometric intuition for corner detection: explore how intensity changes as we shift a window



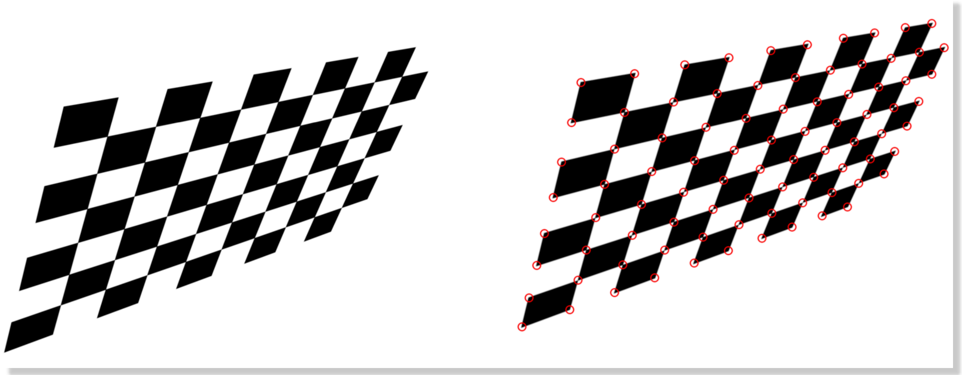
Harris detector: example

Properties of Harris detectors

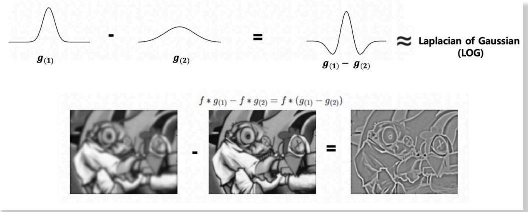
- Widely used
- Detection is invariant to
  - Rotation $\rightarrow$ geometric invariance
  - Linear intensity changes $\rightarrow$ photometric invariance
- Detection is not invariant to
  - Scale changes
  - Geometric affine changes
- Scale-invariant detection, such as
  1. Harris-Laplacian
  2. in SIFT (specifically, Difference of Gaussians (DoG))

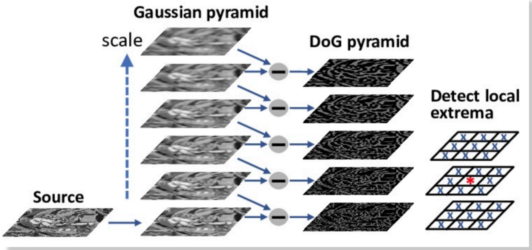


Example Application of Corner Detector



Difference of Gaussians (DoG)

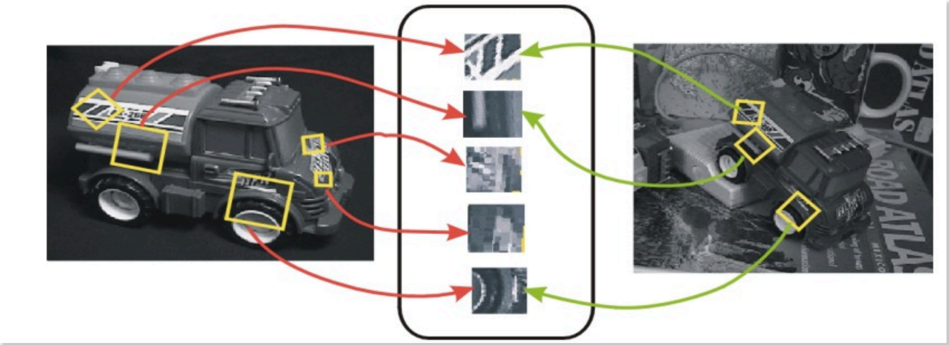




- Features are detected as local extrema in scale and space

Descriptors

- **Goal:** describe keypoints so that we can compare them across images or use them for object detection or matching
- Desired properties:
  - Invariance with respect to pose, scale, illumination, etc.
  - Distinctiviness



Simplest descriptor

- Naive descriptor: associate with a given keypoint an $n \times m$ window of pixel intensities centered at that keypoint
- Window can be normalized to make it invariant to illumination

Main drawbacks
1. Sensitive to pose
2. Sensitive to scale
3. Poorly distinctive

Popular Detectors / Descriptors

- SIFT
  - Invariant to rotation and scale, computationally demanding
  - SIFT descriptor is a 128-dimensional vector!
- SURF
- FAST
- BRIEF
- ORB
- BRISK
- LIFT

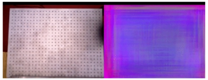
	Corner detector	Blob detector	Rotation invariant	Scale invariant	Affine invariant	Repeatability	Localization accuracy	Robustness	Efficiency
Harris	x		x			+++	+++	++	++
Shi-Tomasi	x		x			+++	+++	++	++
Harris-Laplacian	x	x	x	x		+++	+++	++	+
Harris-Affine	x	x	x	x	x	+++	+++	++	++
SUSAN	x		x			++	++	++	+++
FAST	x		x			++	++	++	++++
SIFT		x	x	x	x	+++	++	+++	+
MSER		x	x	x	x	+++	+	+++	+++
SURF		x	x	x	x	++	++	++	++

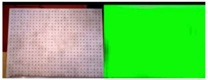
Case study

A case study for learning-based Descriptors Dense Object Nets

Learning *Dense* Visual Object *Descriptors*
By and *For* Robotic Manipulation. CORL 2018

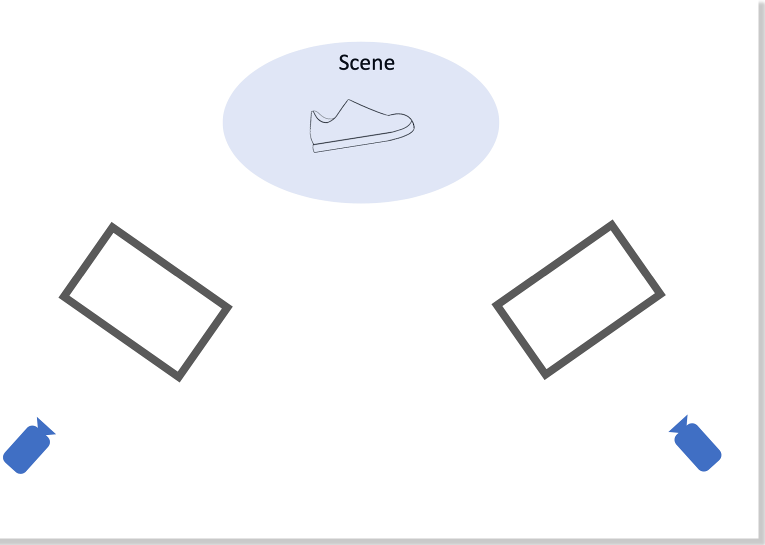
Peter R. Florence, Lucas Manuelli, Russ Tedrake



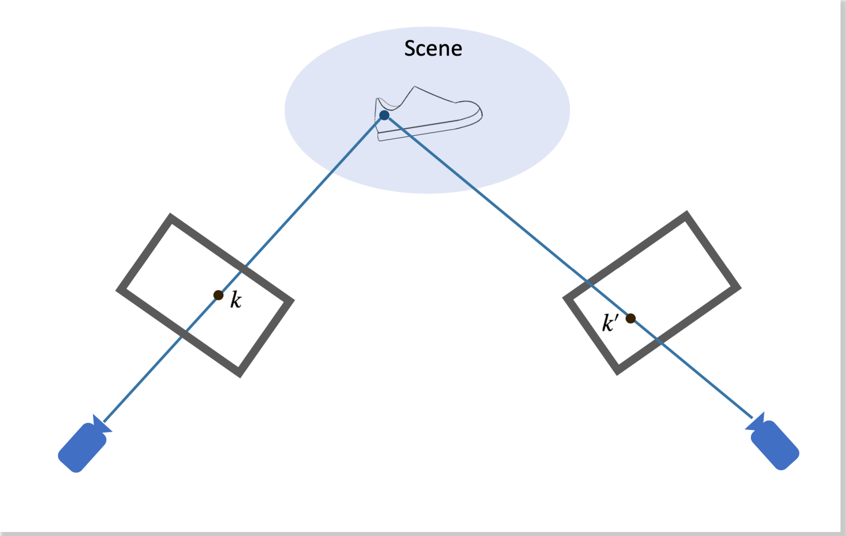


Slides adapted from CS326 by Kevin Zakka and Sriram Somasundaram

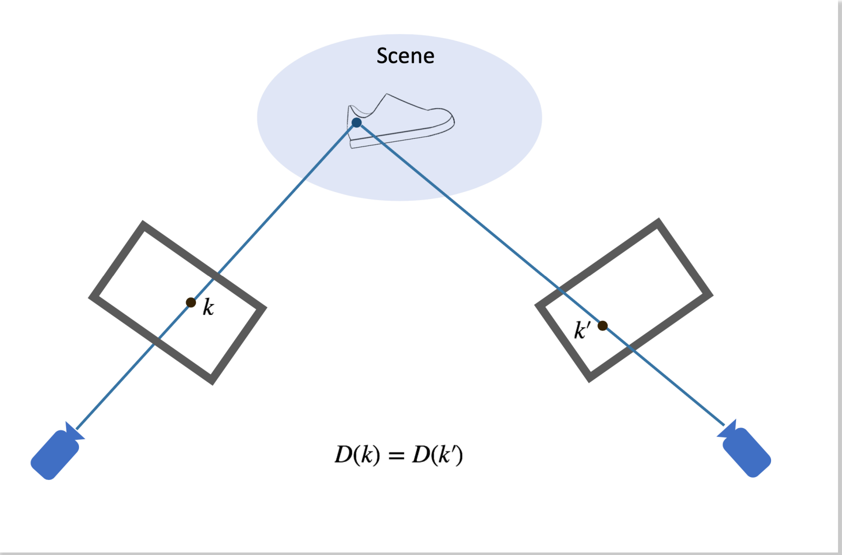
Case study



Case study



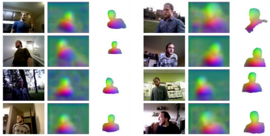
Case study



Case study

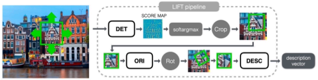
Brief history

Sparse Engineered: SIFT



Dense Learned

Sparse Learned: LIFT



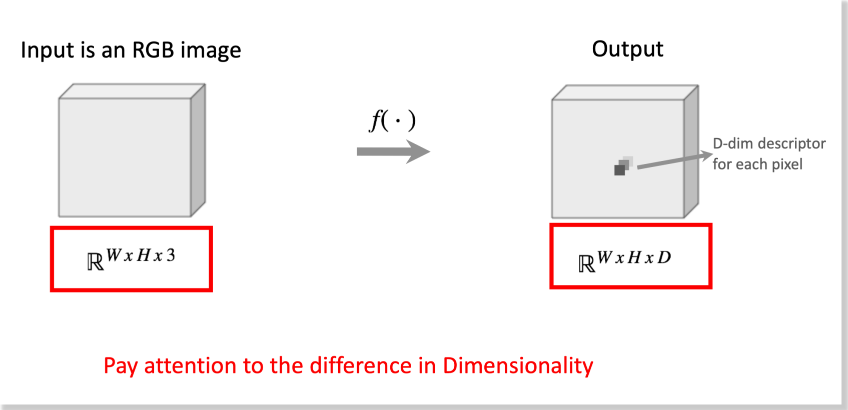
Case study

Why dense?

Bachrach et. al.

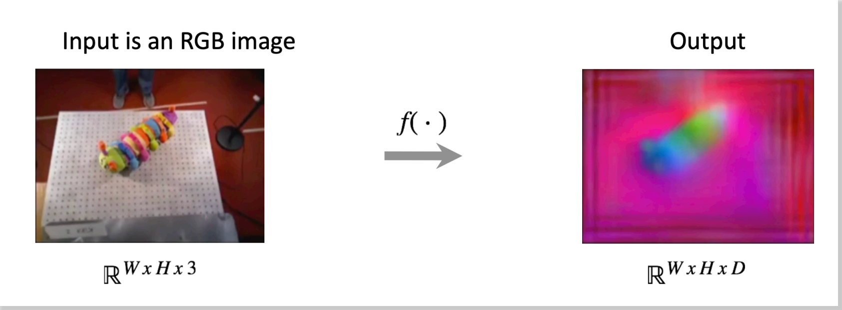
Case study

Dense descriptors



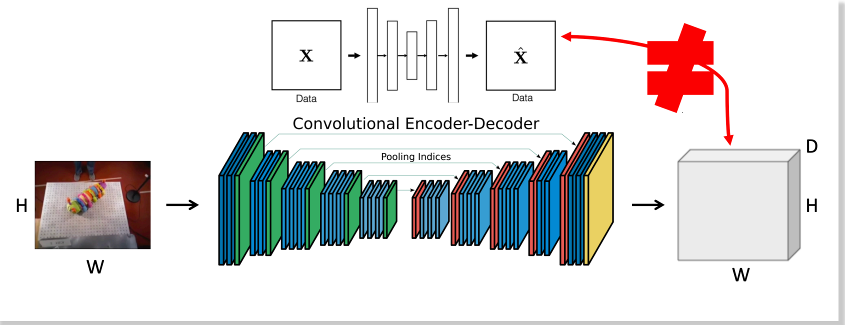
Case study

Dense descriptors



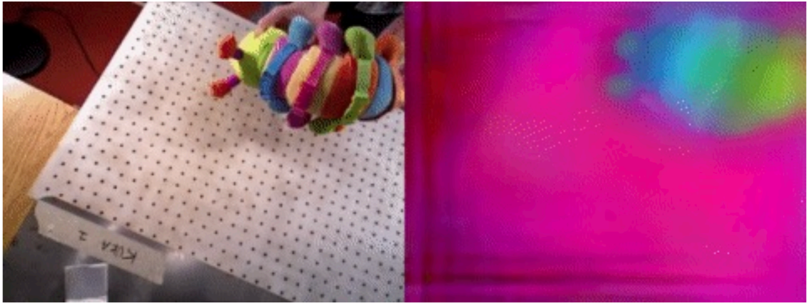
Case study

Network Architecture



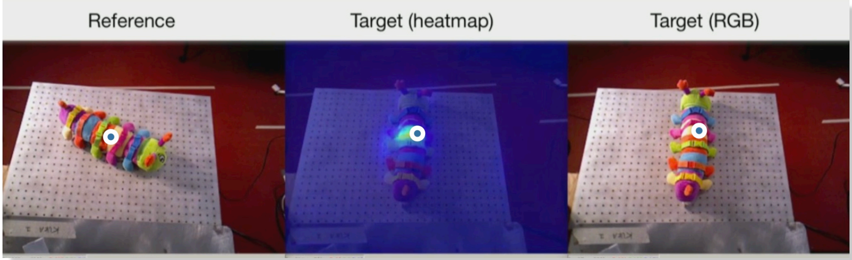
Case study

Single object



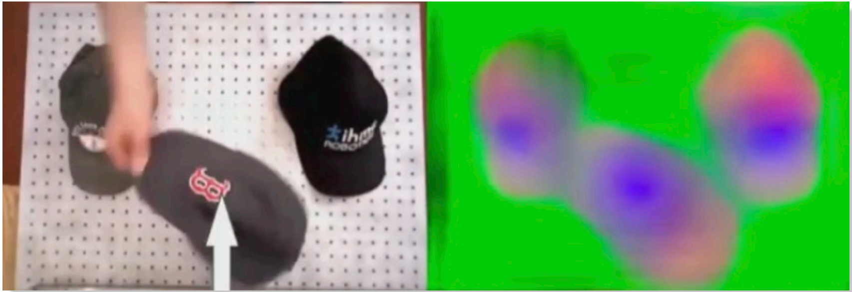
Case study

Learned Dense Correspondences

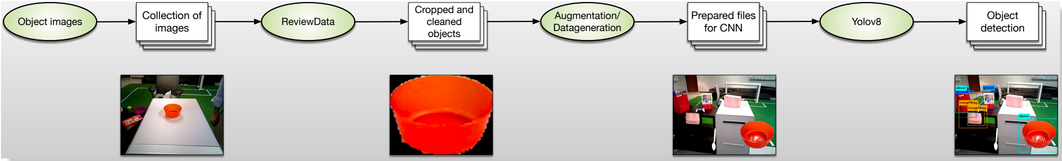


Case study

Class consistent descriptors



RoboCanes vision pipeline, based on Yolov8 (Ultralytics)



Acknowledgements

Acknowledgement

This slide deck is based on material from the Stanford ASL and ETH Zürich

References

D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: speeded up robust features," in *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3951. Springer, 2006, pp. 404–417. [Online]. Available: https://doi.org/10.1007/11744023_32