Convolutional neural networks

AMNLima

DEE/CEEI/UFCG

May 25, 2020

The Mammalian Visual System

- Used to perceive, plan, and act.
- There are not enough neurons (look-up table).
- Visual animals encode visual information using minimal numbers of neurons (hundreds of millions of neurons and billions of connections!).
- An visual image is essentially broken down into simple elements that are reconstructed through a series of processing stages, most of which occur beneath consciousness.
- Process information hierarchically along the visual brain structures.

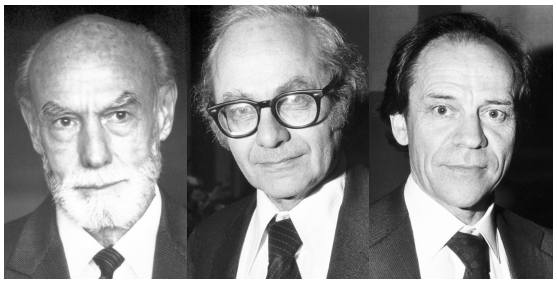


Figure 1: The Nobel Prize in Physiology (Medicine) 1981: Roger W. Sperry (1/2-specialization of the cerebral hemispheres) and David H. Hubel and Torsten N. Wiesel (1/2-information processing in the visual system).

Brain layers

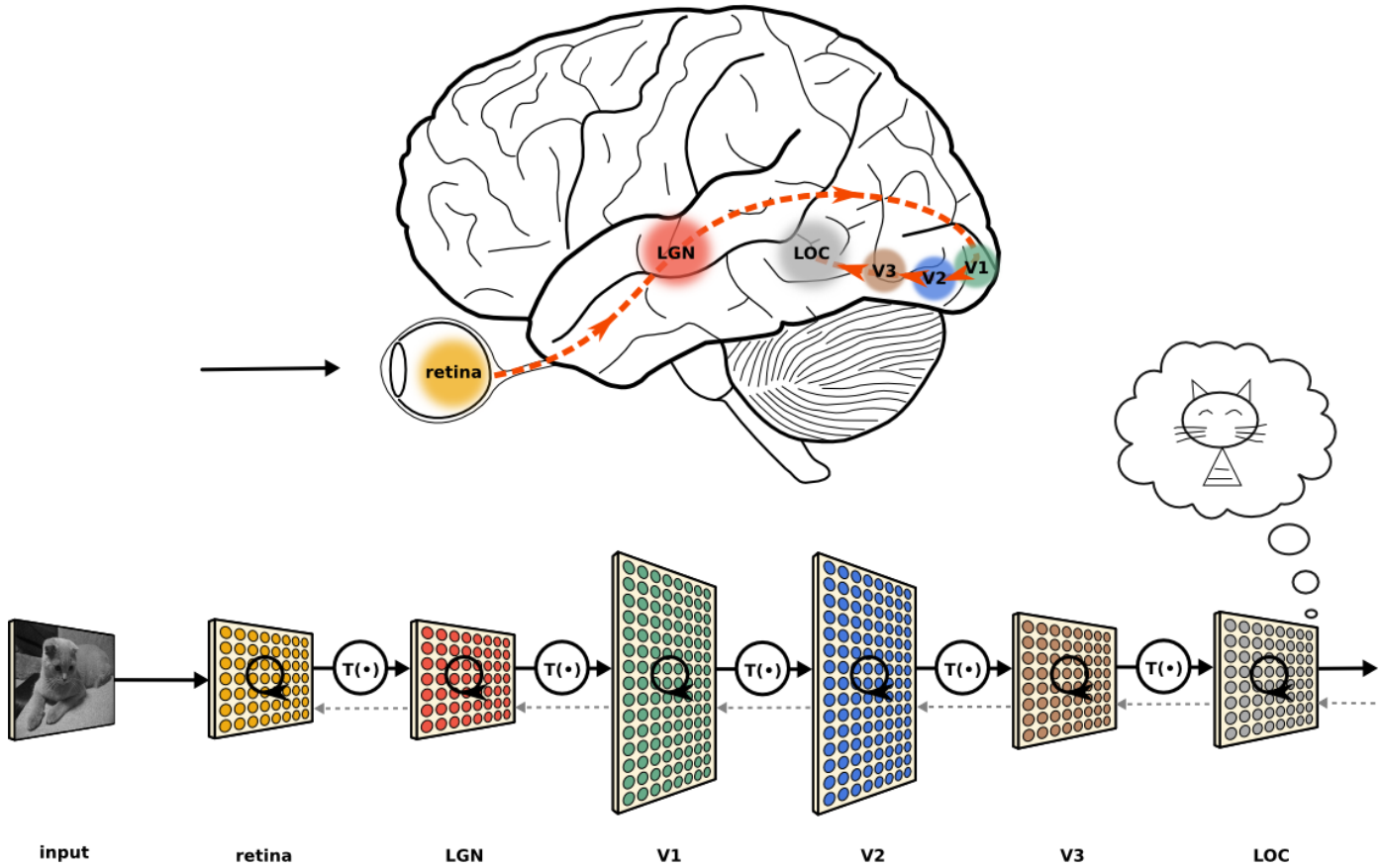


Figure 2: The visual processing in the brain is hierarchical, i.e., one layer feeds into the next, computing progressively more complex features (level of abstraction grows from line, edge, boundary, shape, ...). Lateral geniculate nucleus (LGN), primary visual cortex (V1), secondary visual cortices (V2, V3, V4, ...), and high-level (LOC).

Image formation

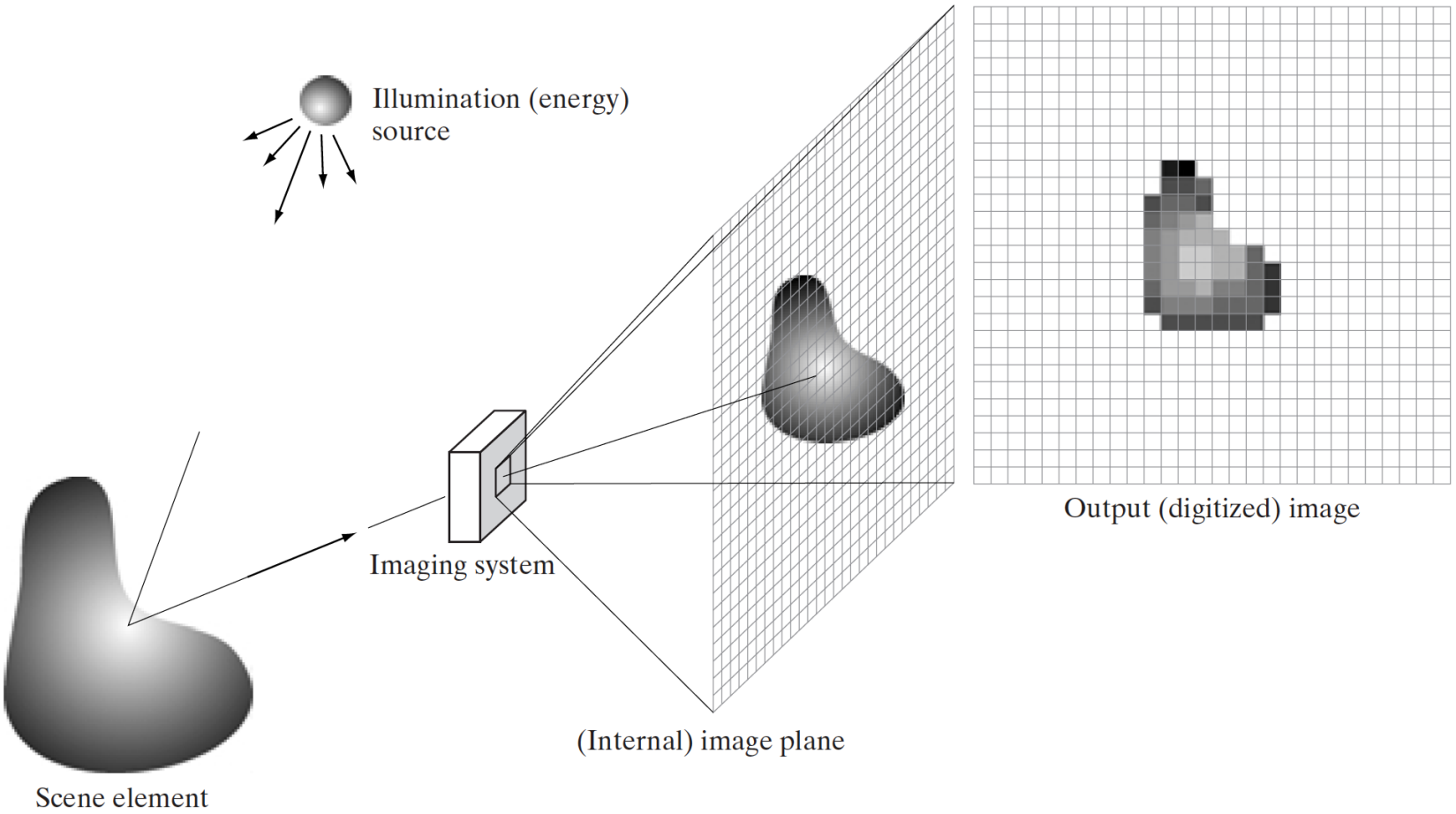


Figure 3: Digital image acquisition process: (a) Energy ("illumination") source. (b) An element of a scene. (c) Imaging system. (d) Projection of the scene onto the image plane. (e) Digitized image.

Image formation model

- An image is a two-dimensional function $f(x, y)$ where (x, y) are spatial coordinates and the value or amplitude of $f(\cdot, \cdot)$ is a positive scalar quantity whose physical meaning is determined by the source of the image.
- When an image is generated from a physical process, the $f(\cdot, \cdot)$ values are proportional to energy radiated by a physical source. Thus, $f(x, y)$ must be nonzero and finite, i.e.,

$$0 < f(x, y) < \infty, \forall x, y \in \mathbb{R}.$$

- The function $f(x, y)$ has two components: (1) the amount of source illumination incident on the scene being viewed $i(x, y)$, and (2) the amount of illumination reflected by the objects in the scene $r(x, y)$, i.e.,

$$f(x, y) = i(x, y)r(x, y),$$

where

$$0 < i(x, y) < \infty \text{ and } 0 < r(x, y) < 1.$$

Monochrome image

- The intensity of an image at any coordinates (x_0, y_0) is given by

$$\mathcal{L} = f(x_0, y_0),$$

where

$$L_{\text{min}} \leq \mathcal{L} \leq L_{\text{max}}.$$

- The interval $[L_{\text{min}}, L_{\text{max}}]$ is called the gray scale.
- The common practice is to shift this interval numerically to the interval $[0, L-1]$, where $\mathcal{L} = 0$ is considered black and $\mathcal{L} = L-1$ is considered white on the gray scale.
- All intermediate values are shades of gray varying from black to white.

Continuous image

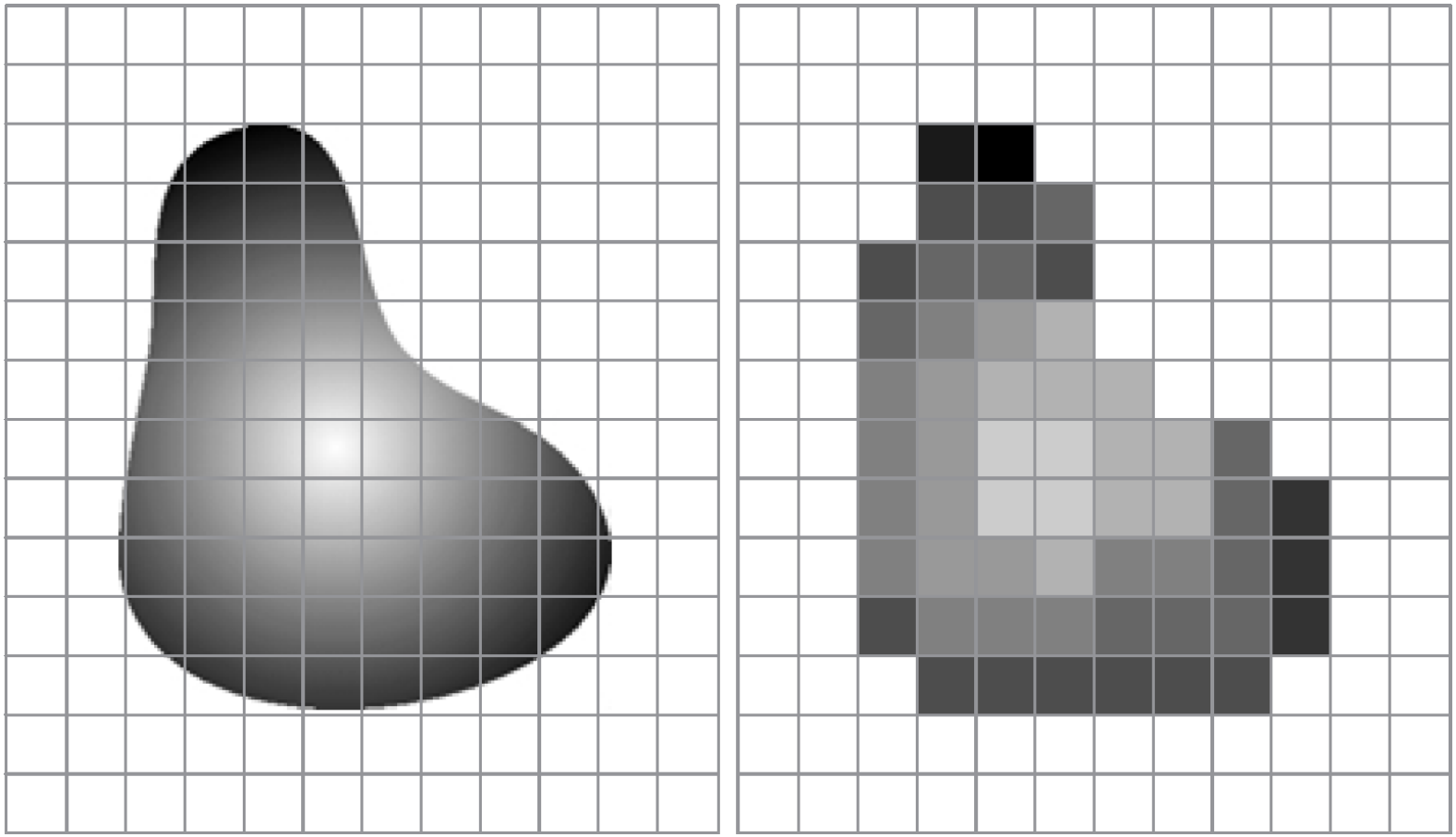


Figure 4: Continuous image projected onto a sensor array. Sampling and quantizated image. The sampling rate in images is the number of samples taken (in both spatial directions) per unit distance. Moiré patterns are observed in digital images when repetitive pattern of high spatial frequency is sampled at low resolution.

Digital image

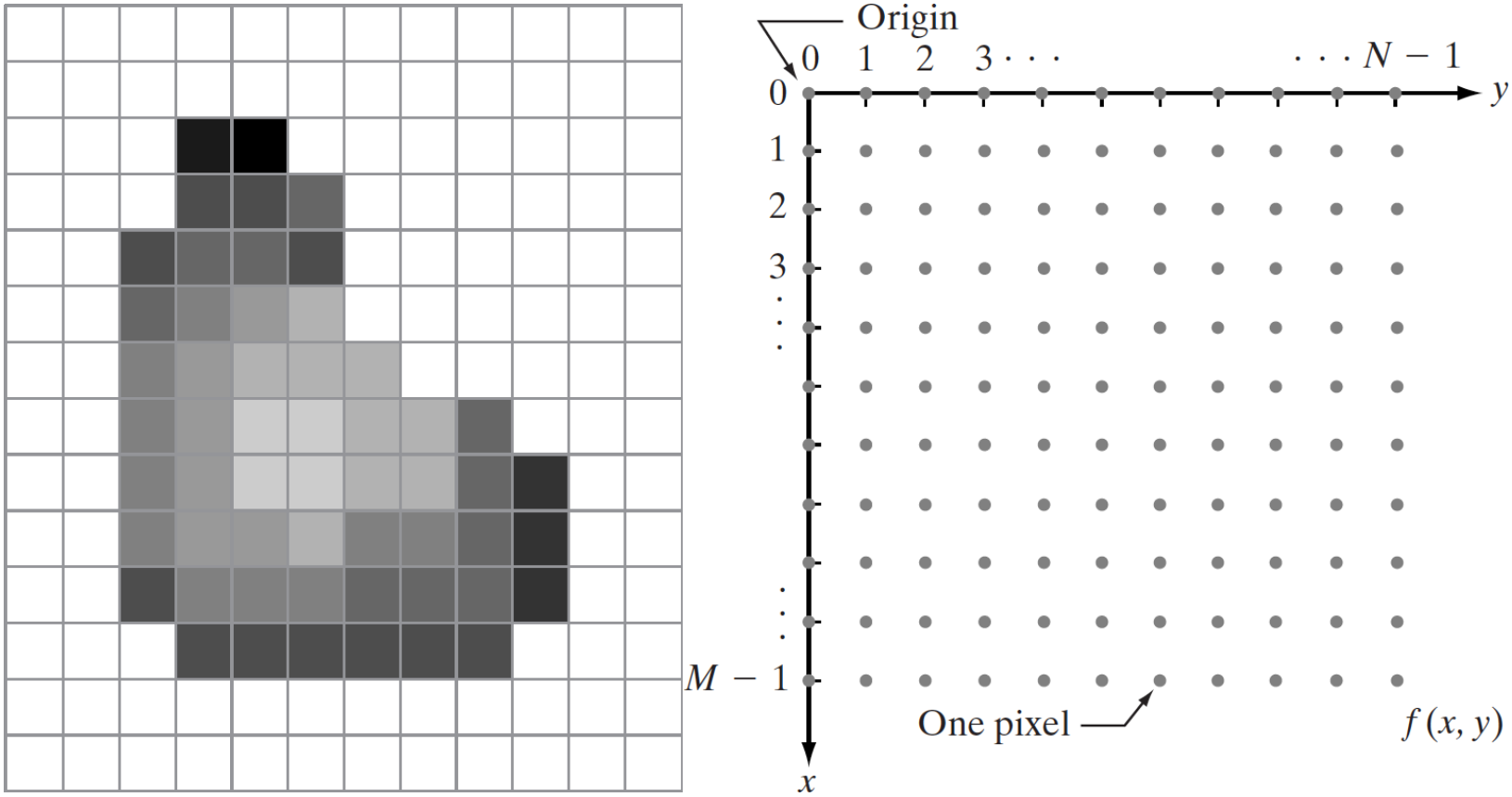


Figure 5: Digital image and a possible coordinate convention used to represent digital images.

Representing a digital image

- The result of sampling and quantization is a matrix of real numbers, and a sampled image has M rows and N columns, i.e.,

$$\mathcal{I} = \begin{bmatrix} f(0,0) & f(0,1) & \cdots & f(0,N-1) \\ f(1,0) & f(1,1) & \cdots & f(1,N-1) \\ \vdots & \vdots & \vdots & \vdots \\ f(M-1,0) & f(M-1,1) & \cdots & f(M-1,N-1) \end{bmatrix},$$

$$L = 2^k, B = M \times N \times k.$$

- Each element of matrix I is called an image element, picture element, or pixel.
- The spatial resolution is the smallest discernible detail in an image, i.e., a line of width W, with the space between the lines also having width W. The width of a line pair is 2W, and there are 1/2/W line pairs per unit distance. The resolution is given by smallest number of discernible line pairs per unit distance, e.g., 100 line pairs per mm.

Image sizes

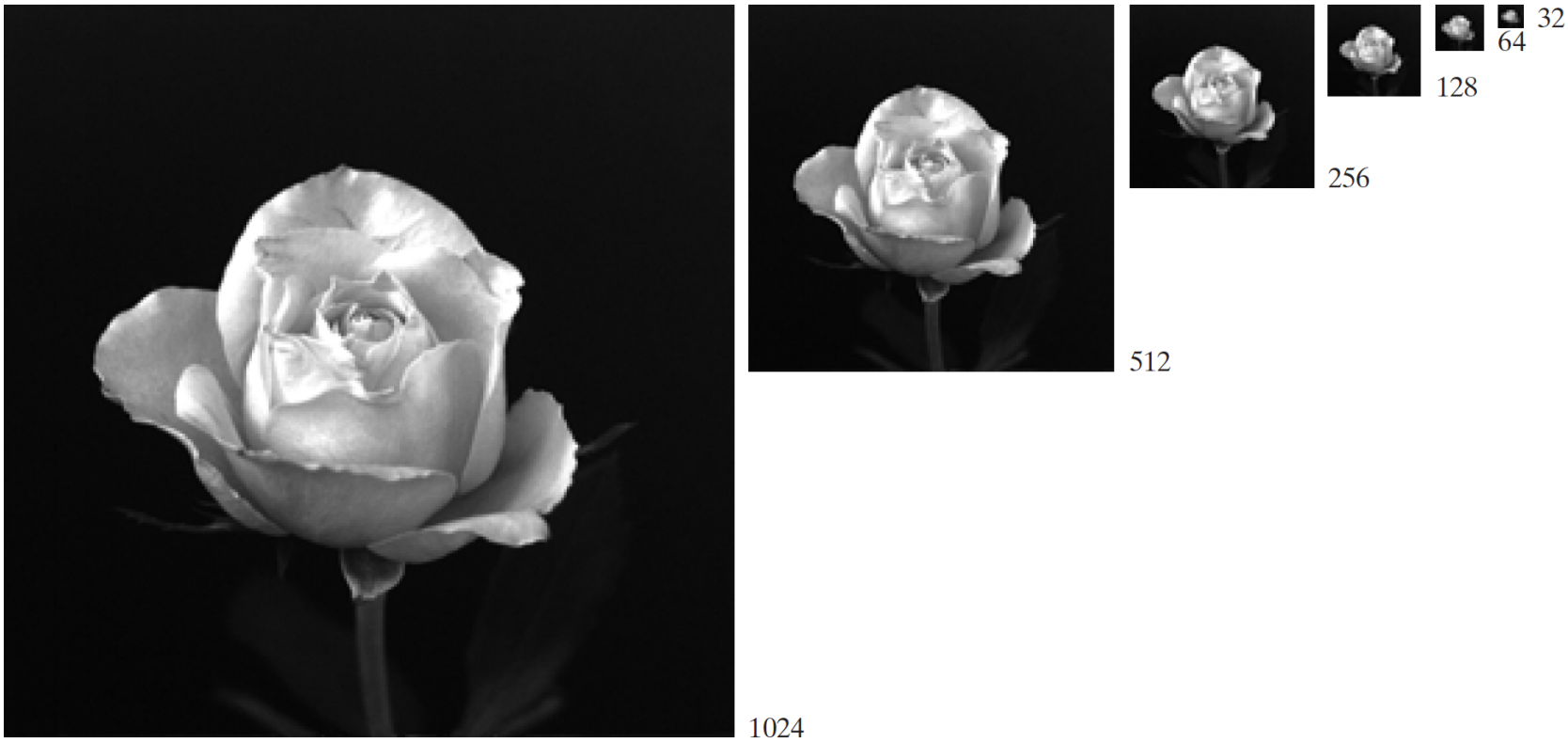
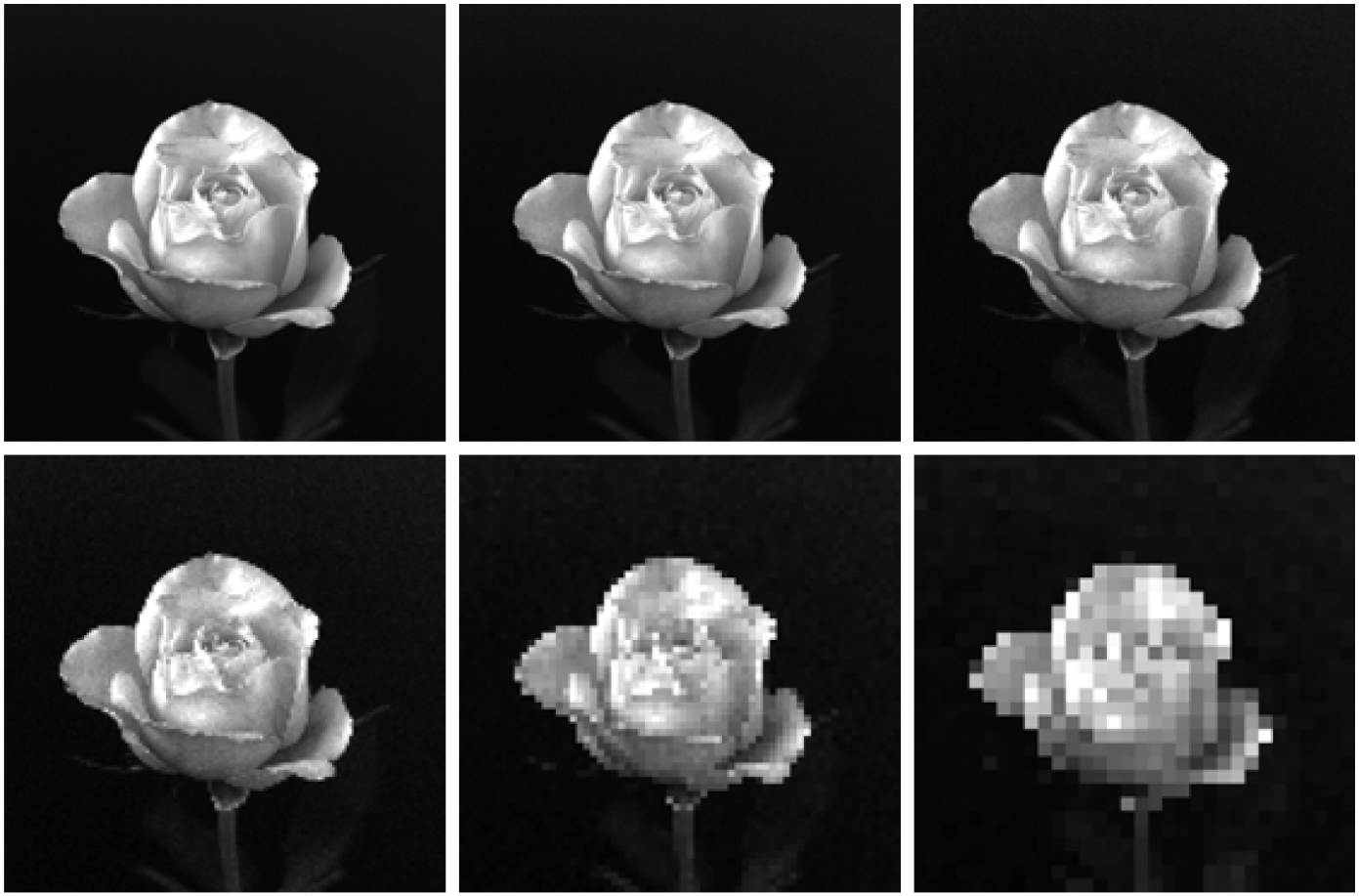


Figure 6: A $1024(M) \times 1024(N)$, 8(k)-bit image subsampled down to size 32×32 pixels for $L = 256$. Zooming may be viewed as oversampling, while shrinking may be viewed as undersampling.

- Gray-level resolution is the smallest discernible change in gray level.

Image sizes+



a b c
d e f

Figure 7: (a) 1024×1024, 8-bit image. (b) 512×512 image resampled into 1024×1024 pixels by row and column duplication. (c) through (f) 256×256, 128×128, 64×64, and 32×32 images resampled into 1024×1024 pixels.

Relationships between pixels

- Neighbors: a pixel p located at (x, y) has four horizontal and vertical neighbors as well as four diagonal neighbors, i.e.,

$$N_4(p) = \{(x+1, y), (x-1, y), (x, y+1), (x, y-1)\},$$

$$N_D(p) = \{(x+1, y+1), (x-1, y-1), (x-1, y+1), (x-1, y-1)\},$$

$$N_8(p) = N_D(p) \cup N_4(p).$$

- Adjacency: V is the set of gray-level values used to define the adjacencies.
  - 4-adjacency: two pixels p and q are 4-adjacent if $q \in N_4(p)$.
  - 8-adjacency: two pixels p and q are 4-adjacent if $q \in N_8(p)$.
  - m-adjacency: two pixels p and q are m-adjacent if (i) $q \in N_4(p)$, or (ii) $q \in N_D(p)$ and the set $N_4(p) \cap N_4(q)$ has no pixels whose values are from V.

Relationships between pixels+

- Connectivity: two pixels p and q are said to be connected in a subset S (subset of pixels in an image) if there exists a path between them consisting entirely of pixels in S.
- A path, $\mathcal{P}$, from a pixel p with coordinates (x, y) to a pixel q with coordinates (s, t) is a sequence of distinct pixels with coordinates

$$\mathcal{P} = \{(x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)\},$$

  where $(x_0, y_0) = (x, y)$, $(x_n, y_n) = (s, t)$, and pixels (x_i, y_i) and (x_{i-1}, y_{i-1}) adjacent for $1 \leq i \leq n$; n is the length of the path, and if $(x_0, y_0) = (x_n, y_n)$ it is said closed.
- Two pixels p and q are said to be connected in S if there exists a path between them consisting entirely of pixels in S. $\forall p \in S$, the set of pixels that are connected to it in S is called a connected component of S; if it only has one connected component, then set S is called a connected set.

Relationships between pixels++

- Boundary: The boundary (also called border or contour) of a region R is the set of pixels that have one or more neighbors that are not in R.
- $R \in \mathcal{I}$ is said a region of $\mathcal{I}$ if R is a connected set.
- If $R = \mathcal{I}$ then its boundary is defined as the set of pixels in the first and last rows and columns of the image.
- Boundary versus Edge: The boundary of a finite region forms a closed path and is thus it is a "global" concept. The edge is formed from pixels with intensity discontinuities and thus, it is a "local" concept that is based on a measure of gray-level at a point.

Distance measure

- For pixels p, q, and z, with coordinates (x, y), (s, t), and (v, w), respectively, D is said a "distance function" or "metric" if

$$\begin{aligned}&\text{(a) } D(p,q) \geq 0, D(p,q) = 0 \text{ iff } p = q,\\ &\text{(b) } D(p,q) = D(q,p), \text{ and,}\\ &\text{(c) } D(p,z) \leq D(p,q) + D(q,z).\end{aligned}$$

- The Euclidean distance is defined as

$$D_e(p,q) = [(x-s)^2 + (y-t)^2]^{\frac{1}{2}},$$

  and the pixels such that $D_e(\cdot,\cdot) \leq r$ are contained in a disk of radius r centered at (x, y).

- The city-block distance is defined as

$$D_c(p,q) = |x-s| + |y-t|,$$

  and the pixels such that $D_c(\cdot,\cdot) \leq r$ are contained in a diamond shape centered at (x, y).

Linear and nonlinear operations

- $\mathcal{H} : \mathbb{R}^{M\times N} \longrightarrow \mathbb{R}^{M\times N}$ is an operator whose input and output are images. $\mathcal{H}$ is said to be a linear operator if, for any two images $\mathcal{I}_0$ and $\mathcal{I}_1$ and any two scalars $a, b \in \mathbb{R}$,

$$\mathcal{H}(a\mathcal{I}_0 + b\mathcal{I}_1) = a\mathcal{H}(\mathcal{I}_0) + b\mathcal{H}(\mathcal{I}_1).$$

An operator that fails such test is by definition nonlinear, like $s = T(r)$ ($r = T^{-1}(s)$) given below.

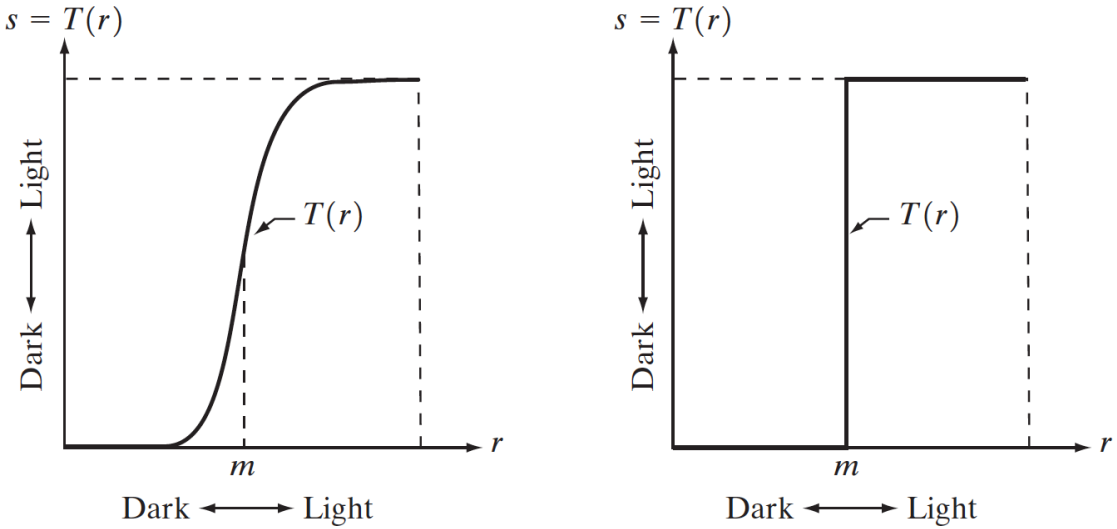


Figure 8: Gray level transformations for contrast enhancement. r is the pixel intensity at (x, y) that is mapped to s. Log transformation $T(r) = c\log(1 + r)$, power-law transformation $T(r) = c(r + \varepsilon)^{\gamma}$.

Spatial filtering

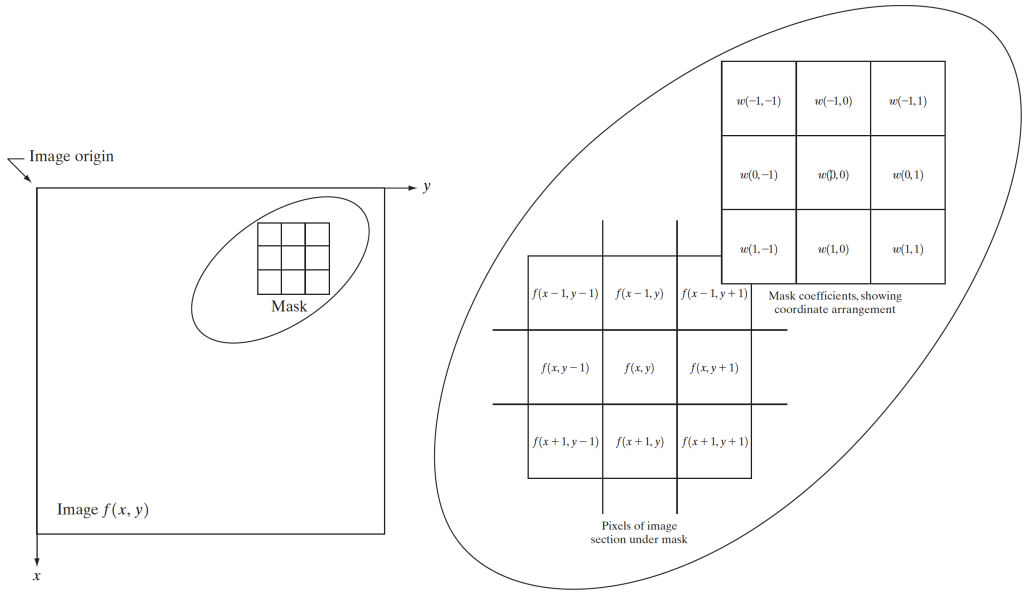


Figure 9: The mechanics of spatial filtering consists simply of moving the filter mask (w) from point to point in an image.

For linear spatial filtering the response is given by a sum of products of the filter coefficients and the corresponding image pixels in the area spanned by the filter mask, i.e.,

$$g(x,y) = w(-1,-1)f(x-1,y-1) + w(-1,0)f(x-1,y) + \cdots + w(0,0)f(x,y) + \cdots + w(1,0)f(x+1,y) + w(1,1)f(x+1,y+1).$$

Discrete linear spatial filtering

The discrete linear filtering of an image f of size $M \times N$ with a filter mask (kernel) w (κ) of size $m \times n$, $m < M$, $n < N$, generates an image g also of size $M \times N$ that is given by

$$g(x, y) = (\kappa \circledast f)(x, y),$$

or

$$g(x, y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} \kappa(s, t) f(x + s, y + t),$$

where $a = (m - 1)/2$ and $b = (n - 1)/2$. The Gaussian blur (also known as Gaussian smoothing) is the result of blurring an image by a Gaussian function given by

$$\kappa_\sigma^k(x, y) = \frac{1}{2\pi\sigma^2} \exp -\frac{x^2 + y^2}{2\sigma^2}, \kappa_g^3 = \begin{bmatrix} 0.1019 & 0.1154 & 0.1019 \\ 0.1154 & 0.1308 & 0.1154 \\ 0.1019 & 0.1154 & 0.1019 \end{bmatrix}.$$

The Gaussian blur is a low-pass filter, attenuating high frequency signals.

Lenna and Lenna blurred

Figure 10: The Lenna (or Lena) picture is one of the most widely used standard test images used for compression algorithms. Lena Soderberg was a Playmate in the November 1972 issue of Playboy magazine. The right hand side image was blurred with a Gaussian kernel.

Lenna and Lenna edged

Figure 11: The image of the right hand side were determined by the Sobel–Feldman kernel that creates an image emphasising the edges.

Image features

A feature is a piece of information which is relevant to a certain application.

Edge A point in the image where intensity changes; its a local intensity discontinuities which does not dependent on an object model.

Boundary A boundary is extensive, is composed of many edge points and may be dependent on an object model.

Knowledge about the boundaries

The ultimate goal of machine vision is to capture the semantic information from the digital images.
The semantic information allows one to determine: what objects are present, what the scene environment is, and what kind of activities are taking place.

If its high One may extract the complete closed contour of the object if its shape is known.

If its medium One may extract pieces of the boundary using general line or curve models.

If it is low One may just try to find a connected series of edge elements, using only heuristics on say, edge curvature.

Image processing operators

1. Arithmetic: pointwise combination of two images
2. Point: functions applied to individual pixels
3. Geometric: image rotation, translation and scaling
4. Analytical: labeling image pixels
5. Morphological: pixel shape based analysis
6. Filtering: noise reduction and other enhancement filters
7. Feature Detectors: edges and others features
8. Transforms: Fourier, Hough and other transforms
9. Synthesis: noise image data

Main problems

The three main problems in line extraction in unknown environment are:

- How many lines are there ?
- Which points belong to which line ?
- Given the points that belong to a line, how to estimate the line model parameters ?

Hough Transform

The Hough transform is a technique which can be used to isolate features of a particular shape within an image.
The classical Hough transform was conceived for the identification of lines in the image.
The Hough transform can be extended for identifying of arbitrary shapes.
The shapes must be specified in some parametric form and thus this transform is most commonly used for the detection of regular curves such circles and ellipses.

Line $x\cos(\theta) + y\sin(\theta) = \rho$

Circle $(x-a)^2 + (y-b)^2 = \rho^2$

Ellipse $\frac{(x-c)^2}{a^2} + \frac{(y-d)^2}{b^2} = 1$

Line space

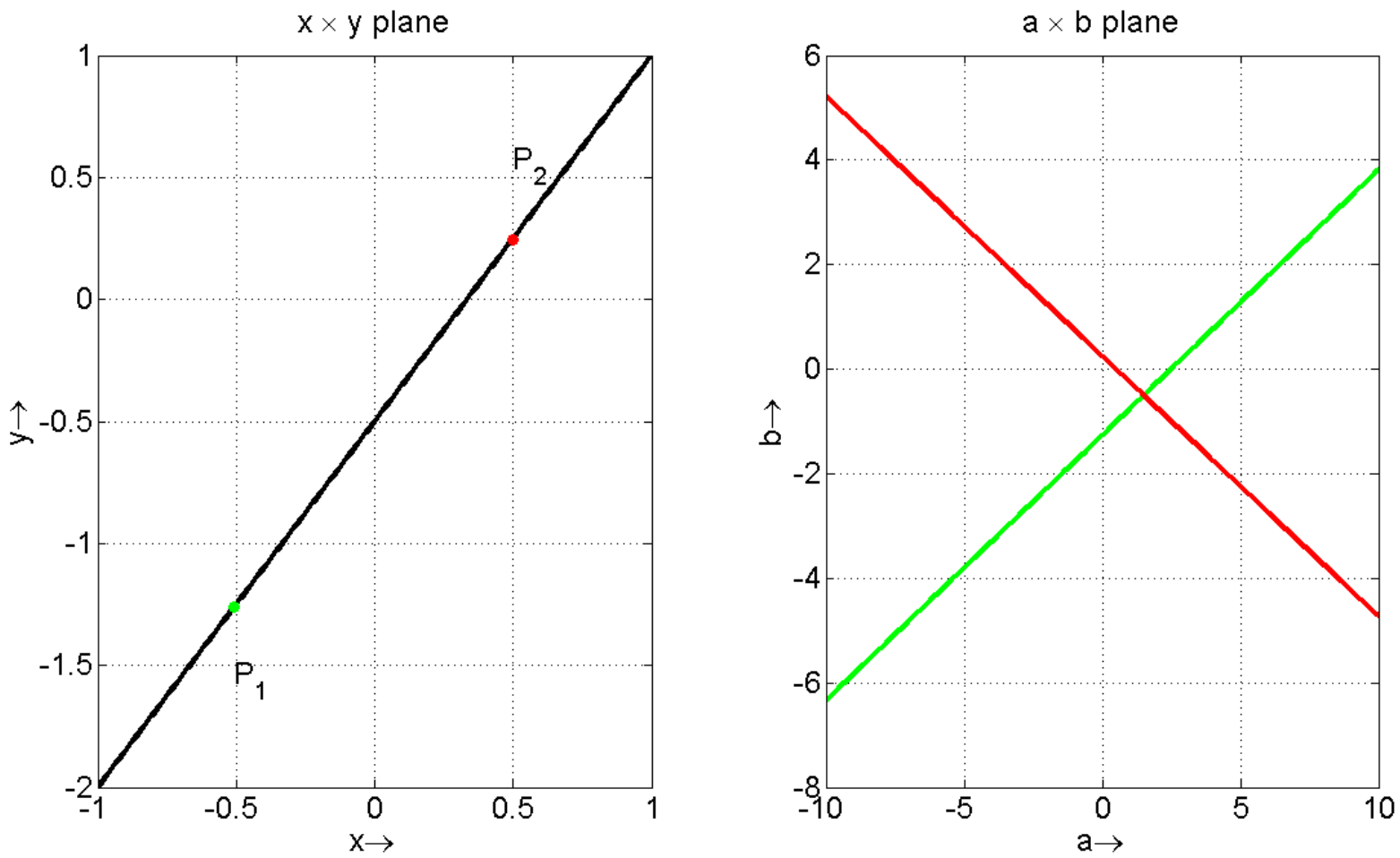


Figure 12: Straight line and line space.

Polar space

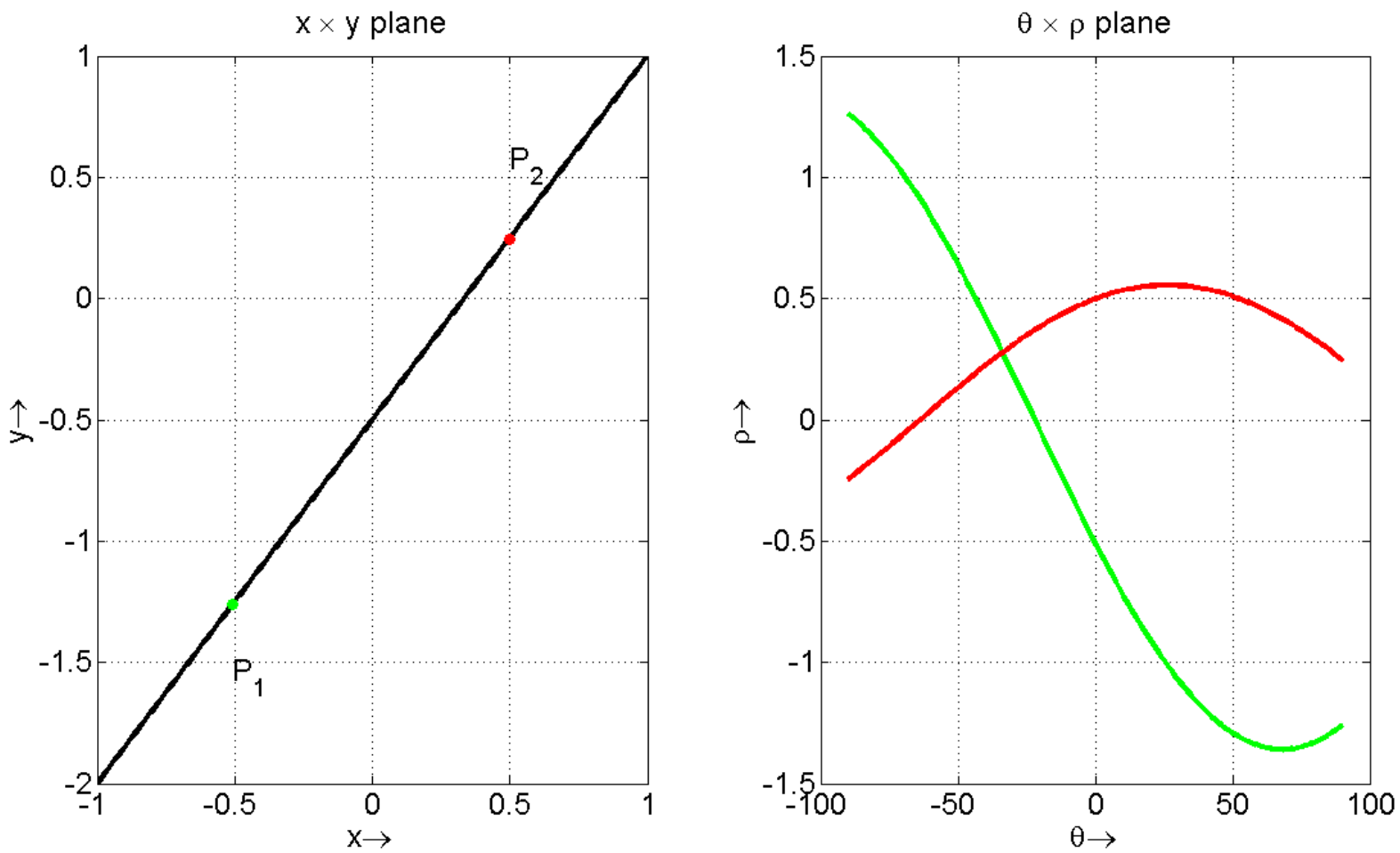


Figure 13: Straight line and polar space. From $y = ax + b$ to $\rho = x\cos(\theta) + y\sin(\theta)$.

Standard Hough Transform Algorithm

- Quantize the Hough parameter space (θ, ρ), i.e.,

$$\underbrace{[\theta_{\text{min}}, \theta_{\text{max}}]}_{A} \times \underbrace{[\rho_{\text{min}}, \rho_{\text{max}}]}_{B} = \{(\theta, \rho) \,|\theta \in A \text{ and } \rho \in B\}$$

- Initialize all of these cells with zero; this quantized space is referred to as the accumulator.
- Count the number of times a line intersects a given cell.
  - For each edge point (x_n, y_n) in the image, find the cell at the quantized Hough parameter space corresponding to that point.
  - Increase the value of the accumulator for such cell (**vote for it**).
  - Proceed with the next edge point (x_{n+1}, y_{n+1}) in the image.
- Cells receiving a minimum number of **votes** are assumed to correspond to lines in (x, y) space.
  - Lines can be found as peaks in the accumulator.

From brain cells to neural network

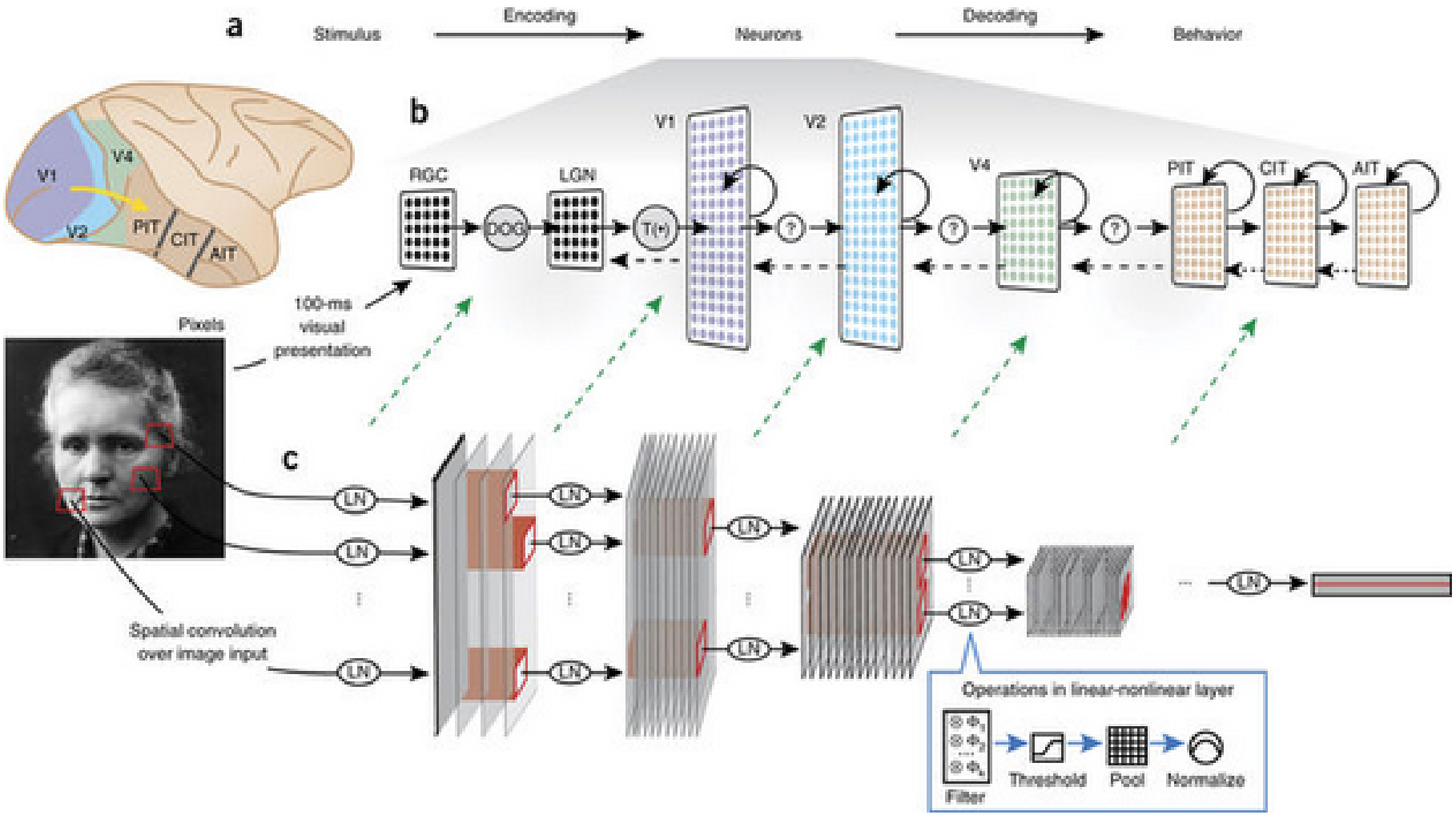


Figure 14: Neural network architecture: local connectivity, layered arrangement, and spatial invariance (size, contrast, rotation, orientation).

Seminal works

26497 LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

62020 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

37648 Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

18440 Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

21366 Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

25420 LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

45547 He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Convolutional neural networks

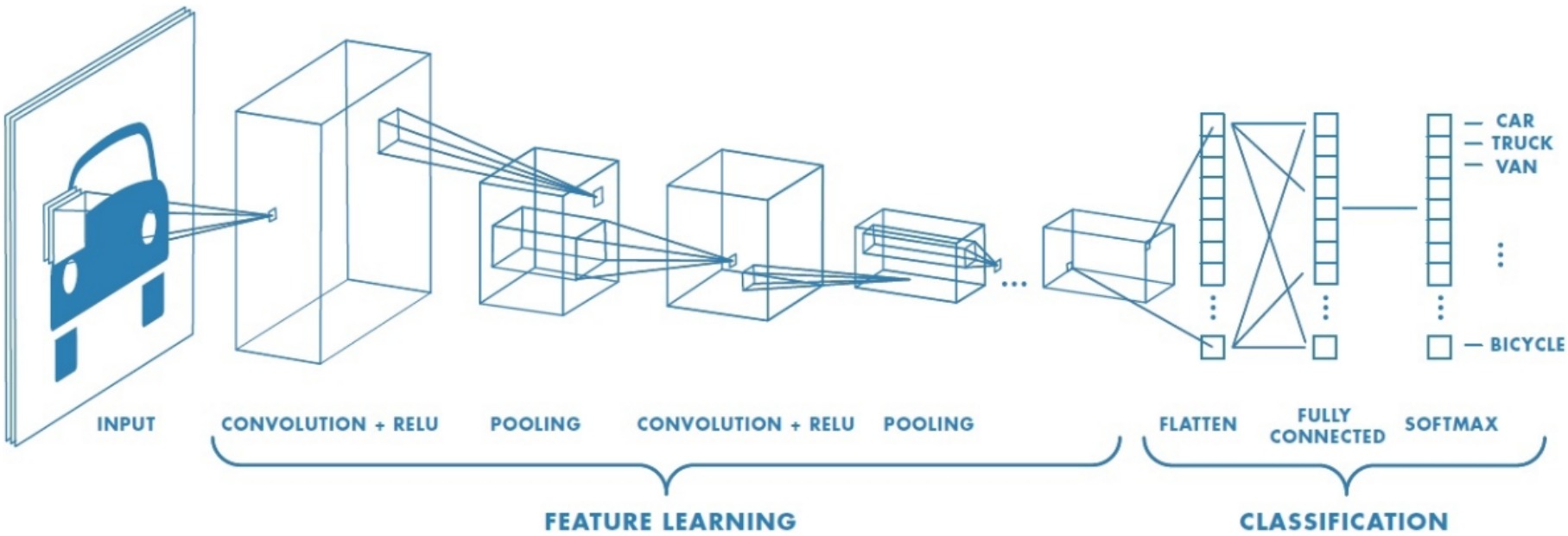


Figure 15: Layers: input (red, green and blue pixels of the image), convolution (filter) + relu (rectified linear unit), pooling (down-sampling), flatten, fully connected, and softmax. Visual image processing is split in feature learning and classification.