Computer Vision

Marc Van Droogenbroeck

Academic year: 2019-2020

Practical information

▶ Instructor: Marc Van Droogenbroeck
▶ Assistant: Philippe Latour
▶ Slides:
  http://orbi.ulg.ac.be
▶ Additional practical information:
  http://www.telecom.ulg.ac.be/assistant
▶ Evaluation
  1. large project (split in several sub-tasks): 4 students, evaluated in December. The project is compulsory!
  2. written exam (also compulsory!): 20 minutes long, after the final presentation
  3. [August only] written examination (open book)
General considerations

▶ Don’t forget the acquisition step
  - technology
  - physics of imaging (color/range data, noise, shadows, etc)
  - understanding your data is essential (also for machine learning applications!)

▶ Linear framework → non-linear frameworks

▶ A world of trade-offs (computational load ↔ framerate, etc)

▶ There is no unique, universal, solution to a computer vision problem

▶ More and more machine learning in computer vision

▶ Lessons in the spirit of a “catalog”

▶ Hands on computer vision ⇒ “practice” computer vision

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
14. Feature detection and tracking
Outline

1. Fundamentals of 2D imaging
   - Color perception and representation
   - Data representations
   - Applications

2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction

Topics

- Elements of visual perception
  - Colors: representation and colorspaces
  - Transparency

- Data structure for images

- Examples of industrial applications:
  - Segmentation
  - Optical character recognition
Human visual system, light and colors I

![Diagram of the eye](image)

**Figure:** Lateral view of the eye globe (*rods* and *cones* are receptors located on the retina).
<table>
<thead>
<tr>
<th>color</th>
<th>wavelength interval $\lambda$ [nm]</th>
<th>frequency interval $f$ [THz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>purple</td>
<td>$\sim$ 450–400</td>
<td>$\sim$ 670–750</td>
</tr>
<tr>
<td>blue</td>
<td>$\sim$ 490–450</td>
<td>$\sim$ 610–670</td>
</tr>
<tr>
<td>green</td>
<td>$\sim$ 560–490</td>
<td>$\sim$ 540–610</td>
</tr>
<tr>
<td>yellow</td>
<td>$\sim$ 590–560</td>
<td>$\sim$ 510–540</td>
</tr>
<tr>
<td>orange</td>
<td>$\sim$ 635–590</td>
<td>$\sim$ 480–510</td>
</tr>
<tr>
<td>red</td>
<td>$\sim$ 700–635</td>
<td>$\sim$ 430–480</td>
</tr>
</tbody>
</table>

Figure: Visible colors (remember that $\lambda = \frac{3 \times 10^8}{f}$).

---

Figure: Colors on the visible spectrum.
Frequency representation of colors

\[ \int_{\lambda} L(\lambda) \, d\lambda \tag{1} \]

Impossible from a practical perspective because this would require one sensor for each wavelength.

Solution: use colorspaces

Figure: Equalization experiment for colors. The aim is to mix \( A \), \( B \), and \( C \) to get as close as possible to \( X \).
Three fundamental colors: red $R (700 \text{ [nm]})$, green $G (546, 1 \text{ [nm]})$ and blue $B (435, 8 \text{ [nm]})$,

\[ r(\lambda) \quad b(\lambda) \quad v(\lambda) \]

**Figure:** Equalization curves obtained by mixing the three fundamental colors to simulate a given color.

CIE chromatic diagram for RGB
Notion of intensity

Figure: Pyramid derived from an RGB color representation.

Towards other colorspaces: the XYZ colorspace

\[
\begin{pmatrix}
X \\
Y \\
Z
\end{pmatrix} =
\begin{pmatrix}
2,769 & 1,7518 & 1,13 \\
1 & 4,5907 & 0,0601 \\
0 & 0,0565 & 5,5943
\end{pmatrix}
\begin{pmatrix}
R \\
G \\
B
\end{pmatrix}
\] (2)

\[
x = \frac{X}{X + Y + Z}
\] (3)

\[
y = \frac{Y}{X + Y + Z}
\] (4)

\[
z = \frac{Z}{X + Y + Z}
\] (5)

\[
x + y + z = 1
\] (6)
Towards other colorspaces: the XYZ colorspace II

Figure: Approximative chromatic colorspace defined by two chrominance variables $x$ and $y$.

Luminance

Luminance: $Y = 0.2126 \times R + 0.7152 \times G + 0.0722 \times B$

Figure: $xy$ chromatic diagram and maximal luminance for each color.
### The HSI colorspace

Colorspace that has a better physical meaning:

- **hue**
- **saturation**
- **intensity**

![Hue, Saturation, Intensity](image)

### Other colorspaces

- A subtractive colorspace: Cyan, Magenta, and Yellow (CMY)
- Luminance + chrominances (YIQ, YUV or YCbCr)

In practice, most of the time, we use 8 bits (≡ 1 byte) to describe a color:

<table>
<thead>
<tr>
<th>Hexadecimal</th>
<th>Decimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 00 00</td>
<td>0 0 0</td>
</tr>
<tr>
<td>00 00 FF</td>
<td>0 0 255</td>
</tr>
<tr>
<td>00 FF 00</td>
<td>0 255 0</td>
</tr>
<tr>
<td>00 FF FF</td>
<td>0 255 255</td>
</tr>
<tr>
<td>FF 00 00</td>
<td>255 0 0</td>
</tr>
<tr>
<td>FF 00 FF</td>
<td>255 0 255</td>
</tr>
<tr>
<td>FF FF 00</td>
<td>255 255 0</td>
</tr>
<tr>
<td>FF FF FF</td>
<td>255 255 255</td>
</tr>
</tbody>
</table>

**Table:** Definition of RGB color values (8 bits) and conversion table between an hexadecimal notation and decimal notation.
There are variations, such as the $Y U V$ colorspace, mainly developed for compression:

1. information concentrated in the $Y$ channel $\Rightarrow$ better compression.
2. better decorrelation between channels.

A Bayer filter mosaic is a color filter array for arranging RGB color filters on a square grid of photo sensors.

Figure: The Bayer arrangement of color filters on the pixel array of an image sensor.
Bayer filter: practical considerations

- Most mono-sensor cameras use the Bayer pattern, except for professional 3CCD cameras (three sensor planes + prism to divide the incoming light).
- The filter pattern is 50% green, 25% red and 25% blue. Why?
- We only have one value per pixel. Other values are re-built by interpolation, but they might not even exist...!
- For compression or processing,
  - 1 sensor plane ⇒ normally only one byte to process. Possible if the processing is very close to the sensor. Otherwise, there is no information about the real observed values.
  - 3 sensor planes ⇒ 3 planes to process or to compress. Expected compression rate: 3 × lower than for a single sensor.
- It might be wiser, for processing, to have a black-and-white (or a monochromatic, such as red) camera, instead of a color camera.
What defines the color of an object?

**Figure:** A synthetic 3D object. Shadows and surfaces with varying reflective coefficients model a 3D object.

Visual effects

**Figure:** Illustration of a masking visual effect.

Psycho-visual effects are taken into account. For example, for evaluating the *subjective quality* of a compression algorithm.
Sampling grid and frame organization

- Each sample located on a grid is named a *pixel* (which stands for *picture element*).
- There are two common sampling grids and they induce certain types of connectivity (defined the *number of neighbors*).

<table>
<thead>
<tr>
<th></th>
<th>Square grid</th>
<th>Hexagonal grid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4-connectivity</strong></td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>8-connectivity</strong></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>6-connectivity</strong></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
</tbody>
</table>

**Table**: Types of grid and associated connectivities.
Typical data structure for representing images: **matrices** (or 2D tables), vectors, trees, lists, piles, ...  

For **matrices**, there are several ways to organize the memory:  
- 1 channel (for example, luminance):  
  - upper left corner coordinate correspond to the (0, 0) location in the matrix.  
  - one byte per pixel  
- 3 channels, such **RGB** images. We have two options:  
  1. store 3 matrices separately (according to the 1 channel storage scheme)  
  2. store the RGB values consecutively (intermixed) for each pixel. For example, if $R_1G_1B_1$ and $R_2G_2B_2$ are the color values of two consecutive pixels, then we could store the values as $R_1G_1B_1R_2G_2B_2$ in a single matrix.
The bitplanes of an image

Table: An original image and its 8 bitplanes starting with the Most Significant Bitplane (MSB). The last plane is that with the Least Significant Bit (LSB).

Typology of images and videos

2D. This type refers to a “classic” image and is usually expressed as a 2D array of values. It might represent the luminance, a color, depth, etc.

3D. 3D images are obtained with devices that produce 3D images (that is with \(x, y, z\) coordinates). Medical imaging devices produce this type of images.

2D+\(t\). \(t\) refers to time. Therefore, \(2D + t\) denotes a video composed over successive 2D images, indexed by \(t\).

3D+\(t\). \(3D + t\) images are in fact animated 3D images. A typical example is that of animated 3D graphical objects, like that produced by simulations.
Compression I

Compression ratios: ITU-R 601 (ex CCIR 601)

- $YC_bC_r$
- $4:4:4 (720/720/720) = 270 \, [Mb/s]$
- $4:2:2 (720/360/360) = 180 \, [Mb/s]$

Two main principles:
1. remove some redundancies inside the frame (intraframe coding)
2. remove some redundancies between frames (interframe coding)

MPEG-2 levels and profiles

<table>
<thead>
<tr>
<th>Profile</th>
<th>Simple</th>
<th>Main</th>
<th>SNR</th>
<th>Spatial</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low level (235 x 288 x 30Hz)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main level (720 x 576 x 30Hz)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>High-1440 level (1440 x 1152 x 60Hz)</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>High level (1920 x 1152 x 60Hz)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

A 50 Hz video stream of size 720 x 576 (PAL) corresponds to 270 [Mb/s] uncompressed

- Lossless compression: $\approx 100 \, [Mb/s]$
- Lossy compression (but intraframe only): $\approx 27 \, [Mb/s]$
- Lossy compression (including interframe coding): MPEG-2 format $\approx 5 \, [Mb/s]$ or less
Outline

1 Fundamentals of 2D imaging
   - Color perception and representation
   - Data representations
   - Applications

2 3D vision
3 Linear filtering
4 Edge detection
5 Introduction to machine learning
6 Performance analysis
7 Segmentation
8 Object description and analysis
9 Deep learning for visual understanding
10 Template Matching & Image Registration
11 Mathematical morphology
12 Non-linear filtering
13 Motion analysis and background subtraction

Image segmentation
Character recognition

Several successive stages:
- Selection of a Region of Interest (ROI). Processing is limited to that area.
- Detection of edges (contours).
- Identification and classification of characters.

Example of classification task: classify handwritten digits

Goal: learn to recognize handwritten digits (by machine learning)

Dataset: the MNIST (Mixed National Institute of Standards and Technology) database of handwritten digits has
- a training set of 60,000 examples.
- a test/validation set of 10,000 examples.
- normalization: the digits have been size-normalized and centered in a fixed-size image.
Motivation

The physical world is in 3D (but flattened/2D by color cameras)

- Acquire 3D information
- Understand 3D geometry (in the eyes of a camera)
- Exploit 3D information for computer vision tasks
- Change the viewpoint of an observer
Technologies for dealing with 3D information

Acquisition
- Single monochromatic/color camera
- Multiple cameras (stereoscopy, network of cameras)
- Depth (range) cameras

Rendering
- Glasses
  - Color anaglyph systems
  - Polarization systems
- Display
  - Autostereoscopic display technologies

3D reconstruction

There are several ways to reconstruct the 3D geometry of a scene:

- **Direct** methods:
  - depth cameras (also named range or 3D cameras)
  - stereoscopic vision or multiview camera systems

- **Indirect** methods (more or less out of fashion due the existence of 3D cameras):
  - “depth from motion”
  - “depth from focus”
  - “depth from shadow”
  - …
There are two acquisition technologies for depth-cameras, also called range- or 3D-cameras:

1. measurements of the deformations of a pattern sent on the scene (structured light).

2. first generation of the Kinects
   - measurements by time-of-flight (ToF). Time to travel forth and back between the source led (camera) and the sensor (camera).

   - Mesa Imaging, PMD cameras
     - second generation of Kinects
Illustration of a depth map acquired with a range camera

Outline

1. Fundamentals of 2D imaging
2. 3D vision
   - A model for cameras
   - Calibration
   - 3D reconstruction
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
A camera projects a 3D world onto a 2D image. ⇒ this is the concept of *projection*.

Remarks:

- when we apply a *transformation*, some characteristics are preserved. For example, when you translate an object, its dimensions and angles are preserved.
- translations and rotations (which can be expressed as the product of a matrix on the \((X, Y, Z)^T\) real world coordinates) preserves distances, angles, etc. These leads to so-called *Euclidean transformations*.
- But what is preserved in general?
  - distances? angles? parallelism? *alignment of pixel* (≡ lines)?

The real problem if that we have some *ambiguities* when we project an 3D object to a 2D plane.

One of the simplest model (and most common): *pinhole camera* (central projection)

What is preserved?

- *alignment of pixel* (≡ lines)? Yes
Pinhole camera model

- All the light rays converge to a unique point (camera center) before arriving on the sensor plane.

Vocabulary terms:
- **camera center**: center of the central projection.
- **image plane**: plane of the sensor.
- **principal axis**: line passing through the camera center and orthogonal to the image plane.
- **principal point**: intersection between the principal axis and the image plane.
- \( f \) is called the focal length.

Mathematical model of the pinhole camera I

If \( \mathbf{X} = (X, Y, Z)^T \) is a point in space and \( \mathbf{x} = (x, y)^T \) is its projection on the image plane, then similar triangles gives

\[
\frac{x}{f} = \frac{X}{Z} \quad \text{and} \quad \frac{y}{f} = \frac{Y}{Z}
\]

(7)

Ambiguity: although \( f \) is fixed for a camera, we cannot derive the absolute values of \( X \) and \( Y \) from the measured pair of values \( (x, y) \).

⇒ a convenient representation is that of **homogeneous coordinates**

---

\(^1\)We use capital/uppercase letters to denote coordinates in the real world.
Homogeneous coordinates

Idea: new representation, by adding a trailing 1: \((x, y, 1)\)

\[(x, y) \equiv (x, y, 1) \quad (8)\]

By definition, we assume that

\[(x, y) \equiv (\lambda x, \lambda y, \lambda) \quad (9)\]

so that all pixels with varying \(\lambda\) are equivalent. In other words, in homogeneous coordinates, a point may be expressed by an infinite number of homogeneous coordinates.

Note: this is similar to how we define rational numbers, that have many different equivalent representations:

\[
\frac{1}{3} = \frac{2}{6} = \frac{13}{39} \quad (10)
\]

Homogeneous coordinates II

Properties:

- \((x, y, 0)\) is not equivalent to \((x, y)\). It is a special point.
- A point of \(\mathbb{R}^n\) is represented by a vector of size \(n + 1\). Example:
  
  \[
  \mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3 \mapsto \left(\frac{x_1}{x_3}, \frac{x_2}{x_3}\right)^T \in \mathbb{R}^2
  \]

- A line in a plane is represented by the following equation \(ax + by + c = 0\). With homogeneous coordinates, this becomes:
  
  \[
  \mathbf{l} = (a, b, c)^T
  \]

- A point belongs to a line if and only if \(\mathbf{x}^T \mathbf{l} = 0\).
- Intersection between two lines: \(\mathbf{x} = \mathbf{l} \times \mathbf{l}'\) where \(\times\) is the product between vectors.
Camera matrix I

With the help of homogeneous coordinates, the relationships of a pinhole camera model can be expressed in a convenient matrix form:

$$
\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}
$$

where $\lambda$ is equal to the depth $Z$. Remember that we don’t know the depth $Z$ from the observation in the place image (depth ambiguity).

$$
\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \mapsto \lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = PX
$$

Camera matrix II

- $P$ is a camera (projection) matrix
- For convenience, $P$ is decomposed as follows:

$$
P = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}
$$

$$
= K[I_{3 \times 3} | 0_{3 \times 1}]
$$

$K = \text{diag}(f, f, 1)$ is the calibration matrix.
(1) The central point might not be at (0, 0) in the image plane

⇒ two additional parameters in the calibration matrix

If \((x_0, y_0)^T\) are the coordinates of the principal point, the projection becomes:

\[
(X, Y, Z)^T \mapsto \left( f \frac{X}{Z} + x_0, f \frac{Y}{Z} + y_0 \right)^T
\]  (14)

The matrix form is

\[
\begin{bmatrix}
    x + Zx_0 \\
    y + Zy_0 \\
    1
\end{bmatrix} =
\begin{bmatrix}
    f & 0 & x_0 & 0 \\
    0 & f & y_0 & 0 \\
    0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
    X \\
    Y \\
    Z \\
    1
\end{bmatrix}
\]  (15)

The matrix \(K\) becomes

\[
K =
\begin{bmatrix}
    f & x_0 \\
    f & y_0 \\
    1
\end{bmatrix}
\]  (16)
(2) Non rectangular light-elements on the sensor

→ *skew parameter* \( s \)

(3) Non squared elements

This modifies the aspect ratio \( \alpha_x \) and \( \alpha_y \).

Finally,

\[
K = \begin{bmatrix}
\alpha_x & s & x_0 \\
\alpha_y & y_0 & 1
\end{bmatrix}
\] (17)

---

Intrinsic parameters

In a refined camera model,

\[
K = \begin{bmatrix}
\alpha_x & s & x_0 \\
\alpha_y & y_0 & 1
\end{bmatrix}
\] (18)

This matrix has 5 parameters. These parameters are called *intrinsic parameters*.

They characterize a camera and should be estimated for each camera separately. But once they are known, there is no need to estimated them again!
Calibration

Definition

A camera is said to be *calibrated* if

\[
K = \begin{bmatrix}
\alpha_x & s & x_0 \\
\alpha_y & y_0 & 1
\end{bmatrix}
\]

is known.

In general:

- \(\alpha_x \simeq \alpha_y\)
- \(s \simeq 0\)
- \(x_0\) and \(y_0\) close to 0 pixel (typically a few pixels maximum).

But we have a problem: *we don’t know where the center of camera is located in the real world*. So there is no way to measure \((X, Y, Z, 1)^T\).

Towards *extrinsic* parameters I

Points in the 3D world need to be expressed in a system coordinate different from that of the camera (which is not known).

- Both coordinate systems are related by a rotation and a translation:
Towards **extrinsic** parameters II

For a translation:

\[
\begin{bmatrix}
X_c \\
Y_c \\
Z_c \\
1
\end{bmatrix} = \begin{bmatrix}
X_0 - t_1 \\
Y_0 - t_2 \\
Z_0 - t_3 \\
1
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & -t_1 \\
0 & 1 & 0 & -t_2 \\
0 & 0 & 1 & -t_3 \\
0 & 0 & 0 & 1
\end{bmatrix}\begin{bmatrix}
X_0 \\
Y_0 \\
Z_0 \\
1
\end{bmatrix}
\] (19)

\[
= \begin{bmatrix}
1 & 0 & -t \\
0 & 1 & 0
\end{bmatrix}\begin{bmatrix}
X_0 \\
Y_0 \\
Z_0 \\
1
\end{bmatrix}
\] (20)

More generally (translation + rotation):

\[
\begin{bmatrix}
X_c \\
Y_c \\
Z_c \\
1
\end{bmatrix} = \begin{bmatrix}
R^T & 0 \\
0 & 1
\end{bmatrix}\begin{bmatrix}
1 & -t \\
0 & 1
\end{bmatrix}\begin{bmatrix}
X_0 \\
Y_0 \\
Z_0 \\
1
\end{bmatrix}
\] (21)

Towards **extrinsic** parameters III

In conclusion:

\[
X_c = \begin{bmatrix}
R^T & -R^T t \\
0 & 1
\end{bmatrix}X_0
\] (22)

where \(R\) is a rotation matrix and \(t\) a translation vector. This is not a projection, but an Euclidean transform between coordinate systems.

**Extrinsic parameters**

There are **6 degrees of freedom/parameters** (called **extrinsic** parameters):

- a translation vector (3 values)
- 3 rotation angles
Conclusions on $\mathbf{P}$ I

Pinhole camera model:

$$\lambda \mathbf{x} = \mathbf{K} \begin{bmatrix} \mathbf{I}_{3 \times 3} & \mathbf{0}_{3 \times 1} \end{bmatrix} \mathbf{X}_c$$  \hspace{1cm} (23)

Link between the coordinate system of the camera and an arbitrary coordinate system:

$$\mathbf{X}_c = \begin{bmatrix} \mathbf{R}^T & -\mathbf{R}^T \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \mathbf{x}$$ \hspace{1cm} (24)

By combination:

$$\lambda \mathbf{x} = \mathbf{K} \mathbf{R}^T [\mathbf{I} | - \mathbf{t}] \mathbf{x} = \mathbf{P} \mathbf{x}$$ \hspace{1cm} (25)

where $\mathbf{P} = \mathbf{K} \mathbf{R}^T [\mathbf{I} | - \mathbf{t}]$.

Conclusions on $\mathbf{P}$ II

Definition

A camera represented with a camera matrix of the form $\mathbf{P} = \mathbf{K} \mathbf{R}^T [\mathbf{I} | - \mathbf{t}]$ is called **normalized camera**.

$\mathbf{P}$ is a $3 \times 4$ matrix, which is subject to scale ambiguity. Therefore, $\mathbf{P}$ has 11 degrees of freedom (unknown parameters):

- **5 intrinsic** parameters (related to camera itself; the manufacturer can computed them and give them to the user who buys the camera).
- **6 extrinsic** parameters (related to the choice of an external coordinate system).
Calibration procedure I

How do we find the parameters of $P$?

- We need correspondences between some 3D positions on their location on the projected 2D image $X_i \leftrightarrow x_i = (x_i, y_i, z_i)$.  

- As a result

$$x_i = PX_i \Rightarrow x_i \times PX_i = 0$$

Let $P^j$ be the vector with the $j$th line of $P$,

$$
\begin{bmatrix}
0^T & -z_iX_i^T & y_iX_i^T \\
z_iX_i^T & 0^T & -x_iX_i^T \\
-y_iX_i^T & x_iX_i^T & 0^T
\end{bmatrix}
\begin{pmatrix}
P^1 \\
P^2 \\
P^3
\end{pmatrix} = 0
$$

- Two of these equations are linearly independent. But how many correspondences do we need to solve this linear systems?

\textit{At least 6}
In practice, we have an over-determined system of linear equations $A\mathbf{p} = \mathbf{0}$ if there is no noise. But, due to noise, we use an optimization procedure, such as the minimum square optimization.

Use of a precise and well-known 3D calibration pattern.
Reconstruction involves several aspects:

- **3D reconstruction** of the scene
  - **dense** reconstruction (we have to find the depth for every pixel of the scene)
  - **non-dense** reconstruction

- **detection of features**. Usually, these are special points of objects (corners, centers, etc)

- the scene content might be *moving*

Feature detection and motion are often related to each other!

---

**Systems composed of multiple cameras**

Three elements interact:

- **calibration**
- possible **motion** of the cameras
- **3D structure** of the scene

**Theorem**

Assume a series of images taken with an uncalibrated moving camera for which we have establish correspondences between pairs of points, then it is only possible to reconstruct the 3D scene up to a projective transform.

In the following, we only consider fixed cameras.
Steps to reconstruct a scene from multiple cameras (stereoscopic system)

Generally, we have to find correspondences between points of the **views of two color cameras**.

There are several possibilities:

1. **cameras are aligned** (mechanically or numerically, after a transformation) $\rightarrow$ computation of the **disparity map**.
2. **cameras are not aligned** $\rightarrow$ corresponding points in the two camera planes are related to each other via the **fundamental matrix**. The fundamental matrix imposes a constraint; it is not sufficient to reconstruct a scene (calibration is needed, etc). This is studied under the name of **epipolar geometry**.
3. **all the points are in the same plane** $\rightarrow$ this constraint (knowledge) facilitates the correspondence finding between two views. The transformation is named an **homography**.

---

**Stereoscopic cameras (aligned cameras)**

*If cameras are aligned, the disparity is the inverse of depth.*
If two cameras only differ in the horizontal direction, the horizontal difference between two locations provides an information about the depth.

**Attention:**
- cameras and alignments are not perfect!
The computation is disparity is difficult:
- close to borders (diffraction effects)
- in texture-less zones

Discussion

Advantages (+) or drawbacks (-) of a reconstruction (≡ depth computation) by the technique of disparity maps:
- (+) the process is simple. One only has to estimate the distance between two corresponding points along a line.
- (-) cameras must be aligned,
  - either mechanically,
  - or by software. This usually involves re-sampling the image, leading to precision losses.

We need an alternative when the cameras are not aligned.
Unaligned cameras: epipolar geometry and fundamental matrix I

![Diagram](image)

C and C’ are the center of the two cameras.

Unaligned cameras: epipolar geometry and fundamental matrix II

![Diagram](image)

The use of a second camera is supposed to help determining the location of a point $X$ along the projection line of the other camera.
Major challenges I

Objective: find a way to use the correspondences between points in a pair of views to be able to find \((X, Y, Z)\) for every point in the scene.

▶ Calibration:
  - each camera has its own calibration matrix:
    - solution: find the intrinsic and extrinsic parameters
  - when there are two cameras
    - solution: there exists a way to build the relationship between the projections in the camera planes that is expressed by the fundamental matrix or the essential matrix.

▶ Camera placement:
  - fixed: the manufacturer can calibrate the camera.
  - changing or unknown: one can isolate the intrinsic parameters, but this requires a calibration procedure.

▶ Object structures: they might lead to ambiguities (holes, shadows, etc).

Major challenges II

“realistic” solution: do some prior assumptions about the objects. It is impossible to solve all the ambiguities, even with an infinite number of cameras.
Towards the fundamental matrix

Consider a point $X$ in space. It is viewed by the two cameras as:

$$\lambda_1 x_1 = P_1 X = [A_1 \mid b_1] X \quad (26)$$
$$\lambda_2 x_2 = P_2 X = [A_2 \mid b_2] X \quad (27)$$

We find $X, Y, Z$ with the help of the equations of the first camera:

$$\lambda_1 x_1 = P_1 X = [A_1 \mid b_1] X = A_1 \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + b_1 \quad (28)$$

therefore,

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = A_1^{-1}(\lambda_1 x_1 - b_1) \quad (29)$$

By substitution, in the second equation, we get

$$\lambda_2 x_2 = A_2 A_1^{-1}(\lambda_1 x_1 - b_1) + b_2 = \lambda_1 A_{12} x_1 + (-A_{12} b_1 - b_2) \quad (30)$$

Consequently, $x_1$ and $x_2$ are linearly independent.

Fundamental matrix

Epipolar constraint

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}^T F \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0 \quad (31)$$

- The fundamental matrix $F$ is a $3 \times 3$ matrix that constraints the values of the coordinates in the image planes of two cameras.
- $x'Fx = 0$ for all the corresponding pairs $x \leftrightarrow x'$.

Characteristics:

- 7 degrees of freedom (rank 2 matrix)
- the determinant is null
- it is defined up to a scaling factor
We can derive the fundamental matrix, starting from the two camera models. Assume that:

1. we are able to isolate the intrinsic and extrinsic parameters of both cameras.
2. the absolute 3D coordinate systems is placed on the internal coordinate of one camera. Then, we have

\[ P = K [I|0] \quad \text{and} \quad P' = K' [R|t] \]  \hspace{1cm} (32)

Then, it can be shown that the fundamental matrix is given by

\[ F = K'^{-T} [t] \times RK^{-1} = K'^{-T} R[R^T t] \times K^{-1} \]  \hspace{1cm} (33)

where

\[ [a]_\times = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \]  \hspace{1cm} (34)

If the intrinsic parameters are known, the coordinates can be normalized to remove the effect of the intrinsic parameters (according to transformation such as \( \hat{u} = K^{-1} u \) and \( \hat{u}' = K'^{-1} u' \)), leading to

\[ P = [I|0] \quad \text{and} \quad P' = [R|t] \]  \hspace{1cm} (35)
The fundamental matrix simplifies, which results in the *essential matrix*.

\[ E = [t] \times R = R[R^T t] \times \]  

The major advantage of this matrix is that only the extrinsic parameters play a role (the essential matrix has 5 degrees of freedom).

---

**Computation of the fundamental matrix: similar ideas to that of the computation of the calibration matrix**

- We have to find a number of pairwise correspondences \( x_i \leftrightarrow x'_i \) (minimum of 4)
- We build a system of linear equations by means of \( x'_i F x_i = 0 \)
- The resolution of the system leads to \( F \)

There exist several algorithms to proceed to the determination of \( F \).

Remember that, if the fundamental matrix links the points between two planes, it does not suffice to reconstruct a scene.
Reconstruction of a 3D: steps I

1. Find correspondences between pairs of points in the two views
2. Calculate the fundamental matrix
3. Find the camera parameters, with the help of the fundamental matrix:

\[ P = [I|O] \quad P' = [[e'] \times F|e'] \]

where

\[
[a]_× = \begin{bmatrix}
0 & -a_3 & a_2 \\
a_3 & 0 & -a_1 \\
-a_2 & a_1 & 0
\end{bmatrix}
\]

and

\[ F^T e' = 0 \]

Reconstruction of a 3D: steps II

4. For each pixel \( x_i \leftrightarrow x'_i \), triangulate \( X_i \):

\[ PX_i \times x_i = 0 \quad \text{et} \quad P'X_i \times x'_i = 0 \]

We have 4 equations for 3 unknowns.

Attention! Remember that this reconstruction is valid up to a projective transform.
Outline

1 Fundamentals of 2D imaging
2 3D vision
3 Linear filtering
4 Edge detection
5 Introduction to machine learning
6 Performance analysis
7 Segmentation
8 Object description and analysis
9 Deep learning for visual understanding
10 Template Matching & Image Registration
11 Mathematical morphology
12 Non-linear filtering
13 Motion analysis and background subtraction
14 Feature detection and tracking

What is filtering?

<table>
<thead>
<tr>
<th>$f(x)$</th>
<th>25</th>
<th>27</th>
<th>30</th>
<th>24</th>
<th>17</th>
<th>15</th>
<th>22</th>
<th>23</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>$\min{f(x + 1), f(x), f(x - 1)}$</td>
<td>25</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\max{f(x + 1), f(x), f(x - 1)}$</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>24</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“better” version of $f(x)$?

But, why do we filter an image?
Linear filtering

Fourier transform and filtering operation

The notion of “ideal” filter

Categories of ideal filters

Typical filters:
- Low-pass filters
- High-pass filters
- Gabor filters
Let us take $f(x, y)$ as the values of a single channel image (for example, the grayscale component):

- $(x, y)$ are the coordinates
- $f(x, y)$ is the single-valued image of the pixel located at $(x, y)$

**Definition ((Continuous) Fourier transform)**

The (continuous) Fourier transform $\mathcal{F}(u, v)$ of $f(x, y)$ is defined by

$$
\mathcal{F}(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)e^{-2\pi j(ux+vy)}dx\,dy
$$

**Example**

The Fourier transform is a complex image (it has a real component and an imaginary component). Often, we are interested in the amplitude.

*Figure: (a) original image and (b) centered Fourier transform of its amplitude.*
Fourier transform: definition and interpretation III

What is a spatial frequency?

Simple waveforms and profiles

Complex waveform

Fourier transform: definition and interpretation IV

The Fourier transform is a representation of $f(x, y)$ in terms of spatial frequencies:

- How the Image information is stored in the Fourier Transform
- How the Image information is stored in the Fourier Transform

Spatial Frequency

Orientation
**Fourier transform: definition and interpretation V**

#### Definition (Continuous inverse Fourier transform)

The inverse (continuous) Fourier transform is defined by

$$f(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathcal{F}(u, v)e^{2\pi j(ux + vy)} du dv$$  \hspace{1cm} (38)

There is a one-to-one relationship between a $f(x, y)$ and $\mathcal{F}(u, v)$:

$$f(x, y) \iff \mathcal{F}(u, v)$$  \hspace{1cm} (39)

#### Definition (Filtering)

Applying a filter $\mathcal{H}(u, v)$ corresponds to modify the Fourier coefficient of some frequencies

$$\mathcal{F}(u, v) \rightarrow \mathcal{F}(u, v)\mathcal{H}(u, v)$$  \hspace{1cm} (40)

---

**Notion of “ideal” filter**

A filter is said to be “ideal” if every transform coefficient is multiplied by 0 or 1.

#### Definition (Ideal filter)

An *ideal* filter is such that its transfer function is given by

$$\forall(u, v), \mathcal{H}(u, v) = 0 \text{ or } 1.$$  \hspace{1cm} (41)

The notion of ideal filter is closely related to that of *idempotence*. The idempotence for a filter is to be understood such that, for an image $f(x, y)$,

$$\mathcal{F}(u, v)\mathcal{H}(u, v) = \mathcal{F}(u, v)\mathcal{H}(u, v)\mathcal{H}(u, v)$$  \hspace{1cm} (42)
Typology of ideal filters I

For one-dimensional signals (such as the image function along an image line):

![Diagram showing one-dimensional filters]

Figure: One-dimensional filters.

Typology of ideal filters II

There are three types of circular ideal filters:

- **Low-pass filters:**

  \[
  H(u, v) = \begin{cases} 
  1 & \sqrt{u^2 + v^2} \leq R_0 \\
  0 & \sqrt{u^2 + v^2} > R_0 
\end{cases}
  \]  
  (43)

![Images of original and low-pass filtered images]

(a) Original image  (b) Low-pass filtered image
Typology of ideal filters III

- **High-pass filters:**

\[
\mathcal{H}(u, v) = \begin{cases} 
1 & \sqrt{u^2 + v^2} \geq R_0 \\
0 & \sqrt{u^2 + v^2} < R_0 
\end{cases} 
\]  

(44)

- **Band-pass filters.** There are equivalent to the complementary of a low-pass filter and a high-pass filter:

\[
\mathcal{H}(u, v) = \begin{cases} 
1 & R_0 \leq \sqrt{u^2 + v^2} \leq R_1 \\
0 & \text{otherwise}
\end{cases} 
\]  

(45)

Typology of ideal filters IV

**Figure:** Transfer function of pass-band filters.
Effects of filtering

A typical low-pass filter is the Butterworth filter (of order $n$) defined as

$$
\mathcal{H}(u, v) = \frac{1}{1 + \left(\frac{\sqrt{u^2 + v^2}}{R_0}\right)^{2n}}.
$$

(46)

**Figure:** Transfer function of a low-pass Butterworth filter (with $n = 1$).
(a) Input image  
(b) Spectrum of the filtered image

**Figure:** Effects of an order 1 Butterworth filter (cut-off frequency: $f_c = 30$).

**Effect of a low-pass filter (with decreasing cut-off frequencies)**
High-pass filters I

\[ H(u, v) = \frac{1}{1 + \left(\frac{R_0}{\sqrt{u^2 + v^2}}\right)^{2n}} \]  

(a) Filtered image \hspace{1cm} (b) Spectrum of (a)

**Figure:** Effects of an order 1 Butterworth filter (cut-off frequency: \( f_c = 50 \)).

Effect of a high-pass filter (with increasing cut-off frequencies)
Linear filtering

Gabor filters I

Definition

Gabor filters belong to a particular class of linear filters. There are directed filters with a Gaussian-shaped impulse function:

\[ h(x, y) = g(x', y')e^{2\pi j(Ux+Vy)} \]  

\[ (x', y') = (x \cos \phi + y \sin \phi, -x \sin \phi + y \cos \phi) \], these are the \((x, y)\) coordinates rotated by an angle \(\phi\), and

\[ g(x', y') = \frac{1}{2\pi\sigma^2}e^{-\left((x'/\lambda)^2+y'^2\right)/2\sigma^2}. \]

The corresponding Fourier transform is given by

\[ \mathcal{H}(u, v) = e^{-2\pi^2\sigma^2[(u'-U')^2\lambda^2+(v'-V')^2]} \]  

\[ (u', v') = (u \cos \phi + v \sin \phi, -u \sin \phi + v \cos \phi) \], and

\[ (U', V') \] is obtained by rotating \((U, V)\) with the same angle \(\phi\).

Gabor filters II

Figure: Transfer function of Gabor filter. The white circle represents the \(-3 \, [\text{dB}]\) circle (= half the maximal amplitude).
Implementation

There are mainly 4 techniques to implement a Gaussian filter:

1. **Convolution** with a restricted Gaussian kernel. One often chooses $N_0 = 3\sigma$ or $5\sigma$

   \[
   g_{1D}[n] = \begin{cases} 
   \frac{1}{\sqrt{2\pi}\sigma} e^{-(n^2/2\sigma^2)} & |n| \leq N_0 \\
   0 & |n| > N_0 
   \end{cases}
   \]  
   \hspace{2cm} (50)

2. **Iterative convolution** with a uniform kernel:

   \[
   g_{1D}[n] \simeq u[n] \otimes u[n] \otimes u[n]
   \]  
   \hspace{2cm} (51)

   where

   \[
   u[n] = \begin{cases} 
   \frac{1}{(2N_0+1)} & |n| \leq N_0 \\
   0 & |n| > N_0 
   \end{cases}
   \]  
   \hspace{2cm} (52)

3. Multiplication in the Fourier domain.

4. Implementation as a recursive filter.
Outline

1 Fundamentals of 2D imaging
2 3D vision
3 Linear filtering
4 Edge detection
   - Linear operators
   - Non-linear operators
   - Hough’s transform
5 Introduction to machine learning
6 Performance analysis
7 Segmentation
8 Object description and analysis
9 Deep learning for visual understanding
10 Template Matching & Image Registration
11 Mathematical morphology
12 Non-linear filtering
13 Motion analysis and background subtraction

Border/contour/edge detection

- Linear operators
  - First derivative operators
  - Second derivative operators
  - Sampling the derivative
    - Residual error
    - Synthesis of operators for a fixed error
  - Practical expressions of gradient operators and convolution masks
- Non-linear operators
  - Morphological gradients
- Hough’s transform
What's a border/contour/edge?

Figure: An image (diagonal gradient) and its contours (in black).

Edges in depth images

Kinect 1

Kinect 2

CamCube

Range image

Ground truth

Sobel

PED-2 (Lejeune)
Can we locate edge points?

![Diagram](image)

**Figure**: Problem encountered in locating an edge point.

**Outline**

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. **Edge detection**
   - Linear operators
   - Non-linear operators
   - Hough’s transform
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
For derivate operators, we have to address two problems:

1. find the best **approximate** for the derivate.
2. **avoid** an excessive amplification of the **noise**.

These are two apparent contradictory requirements ⇒ trade-offs
Fourier transform (quick reminder) I

(A) (B) (C) (D) (E) (F)

(G) (H) (I) (J) (K) (L)

Fourier transform (quick reminder) II

(A) (B) (C) (D) (E) (F)

(G) (H) (I) (J) (K) (L)
First derivate operator I

Let us consider the partial derivate of a function \( f(x, y) \) with respect to \( x \). Its Fourier transform \( \mathcal{F}(u, v) \) becomes

\[
\frac{\partial}{\partial x} f(x, y) \Leftrightarrow 2\pi j u \mathcal{F}(u, v)
\]  

(53)

In other words, deriving with respect to \( x \) consists of multiplying the Fourier transform of \( f(x, y) \) by the following transfer function \( \mathcal{H}_x(u, v) = 2\pi j u \), or of filtering \( f(x, y) \) with the following impulse function:

\[
h_x(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (2\pi j u) e^{2\pi j (xu + yv)} du dv
\]  

(54)

If we adopt a vectorial notation of the derivate, we define the gradient \( \nabla f \) of image \( f \) by

\[
\nabla f = \frac{\partial f}{\partial x} \overrightarrow{e_x} + \frac{\partial f}{\partial y} \overrightarrow{e_y} = (h_x \otimes f) \overrightarrow{e_x} + (h_y \otimes f) \overrightarrow{e_y}
\]  

(55)
Definition (Gradient amplitude)

\[ |\nabla f| = \sqrt{(h_x \otimes f)^2 + (h_y \otimes f)^2} \]  
(56)

The amplitude of the gradient is sometimes approximated by

\[ |\nabla f| \simeq |h_x \otimes f| + |h_y \otimes f| \]  
(57)

which introduces a still acceptable error (in most cases) of 41%!

But we still have to find a way to calculate \( h_x \otimes f \) and \( h_y \otimes f \) ...

Definition (Gradient orientation)

\[ \varphi_{\nabla f} = \tan^{-1} \left( \frac{h_y \otimes f}{h_x \otimes f} \right) \]  
(58)

Definition (Laplacian \( \nabla^2 f \))

The Laplacian is the scalar defined as

\[ \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = (h_{xx} \otimes f) + (h_{yy} \otimes f) \]  
(59)

As the first derivative, it can be shown that in the Fourier domain, the Laplacian consists to apply the following filter

\[ \nabla^2 f \quad \Rightarrow \quad \left( (2\pi ju)^2 + (2\pi jv)^2 \right) \mathcal{F}(u, v) \]  
(60)

\[ \Rightarrow \quad -4\pi^2 (u^2 + v^2) \mathcal{F}(u, v) \]  
(61)

As can be seen, high frequencies tend to be amplified. Is this suitable?
In order to derive practical expressions for the computation of a derivate, we adopt the following approach:

1. develop some approximations and derive the error due to the approximation,
2. study the spectral behavior of these approximations, and
3. discuss some practical approximations expressed in the terms of convolution masks.

**Centered approximations?**
An approximation of the first derivate is given by

\[ f'_a(x) = \frac{f(x + h) - f(x - h)}{2h} = \frac{(+1)f(x + h) + (-1)f(x - h)}{2h} \]  \hspace{1cm} (62)

where \( h \) is the distance between two samples and index \( a \) denotes that it is an approximation. Please note that this approximation consists to filter \( f(x) \) by the following "multiplicative" mask

\[ \frac{1}{2h} \begin{bmatrix} -1 & 0 & +1 \end{bmatrix} \]  \hspace{1cm} (63)

or the convolution mask

\[ \frac{1}{2h} \begin{bmatrix} +1 & 0 & -1 \end{bmatrix} \]  \hspace{1cm} (64)
Sampling the gradient and residual error III

For the second derivate, one possible approximation is

\[ f''_a(x) = \frac{f(x + h) - 2f(x) + f(x - h)}{h^2} \]  

(65)

Sampling the gradient and residual error IV

We make use of the following representation of a function.

**Theorem (Taylor series)**

The Taylor series of a real valued function \( f(x) \) that is infinitely differentiable at a real number \( a \) is the following power series

\[ f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \ldots \]  

(66)
**Sampling the gradient and residual error V**

**Computation of the residual error.** Let’s consider the following Taylor series

\[
f(x + h) = f(x) + h f'(x) + \frac{h^2}{2} f''(x) + \ldots + \frac{h^n}{n!} f^{(n)}(x) + \ldots \tag{67}
\]

\[
f(x - h) = f(x) - h f'(x) + \frac{h^2}{2} f''(x) + \ldots + (-1)^n \frac{h^n}{n!} f^{(n)}(x) + \ldots
\]

**First derivate.** By subtraction, member by member, these two equalities, one obtains

\[
f(x + h) - f(x - h) = 2 h f'(x) + \frac{2}{3!} h^3 f^{(3)}(x) + \ldots = 2 h f'(x) + O(h^3) \tag{68}
\]

After re-ordering,

\[
f'(x) = \frac{f(x + h) - f(x - h)}{2h} + O(h^2) \tag{69}
\]

**Sampling the gradient and residual error VI**

**Second derivate.**

\[
f(x + h) = f(x) + h f'(x) + \frac{h^2}{2} f''(x) + \ldots + \frac{h^n}{n!} f^{(n)}(x) + \ldots \tag{70}
\]

\[
f(x - h) = f(x) - h f'(x) + \frac{h^2}{2} f''(x) + \ldots + (-1)^n \frac{h^n}{n!} f^{(n)}(x) + \ldots \tag{71}
\]

Like for the first derivate, we use the Taylor extension by add them this time,

\[
f(x + h) + f(x - h) = 2 f(x) + h^2 f''(x) + \frac{2}{4!} h^4 f^{(4)}(x) + \ldots \tag{72}
\]

As a result:

\[
f''(x) = \frac{f(x + h) - 2f(x) + f(x - h)}{h^2} + O(h^2) \tag{73}
\]

The \( f''(x) \) approximation is also of the second order in \( h \).
Synthesis of expressions with a pre-defined error.
Another approximation, of order $O(h^4)$, can be built. It corresponds to

$$f'_a(x) = \frac{-f(x + 2h) + 8f(x + h) - 8f(x - h) + f(x - 2h)}{12h} \quad (74)$$

Spectral behavior of discrete gradient operators I

Consider the one-dimensional continuous function $f(x)$ and the following first derivate:

$$f'_a(x) = \frac{f(x + h) - f(x - h)}{2h} \quad (75)$$

Its Fourier is given by

$$\frac{f(x + h) - f(x - h)}{2h} \Rightarrow \frac{e^{2\pi juh} - e^{-2\pi juh}}{2h} \mathcal{F}(u) \quad (76)$$

which can be rewritten as, given that $\sin(a) = \frac{e^{ja} - e^{-ja}}{2j}$,

$$\frac{f(x + h) - f(x - h)}{2h} \Rightarrow (2\pi ju) \frac{\sin(2\pi hu)}{2\pi hu} \mathcal{F}(u) \quad (77)$$

where the $(2\pi ju)$ factor corresponds to the ideal (continuous) expression of the first derivate.
Let us now consider the approximation of the second derivate

\[ f''_a(x) = \frac{f(x + h) - 2f(x) + f(x - h)}{h^2} \quad (78) \]

Its Fourier transform is given by

\[ \frac{f(x + h) - 2f(x) + f(x - h)}{h^2} \Leftrightarrow (-4\pi^2u^2) \left( \frac{\sin(\pi hu)}{\pi hu} \right)^2 \mathcal{F}(u) \quad (79) \]

**Figure:** Spectral behavior of the derivate approximations (for \( h = 1 \).)

<table>
<thead>
<tr>
<th>ideal</th>
<th>first derivate</th>
<th>second derivate</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2\pi j u )</td>
<td>( (2\pi j u) \frac{\sin(2\pi hu)}{2\pi hu} )</td>
<td>( (-4\pi^2u^2) \left( \frac{\sin(\pi hu)}{\pi hu} \right)^2 )</td>
</tr>
</tbody>
</table>
Practical expressions of gradient operators and convolution/multiplication masks I

Practical expression are based on the notion of convolution masks

\[
\begin{bmatrix}
  +1 & -1 \\
\end{bmatrix}
\]  \hspace{1cm} (80)

corresponds to the following non-centered approximation of the first derivate:

\[
\frac{(-1) \times f(x, y) + (+1) \times f(x + h, y)}{h}
\]  \hspace{1cm} (81)

This “convolution mask” has an important drawback. Because it is not centered, the result is shifted by half a pixel. One usually prefers to use a centered (larger) convolution mask such as

\[
\begin{bmatrix}
  +1 & 0 & -1 \\
\end{bmatrix}
\]  \hspace{1cm} (82)

In the \( y \) (vertical) direction, this becomes

\[
\begin{bmatrix}
  +1 \\
  0 \\
  -1 \\
\end{bmatrix}
\]  \hspace{1cm} (83)

But then, it is also possible to use a diagonal derivate:

\[
\begin{bmatrix}
  +1 & 0 & 0 \\
  0 & 0 & 0 \\
  0 & 0 & -1 \\
\end{bmatrix}
\]  \hspace{1cm} (84)
The use of (centered) convolution masks still has some drawbacks:

- **Border effects.**
  Solutions:
  - (i) put a default value outside the image;
  - (ii) mirroring extension: copy inside values starting from the border;
  - (iii) periodization of the image –pixels locate on the left are copied on the right of the image,
  - (iv) copy border values to fill an artificial added border.

- The range (dynamic) of the possible values is modified.

- It might be needed to apply a normalization factor.
### Prewitt gradient filters

\[
[h_x] = \frac{1}{3} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \tag{85}
\]

\[
[h_y] = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \tag{86}
\]

**Figure:** Original image, and images filtered with a horizontal and vertical Prewitt filter respectively.

### Sobel gradient filters

\[
[h_x] = \frac{1}{4} \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \tag{87}
\]

**Figure:** Original image, and images filtered with a horizontal and vertical Sobel filter respectively.
Second derivate: basic filter expressions

\[
\begin{bmatrix}
1 & -2 & 1 \\
-2 & 1 \\
1 & -2 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 1 & 0 \\
1 & -4 & 1 \\
0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 1 \\
1 & -8 & 1 \\
1 & 1 & 1
\end{bmatrix}
\]
Non-linear operators I

Morphological gradients

- *Erosion gradient* operator:

\[
GE(f) = f - (f \ominus B)
\] (88)

- *Dilation gradient* operator:

\[
GD(f) = (f \oplus B) - f
\] (89)

- Morphological gradient of Beucher: \( GE(f) + GD(f) \).

- *Top-hat* operator: \( f - f \odot B \);

- min/max gradient operators:
  \( \min(GE(f), GD(f)) \), \( \max(GE(f), GD(f)) \)

- Non-linear Laplacian: \( GD(f) - GE(f) \).

Gradient of Beucher

(a) Original image \( f \)  
(b) \( f \oplus B \)
（c）\( f \ominus B \)  
（d）\( (f \oplus B) - (f \ominus B) \) (reverse video)
Different non-linear border detectors

(a) \((f \oplus B) - (f \ominus B)\)

(b) \(f - f \circ B\) (top-hat)

(c) \(\max(\text{GE}(f), \text{GD}(f))\)

(d) \(\text{GD}(f) - \text{GE}(f)\)
Detecting lines

Challenge: detect lines in an image

Towards the Hough transform

- Difficulty: matching a set of points arranged as a line
- Idea: instead of considering the family of points \((x, y)\) that belong to a line \(y = ax + b\), consider the two parameters
  1. the slope parameter \(a\) (but \(a\) is unbounded for vertical lines)
  2. the intercept parameter \(b\) (that is for \(x = 0\))
With the Hough transform, we consider the \((r, \theta)\) pair where

- the parameter \(r\) represents the distance between the line and the origin,
- while \(\theta\) is the angle of the vector from the origin to this closest point.

We have several ways to characterize a line:

1. Slope \(a\) and \(b\), such that \(y = ax + b\).
2. The two parameters \((r, \theta)\), with \(\theta \in [0, 2\pi]\) and \(r \geq 0\).

Link between these characterizations:
The equation of the line becomes

\[
y = \left(-\frac{\cos \theta}{\sin \theta}\right)x + \left(\frac{r}{\sin \theta}\right)
\]

(90)

Check:

- For \(x = 0\), \(r = y \sin \theta \rightarrow\) ok.
- For \(x = r \cos \theta\), \(y = r \sin \theta \rightarrow\) ok.
By re-arranging terms of $y = \left( -\frac{\cos \theta}{\sin \theta} \right)x + \left( \frac{r}{\sin \theta} \right)$, we get that, for an arbitrary point on the image plane with coordinates, e.g., $(x_0, y_0)$, the family of lines passing through it are given by

$$r = x_0 \cos \theta + y_0 \sin \theta$$

(91)

Example

For three points $(x_0, y_0)$, we explore the Hough space. That is, we compute $r$ for a given set of orientations $\theta$:

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>30</td>
<td>69.6</td>
</tr>
<tr>
<td>60</td>
<td>81.2</td>
</tr>
<tr>
<td>90</td>
<td>70</td>
</tr>
<tr>
<td>120</td>
<td>40.6</td>
</tr>
<tr>
<td>150</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>57.1</td>
</tr>
<tr>
<td>30</td>
<td>79.5</td>
</tr>
<tr>
<td>60</td>
<td>80.5</td>
</tr>
<tr>
<td>90</td>
<td>60</td>
</tr>
<tr>
<td>120</td>
<td>23.4</td>
</tr>
<tr>
<td>150</td>
<td>19.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>74.6</td>
</tr>
<tr>
<td>30</td>
<td>89.6</td>
</tr>
<tr>
<td>60</td>
<td>80.6</td>
</tr>
<tr>
<td>90</td>
<td>50</td>
</tr>
<tr>
<td>120</td>
<td>6.0</td>
</tr>
<tr>
<td>150</td>
<td>39.6</td>
</tr>
</tbody>
</table>
Thus, the problem of detecting colinear points can be converted to the problem of finding concurrent curves.
Algorithm for detecting lines

**Algorithm**

1. Detect edges in the original image.
2. **Select only some pixels corresponding to “strong edges”** for which there is enough evidence that they belong to lines.
3. **For each selected pixel, accumulate values in the corresponding bins of the Hough space:**
   - For example, we take each $\theta$ in $[0^\circ, 360^\circ]$ with a step of $1^\circ$.
   - We calculate the corresponding $r$ by $r = x_0 \cos \theta + y_0 \sin \theta$.
   - If $\text{Hough}(r, \theta)$ is one bin of the Hough space, then we do
     \[
     \text{Hough}(r, \theta) \leftarrow \text{Hough}(r, \theta) + 1
     \] (92)
4. **Threshold the accumulator function** $\text{Hough}(r, \theta)$ to select bins that correspond to lines in the original image.
5. **Draw the corresponding lines** in the original image.

**“Toy” example**
Real example

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
   - Introduction to machine learning (ML)
   - Understanding classification
   - Conclusion
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
A starter: silhouette recognition I

Can you decide which silhouettes are those of humans?
Try to write an algorithm to solve this problem!

A starter: silhouette recognition II

After binarization of the silhouettes

Can you decide which silhouettes are those of humans?
Try to write an algorithm to solve this problem!
Observation

- Most of the tasks related to video scene interpretation are complex.
- A human expert can easily take the right decision, but usually without being able to explain how he does it.

One possible solution

Use machine learning techniques that have proven to be a powerful tool in computer science and vision

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
   - Introduction to machine learning (ML)
   - Understanding classification
   - Conclusion
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
Machine learning (ML) techniques\textsuperscript{2} aim is to
\begin{itemize}
  \item build a decision rule automatically.
  \item speed up the decisions.
  \item be able to generalize to unseen objects. Really?!
\end{itemize}

Computational cost:
\begin{itemize}
  \item The model is learned only once.
  \item The model is used many times.
  \item \textbf{[Q]} Which operation should be the fastest?
\end{itemize}

\textsuperscript{2}We consider only “supervised” machine learning techniques.

Examples of techniques are:
\begin{itemize}
  \item Naive Bayes classifier
  \item Nearest neighbor
  \item Artificial neuronal networks (\rightarrow deep learning)
  \item \textit{Support Vector Machines} (SVM) [CV95]
  \item Random forests (\textit{ExtRaTrees} [GEW06])
\end{itemize}

A good reference book on this topic is [HTF09].
Families of machine learning methods

For supervised learning, we have labeled (annotated) training data.

Classification vs regression

Example of classification:

- yes
- yes
- no
- yes
- no
- no

Example of regression:

- 65.2°
- −2.0°
- −71.5°
- 15.4°
- −47.4°
- −5.5°
Introduction to machine learning

Applications I

ML techniques have proven to be successful for many purposes:

- detecting people in images [DT05];
- recognizing people [BHP05];
- analyzing people’s behavior [SFC+11];
- detecting faces with software embedded in cameras [VJ04];

Applications II

- semantic segmentation
- etc
There exists many machine learning libraries. For example,

- scikit-learn (Python)
- libSVM (Matlab, Java, Python, etc)
- Regression trees (C/Matlab)
- Java-ML (Java)
- Shark (C++)
- ...

More specifically for deep learning,

- Caffe2, TensorFlow, Theano, Torch, Keras, CNTK (Python)
**How does it work? What is learned?**

Example of learning database:

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample 1</td>
<td>7.99</td>
<td>6.77</td>
<td>9.75</td>
<td>1.58</td>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>sample 2</td>
<td>2.24</td>
<td>9.51</td>
<td>1.14</td>
<td>8.00</td>
<td>7.66</td>
<td>0</td>
</tr>
<tr>
<td>sample 3</td>
<td>2.18</td>
<td>2.83</td>
<td>2.96</td>
<td>5.14</td>
<td>9.73</td>
<td>0</td>
</tr>
<tr>
<td>sample 4</td>
<td>8.44</td>
<td>7.39</td>
<td>4.57</td>
<td>4.94</td>
<td>2.70</td>
<td>1</td>
</tr>
<tr>
<td>sample 5</td>
<td>9.55</td>
<td>5.92</td>
<td>2.52</td>
<td>0.46</td>
<td>1.53</td>
<td>1</td>
</tr>
<tr>
<td>sample 6</td>
<td>3.32</td>
<td>9.13</td>
<td>0.50</td>
<td>5.07</td>
<td>8.22</td>
<td>2</td>
</tr>
</tbody>
</table>

MODEL $\equiv y(x_1, x_2, x_3, x_4, x_5) = ?$

- $y$ is the **output variable** (the **class**)
- Samples are described by **attributes** (or **features**) $x_1, x_2, \ldots$
- The same number of attributes should be used for all samples.
- The meaning of an attribute should not depend on the sample.

**Example of classification task**

*handwritten character recognition*

- size = 100 samples
- choice: attributes = raw pixels
- the size of the images is $32 \times 32$
  - dimension = 1024 attributes

The intrinsic difficulty of machine learning

The theoretical rule to minimize the error rate is

$$ y(\overrightarrow{x}) = \arg \max_{y_i \in \{0,1,...\}} (p(y = y_i | \overrightarrow{x})) \quad (93) $$

Let $\rho$ be the probability density function (pdf) of all objects in the attributes space, and $\rho_i$ be the pdf of the objects belonging to class $y_i$. Using Bayes’ rule (that is $p(A|B)p(B) = p(B|A)p(A)$):

$$ p(y = y_i | \overrightarrow{x}) = \frac{\rho_i(\overrightarrow{x}) p(y = y_i)}{\rho(\overrightarrow{x})} \quad (94) $$

Therefore,

$$ y(\overrightarrow{x}) = \arg \max_{y_i \in \{0,1,...\}} (\rho_i(\overrightarrow{x}) p(y = y_i)) \quad (95) $$

The intrinsic difficulty is that it is very difficult to estimate $\rho_i$ from the learning database because the space is not densely sampled.

An example of decision rule in 1D

Let us assume that we have to recognize men and women based on a single attribute: the height. In Australia, there are 100 women for 100 men. But in Russia, there are 114 women for 100 men.

$$ y = ( \text{height} < 169.34 )? \text{"woman" : "man"} ; $$
Example of classifier: the nearest neighbors (NN) I

Let us consider a problem in 2 dimensions (2 attributes $x_1$, $x_2$):

$$
\begin{align*}
&x_2 \\
&\downarrow \\
&x_1
\end{align*}
$$

$\text{k-NN}$

$\text{1-NN}$

Advantages:

- The size of the neighborhood is automatically chosen depending on $k$.
- The model is the learning set (or a pruned version of it).

Drawbacks

1. The time needed to take a decision is $O(n)$, where $n$ is the learning set size.
2. Which distance measure should we select? There exists an infinity of possible choices! [DD09]
Choosing the complexity of the model

Which model is the best?

Two questions:

▶ Does the model explain the learning set (LS)?
  → *resubstitution error* = error estimated on the learning set

▶ Is the model able to predict the classes for unknown samples?
  → *generalization error* = error estimated on the test set (TS)
Choosing the complexity of the model

![Graph showing the relationship between model complexity and error](image source: P. Geurts, 'An introduction to Machine Learning')

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
   - Introduction to machine learning (ML)
   - Understanding classification
   - Conclusion
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
Conclusion

ML = automatic + generalization + preprocessing

Machine learning techniques are
- powerful methods.
- a complement to traditional methods.
- essential in computer science.
- adequate for real time computation.
- “easy” to use (not to engineer, nor to optimize\(^3\)).

\(^3\)This is why researchers are still working on machine learning methods.

Bibliography I

Bibliography II


Bibliography III

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. **Performance analysis**
   - General discussion
   - Evaluation criteria and metrics
   - Illustrations
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
What do we want to evaluate?

Segmentation  Edge detection  Background subtraction

It is challenging...

General principles of evaluation I

The process of evaluation usually involves:

- A methodology that comprises:
  - Experiments.
    Experimental conditions have to be defined clearly.
  - Criteria or scores.
  - Reference data:
    - unlabeled data, which is collected “in the wild”.
    - labeled or annotated data, sometimes named as ground truth data.
Typology of evaluation methods:

1. **Subjective.** The evaluation process involves a representative group of human viewers.
   - Takes the user’s experience into account.
   - Is time-consuming (→ expensive), provides subjective evaluation scores, depends on the experimental setup.

2. **Objective.** Based on objective measurements.

---

**Objective quality measures and distortion measures I**

Let $f$ be the input image (whose size is $N \times N$) and $\hat{f}$ be the image after some processing.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not always available</td>
<td>Available $f$</td>
<td>Calculated $\hat{f}$</td>
</tr>
</tbody>
</table>
Objective quality measures and distortion measures II

**Definition (Mean Square Error)**

\[
MSE = \frac{1}{NM} \sum_{j=0}^{N-1} \sum_{k=0}^{M-1} (f(j, k) - \hat{f}(j, k))^2
\]

(96)

**Definitions (Signal to noise ratios)**

[**Signal to Noise Ratio**]

\[
SNR = \frac{\sum_{j=0}^{N-1} \sum_{k=0}^{M-1} (f(j, k))^2}{\sum_{j=0}^{N-1} \sum_{k=0}^{M-1} (f(j, k) - \hat{f}(j, k))^2}
\]

(97)

[**Peak Signal to Noise Ratio**]

\[
PSNR = \frac{NM \times 255^2}{\sum_{j=0}^{N-1} \sum_{k=0}^{M-1} (f(j, k) - \hat{f}(j, k))^2}
\]

(98)

---

**On the importance of ground truth data**

Depending of the availability of ground truth, we have:

- **Standalone evaluation.**
  When a reference is not available.

- **Relative evaluation.**
  When ground truth data is available for comparison.

**Properties of “good” ground truth data**

- representative for the application.
- different conditions for the acquisition (large variety of samples).
- correctly annotated (→ there is some agreement on the quality of the reference).
On the importance of ground truth data

Depending of the availability of ground truth, we have:

- **Standalone evaluation.**
  When a reference is not available.

- **Relative evaluation.**
  When ground truth data is available for comparison.

### Properties of “good” ground truth data

- representative for the application.
- different conditions for the acquisition (large variety of samples).
- correctly annotated (→ there is some agreement on the quality of the reference).

### Difficulties specific to ground truth data

#### Annotation process

- the “annotation” process (which consists to build the ground truth data) is complex and sometimes debatable.
  Two “flavors”:
  - synthetic data (which comes with synthetic annotations)
  - real data with manually annotated data.

- involves some averaging.
  - the averaging process is often wrong...
  - it smooths the results and complicates the interpretation.

- the annotation process is partly subjective.
**ChangeDetection.NET (CDNET) dataset:** a dataset for testing background subtraction algorithms

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Camera Jitter</th>
<th>Dynamic Back.</th>
<th>Intern. Motion</th>
<th>Shadow</th>
<th>Thermal</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Baseline" /></td>
<td><img src="image2.png" alt="Camera Jitter" /></td>
<td><img src="image3.png" alt="Dynamic Back." /></td>
<td><img src="image4.png" alt="Intern. Motion" /></td>
<td><img src="image5.png" alt="Shadow" /></td>
<td><img src="image6.png" alt="Thermal" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Bad Weather" /></td>
<td><img src="image8.png" alt="Low Frame Rate" /></td>
<td><img src="image9.png" alt="Night Video" /></td>
<td><img src="image10.png" alt="PTZ" /></td>
<td><img src="image11.png" alt="Turbulence" /></td>
<td></td>
</tr>
</tbody>
</table>

**Main characteristics:**

- 11 categories: about 5 videos per category
- Long video sequences
- Thousands of annotated images (ground truth data)
  - 5 labels: static (≡ background), shadow, non-ROI, unknown, and moving (≡ foreground)
- 7 performance metrics are computed
- Ranking per video, category, and globally
We can distinguish 3 main evaluation tasks:

1. **Optimization/analysis** of one technique. We have different parameter sets.
   - *What are the best parameters?*

2. **Comparison** between techniques (a same technique with different parameter sets or different techniques).
   - *Is this technique better than this other one?*

3. **Ranking**.
   - *Which one is the best?*

Ideally, evaluation tools should be different for these tasks. In practice, there is a confusion.
A framework for developing evaluation criteria

Terminology from the *classification theory*.

### Definition (Classification)

In machine learning and statistics, *classification* is the process of identifying the *category/class of a new observation*, on the basis of a training set of data containing observations (or instances) whose category membership is known.

The categories are named “*classes*”.

### Example (In background subtraction, there are two classes:)

1. Foreground
2. Background

But there could be more: shadows, sky, road, etc.

---

The notion of positive and negative

Let us consider a two-class problem and denote the “*negative*” and “*positive*” classes by \( c^- \) and \( c^+ \), respectively. In background subtraction, \( \text{background} \equiv \text{negative} \) and \( \text{foreground} \equiv \text{positive} \).

*Figure:* Background subtraction identifies pixels belonging to the *foreground* (positive class \( c^+ \), white pixels) and pixels belonging to the *background* (negative class \( c^- \), black pixels).
A classifier estimates the class of each new sample (pixel in the case of background subtraction).

There are two notions of classes:

- [true class or ground truth] it is denoted by $y \in \{c^-, c^+\}$.
- [estimated class] the class estimated by the classifier is denoted by $\hat{y} \in \{c^-, c^+\}$

Other notions:

- $\pi^- = p(y = c^-)$ and $\pi^+ = p(y = c^+)$ are the priors.
- the rates of negative and positive predictions, denoted by $\hat{\pi}^- = p(\hat{y} = c^-)$ and $\hat{\pi}^+ = p(\hat{y} = c^+)$ respectively (not to be confused with the priors!).

### Confusion matrix

**Definition (Confusion matrix)**

There are four possibilities, described by the confusion matrix:

<table>
<thead>
<tr>
<th></th>
<th>$c^+$</th>
<th>$c^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c^+$</td>
<td>True Positive (TP)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td>$c^-$</td>
<td>False Positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

**Remarks:**

- There are four quantities: TP, FN, FP, TN.
- It is easy to extend a confusion matrix to more than two classes.
- How do we build metrics/performance indicators?
Evaluation metrics

Definition (Bias)

A classifier is said *unbiased* when \( \hat{\pi}^- = \pi^- \iff \hat{\pi}^+ = \pi^+ \).

Definitions (Conditional probabilities)

They involve the true class:

- **True Positive Rate**, also named *sensitivity* or *recall*, is defined as
  \[
  TPR = p(\hat{y} = c^+ | y = c^+) \tag{99}
  \]

- **True Negative Rate** (also named *specificity*) is
  \[
  TNR = p(\hat{y} = c^- | y = c^-) \tag{100}
  \]

These quantities are linked to the

- **False Negative Rate** \( \text{FNR} = (1 - TPR) \)
- **False Positive Rate** \( \text{FPR} = (1 - TNR) \).

Receiver Operating Characteristic (ROC) evaluation space

Definition (Receiver Operating Characteristic (ROC))

The \((\text{FPR}, \text{TPR})\) pair defines the ROC evaluation space.
Performance analysis
Evaluation criteria and metrics

Precision/recall evaluation space

Definition (Precision)

\[
P = p\left(y = c^+ \mid \hat{y} = c^+\right) = \frac{TP}{TP + FP}
\]  

(101)

Definitions (Recall \[\equiv \text{TPR}\])

\[
R = p\left(\hat{y} = c^+ \mid y = c^+\right) = \frac{TP}{TP + FN}
\]  

(102)

Unachievable zone in the PR space, but only the upper part of the ROC curve makes sense.

In the ROC space, the diagonal represents the random classifiers.

Figure: Isoperformance lines for the F metric \(F = \frac{2P \times R}{P + R}\).
Evaluation of segmentation quality I

Problem statement
Evaluation of segmentation quality II

Practice:

- Does not nicely fit into the framework of classification
- Some difficulties:
  - *individual* object segmentation quality vs *overall* segmentation quality evaluation
  - subjectivity for defining the ground truth
- Two major classes of metrics for **standalone** segmentation quality evaluation
  - intra-object homogeneity
  - inter-object disparity
- Metrics for **relative** segmentation quality evaluation
  - spatial accuracy
  - temporal accuracy

Evaluation of edge detection I

Problem statement
Performance analysis

Evaluation of edge detection II

Practice:
- Some evaluation methods rely on the classification theory.
- Use of the Precision/Recall evaluation space.
- The F score is a trade-off (harmonic mean between P and R):
  \[ F = \frac{2 \frac{P \times R}{P + R}}{\frac{P}{P} + \frac{R}{R}} = \frac{2P \times R}{P + R} \]

Figure: Maximum Precision and Recall curve obtained by varying the parameter set. Circles indicate where the F measure is maximal.

Problem statement

<table>
<thead>
<tr>
<th>Marc Van Droogenbroeck</th>
<th>Computer Vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance analysis</td>
<td>Illustrations</td>
</tr>
</tbody>
</table>

Evaluation of background subtraction I
Practice:

- Many dataset with ground truth available.
- Background subtraction is often considered as a classification task.
- Many algorithms available \( \Rightarrow \) most people try to rank algorithms.
- Many metrics. No agreement over the best one (if it exists). Therefore, an algorithm is evaluated with respect to multiple metrics.
- It is still hard to evaluate the dynamic (\( \equiv \) over time) behavior. Usually, there is one score for the whole video sequence.
- The role of priors is underestimated or just ignored.

Problems with the ChangeDetection dataset I

Reminder: main characteristics

- 11 categories: about 5 videos per category
- long video sequences
- thousands of annotated images (ground truth data)
  - 5 labels: static (\( \equiv \) background), shadow, non-ROI, unknown, and moving (\( \equiv \) foreground)
- 7 performance metrics are computed
- ranking per video, category, and globally

Are there problems?
Problems with the ChangeDetection dataset II

▶ About the content:
  - Some categories are irrelevant for most targeted applications.
  - Some hand-made annotations are debatable. For example, how do you handle ghosts or static objects? How do you define the shadow?

▶ About the methodology:
  - All the data is available ⇒ it is possible to fine tune all the parameters.
  - Parameters of methods are supposed to be the same for all videos. How do we treat methods that adaptively tune parameters?
  - Problem with the ranking methodology.
  - The averaging process for the metrics (per category and globally) is incorrect.
  - No source code available for some methods ⇒ some results are impossible to check.
  - Some metrics are redundant
Image segmentation I

- Problem statement
- Segmentation by thresholding
- Segmentation by region detection (region growing)
  - Watershed
- Segmentation by classification (semantic classification)

Image segmentation II

General considerations:
- a very specific problem statement is not always easy.
- chicken-and-egg problem; segmentation is an ill-conditioned problem.
Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
   - Problem statement
   - Segmentation by thresholding
   - Segmentation by region growing: illustration with the watershed
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction

Marc Van Droogenbroeck
Computer Vision
Academic year: 2019-2020

Problem statement I

Definition (Segmentation)

Generally, the problem of segmentation consists in finding a set of non-overlapping regions $R_1, \ldots, R_n$ such that

$$\mathcal{E} = \bigcup_{i=1}^{n} R_n \quad \text{and} \quad \forall i \neq j, \ R_i \cap R_j = \emptyset$$  \hspace{1cm} (103)

Definition (Alternative definition)

More formally, the segmentation process is an operator $\phi$ on an image $I$ that outputs, for example, a binary image $\phi(I)$ that differentiates regions by selecting their borders.

An alternative consists to attribute a different label to each pixel of different regions (this is called scene labeling).
Segmentation is a spatial process

As any similar operator, segmentation can be *local* or *global*:

- for *local* segmentation techniques, the results for one given pixel does not impact the segmentation result outside a close neighborhood.
- for *global* techniques, changing one pixel value can impact the whole result.
A first typology of segmentation techniques and comparisons

<table>
<thead>
<tr>
<th>Family of segmentation techniques</th>
<th>input</th>
<th>local/global</th>
<th>markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thresholding</td>
<td>image</td>
<td>local (pixel)</td>
<td>no</td>
</tr>
<tr>
<td>Watershed</td>
<td>image, gradient, etc</td>
<td>global</td>
<td>yes</td>
</tr>
</tbody>
</table>

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
   - Problem statement
   - Segmentation by thresholding
   - Segmentation by region growing: illustration with the watershed
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
Rationale
There are two classes of pixels:

1. background pixels
2. foreground pixels

[Note that background and foreground do not refer to motion in this case!]. Sometimes, there is a “don’t know” class.

Assumptions to solve the segmentation problem:

1. the probability density functions of the two content types are different.
2. one threshold or two thresholds (Otsu’s method) are sufficient.
Segmentation by thresholding III

Segmentation by thresholding

![Graphs showing optimal threshold and distributions](image)

Figure: Optimal threshold.

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
   - Problem statement
   - Segmentation by thresholding
   - Segmentation by region growing: illustration with the watershed
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
Images can be seen as topographic surfaces

**Figure**: An image (left-hand side) and a view of its corresponding topographic surface (right-hand side).

**Segmentation by watershed**

In the terms of a topographic surface, a catchment basin $C(M)$ is associated to every minimum $M$.

**Figure**: Minimums, catchment basins, and watershed.
The general principles of the watershed

Steps:
1. identify valleys (minimums)
2. proceed to flooding
3. construct dams between neighboring catchment basins

Formal expression of a segmentation algorithm based on the watershed

Approach: proceed level by level

- An horizontal “slice” is binary image. Therefore, we first study the case of binary images.
- definition of geodesic path and distance.
- description of an algorithm that handles a stack of thresholded images.
Let $X$ be a binary image.

**Definition (Geodesic path)**

A **geodesic path**, of length $l$, between two points $s$ and $t$ is a series of $l + 1$ pixels $x_0 = s, x_1, \ldots, x_l = t$ such that

$$\forall i \in [0, l], x_i \in X \text{ and } \forall i \in [0, l], x_{i-1}, x_i \text{ are neighbors}$$

(104)

**Geodesic distance**

![Figure: The shortest path between $x$ and $y$.](image)

**Definition (Geodesic distance)**

The **geodesic distance** between two points $s$ and $t$ is the length of the shortest geodesic path linking $s$ to $t$; the distance is infinite if such a path does not exist.
Algorithm for the construction of the geodesic skeleton by growing the zone of influence

**Notations:**
The zone of influence of a set $Z_i$, is denoted by $ZI(domain=X, center=Z_i)$ and its frontier by $FR(domain=X, center=Z_i)$. The skeleton by zone of influence (SZI) is obtained via the following algorithm:

- first, one delineates the $Z_i$ zones of each region;
- for remaining pixels, an iterative process is performed until stability is reached: if a pixel has a neighbor with an index $i$, then this pixel gets the same index; pixels with none or two different indices in their neighborhood are left unchanged;
- after all the iterations, all the pixels (except pixels at the interface) are allocated to one region of the starting regions $Z_i$.

**Example**

![Geodesic skeleton diagram](image)

*Figure:* Geodesic skeleton.
The case of grayscale images (a gradient image for example) I

Figure: A dam is elevated between two neighboring catchment basins.

Notations:

- $f$ is the image.
- $h_{min}$ and $h_{max}$ are the limits of the range values of $f$ on the function support (typically, $h_{min} = 0$ and $h_{max} = 255$).
- $T_h(f) = \{x \in \text{dom } f : f(x) \leq h\}$ is a set obtained by thresholding $f$ with $h$. For $h$ growing, we have a stack of decreasing sets.
- $M_i$ are the minimums and $C(M_i)$ are the catchment basins.
Step by step construction

Let $C_h(M_i)$ be the subset of the $M_i$ basin filled at “time” (or “height”) $h$. Then

$$C_h(M_i) = C(M_i) \cap T_h(f)$$ (105)

In this expression, $C(M_i)$ is unknown.

**Initialization:**

- $C_{h\min}(M) = T_{h\min}(f)$; the initialization considers that all the local minimums are valid catchment basin originators.

**Construction**

$$\forall h \in [h_{\min} + 1, h_{\max}]: C_h(M) = ZI_h \cup Min_h$$ (106)

with

- $ZI_h$ = influence zone (with domain $T_h(f)$);
- $Min_h$ is the set of all the points of $T_h(f)$ that have no label after the growing process of influence zones. They correspond to minimums that are introduced at level $h$.

Illustration: detection of pores in gypsum

![Microtomography image](image1)
![Labeled image](image2)
Marking is a process that allows to select only some of the local minimums.

Watershed has the following advantages with respect to other techniques (such as thresholding):

- the possibility to be applicable to any sort of input image (original image, gradient, etc).
- the flexibility to put some markers to select only a few local minimums. With markers, the amount of regions is exactly equal to the number of markers put in the image.

Illustration: segmentation of cells
Semantic segmentation (based on deep learning)

- Based on classification techniques and machine learning
- Pixel-based
- A series of semantic notions (persons, cars, bicycles, etc)

Panoptic segmentation (based on deep learning)

- Panoptic segmentation ≡ *semantic* segmentation + *instance* segmentation
Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
   - General considerations
   - Texture analysis
   - Shape description
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction

Marc Van Droogenbroeck

Computer Vision

Academic year: 2019-2020

245 / 600
General considerations

Objects have:
- a **texture** (inside)
- a **shape** (border)

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. **Object description and analysis**
   - General considerations
   - Texture analysis
   - Shape description
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
Texture analysis

Definition?

Statistical characterization of textures
- Local mean
- Local standard deviation
- Local histogram
- Co-occurrence matrix of a grayscale image

Geometrical characterization of textures
- Spectral approach
- Texture and energy

Goals of texture analysis?

The major question related to texture are:
- texture analysis. The purpose is to characterize a texture by a set of parameters called “texture descriptors”.
- texture recognition
- image segmentation
Definition

Definition (Tentative definition)

A texture is a signal that can be extended naturally outside of its domain.

Figure: One possible texture (according to Lantuéjoul).

Examples of “real” textures
Simple analysis of a grayscale image

**Figure:** Example of an image with two textures.

---

Statistical descriptors of textures

Simple descriptors:

- mean
- variance

But, there is a problem

**Figure:** Textures with identical means and variances.
Figure: Illustration of texture statistics computed over a circle. (a) grayscale mean (103 and 156 respectively) (b) standard deviation (32 and 66 respectively).

**Local statistics**

**Definition (Local mean)**

The local mean over a spatial window $W$ is defined as

$$
\mu_f = \frac{1}{\#(W)} \sum_{(x,y) \in W} f(x, y) \quad (107)
$$

where $\#(W)$ is the cardinality of $W$.

**Definition (Local standard deviation)**

The standard deviation over a spatial window $W$ is defined as

$$
\sigma_f = \sqrt{\frac{\sum_{(x,y) \in W} [f(x, y) - \mu_f]^2}{\#(W)}} \quad (108)
$$
Object description and analysis

Texture analysis

Global and local histograms I

Definition (Histogram)

The *histogram* of an image is the curve that displays the frequency of each grayscale level.

```
| 0 0 0 0 0 0 |
| 0 2 1 2 2 2 |
| 2 1 1 1 2 2 |
| 2 1 1 1 2 2 |
| 3 2 1 0 0 0 |
| 3 3 3 3 2 0 |
```

**Figure:** Non normalized histogram of an image.

Global and local histograms II

**Figure:** An image and its global histogram (here \(W\) accounts for the whole image domain).

| Mean: 136.6 | Intensity: 0.255 |
| Std Dev: 66.7 | Count: 127092.0 |
| Median: 161.0 | Percentile: 100.00 |
| Pixels: 127092.0 |
Definition (Local histogram)

With a smaller window $W$, it is possible to define a local (normalized) histogram $p(l)$ as

$$p(l) = \frac{\# \{(x, y) \in W \mid f(x, y) = l\}}{\#(W)}$$

Histogram statistics

- **Mean**
  $$\mu_L = \sum_{l=0}^{L-1} l \cdot p(l)$$

  where $L$ denotes the number of possible grayscale levels inside the $W$ window.

- **Standard deviation**
  $$\sigma_L = \sqrt{\sum_{l=0}^{L-1} (l - \mu_L)^2 \cdot p(l)}$$

- **Obliquity**
  $$S_s = \frac{1}{\sigma_L^3} \sum_{l=0}^{L-1} (l - \mu_L)^3 \cdot p(l)$$

- **“Kurtosis”**
  $$S_k = \frac{1}{\sigma_L^4} \sum_{l=0}^{L-1} (l - \mu_L)^4 \cdot p(l) - 3$$
Definition (Co-occurrence matrix)

A co-occurrence matrix is defined by means of a geometrical relationship $R$ between two pixel locations $(x_1, y_1)$ and $(x_2, y_2)$. An example of such a geometrical relationship is

$$x_2 = x_1 + 1 \quad (114)$$
$$y_2 = y_1 \quad (115)$$

for which $(x_2, y_2)$ is at the right of $(x_1, y_1)$.

The co-occurrence matrix $C_R(i, j)$ is squared, with the $L \times L$ dimensions, where $L$ is the range of all possible grayscale values inside of $B$. Indices of the co-occurrence matrix then indicates the amount of grayscale level value pairs as defined by $R$.

Construction of the $C_R(i, j)$ matrix:

1. Matrix initialization: $\forall i, j \in [0, L] : C_R(i, j) = 0$.
2. Filling the matrix. If the relationship/condition $R$ between two pixels $(x_1, y_1)$ and $(x_2, y_2)$ is followed/met, then

$$C_R(f(x_1, y_1), f(x_2, y_2)) \leftarrow C_R(f(x_1, y_1), f(x_2, y_2)) + 1$$
Example I

Let us consider an image with four grayscale levels ($L = 4$, and $l = 0, 1, 2, 3$):

$$f(x, y) = \begin{bmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 \\
0 & 2 & 2 & 3 \\
2 & 2 & 3 & 3 \\
\end{bmatrix} \quad (116)$$

$$P_{0^\circ, d}(i, j) = \#\{(x_1, y_1), (x_2, y_2) \in B \mid y_1 = y_2, |x_2 - x_1| = d, f(x_1, y_1) = i \text{ and } f(x_2, y_2) = j\} \quad (117)$$

The $P_{0^\circ, 1}$ and $P_{90^\circ, 1}$ matrices are $4 \times 4$ matrices respectively given by

$$P_{0^\circ, 1} = \begin{bmatrix}
6 & 2 & 1 & 0 \\
2 & 2 & 0 & 0 \\
1 & 0 & 4 & 2 \\
0 & 0 & 2 & 2 \\
\end{bmatrix} \quad P_{90^\circ, 1} = \begin{bmatrix}
6 & 1 & 2 & 0 \\
1 & 2 & 1 & 1 \\
2 & 1 & 2 & 1 \\
0 & 1 & 1 & 2 \\
\end{bmatrix} \quad (118)$$

Example II

About the use of co-occurrence matrices:

[+] Rich information about the texture

[-] Explosion of the number of features in the case of small textures
Geometrical characterization of textures

Use of the Fourier transform

Figure: Spectral characterization of a texture. Right-hand images are the modules of the Fourier transforms (inverse video).

Textures and energy I

Measures are derived from three simple vectors: (1) \( L_3 = (1, 2, 1) \) that computes the mean, (2) \( E_3 = (-1, 0, 1) \) that detects edges, and (3) \( S_3 = (-1, 2, -1) \) which corresponds to the second derivate. By convolving these symmetric vectors, Laws has derived 9 basic convolution masks:

\[
\begin{bmatrix}
\frac{1}{36} & 1 & 2 & 1 \\
2 & 4 & 2 \\
1 & 2 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & 0 & -1 \\
2 & 0 & -2 \\
1 & 0 & -1 \\
\end{bmatrix}
\begin{bmatrix}
-1 & 2 & -1 \\
-2 & 4 & -2 \\
-1 & 2 & -1 \\
\end{bmatrix}
\]

Laws 1 Laws 2 Laws 3

\[
\begin{bmatrix}
-1 & -2 & -1 \\
0 & 0 & 0 \\
1 & 2 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & 0 & -1 \\
0 & 0 & 0 \\
-1 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
-1 & 2 & -1 \\
0 & 0 & 0 \\
1 & -2 & 1 \\
\end{bmatrix}
\]

Laws 4 Laws 5 Laws 6

\[
\begin{bmatrix}
-1 & -2 & -1 \\
2 & 4 & 2 \\
-1 & -2 & -1 \\
\end{bmatrix}
\begin{bmatrix}
-1 & 0 & 1 \\
2 & 0 & -2 \\
-1 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & -2 & 1 \\
-2 & 4 & -2 \\
1 & -2 & 1 \\
\end{bmatrix}
\]

Laws 7 Laws 8 Laws 9
Textures and energy II

Figure: Laws “residues” (reverse video).

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
   - General considerations
   - Texture analysis
   - Shape description
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
Shape description. There are many of them:
- Lineic shape description
- Quadtree as a shape descriptor
- Morphological skeleton
- Other surfacic shape descriptors

Measures
- Basic geometrical measures
- Shape factors
- Moments
- Morphological measures

There are two main families for describing a shape:
- **lineic** shape descriptors. They follow the border and encode its characteristics.
- **surfacic** shape descriptors. They represent the surface surrounded by the border.

Properties
- Translation, rotation, and scale (affine) invariance
- Insensitive to noise
- Resistance to partial occlusions
There are **two main families** for describing a shape:
- **lineic** shape descriptors. They follow the border and encode its characteristics.
- **surfacic** shape descriptors. They represent the surface surrounded by the border.

**Properties**
- translation, rotation, and scale (affine) invariance
- insensitive to noise
- resistance to partial occlusions

**How to encode a series of pixels?**

**Chain code**

![Diagram](image)

**Figure:** Definitions of directions in (a) 4-connectivity and (b) 8-connectivity.
How to encode a series of pixels? II

Figure: (a) A contour, (b) contour sampled on a digital grid, (c) 4-connected chain code, and (d) 8-connected chain code.

How to encode a series of pixels? III

Figure: Contour with a crossing border pixel.
Polygonal approximation

Figure: Polygonal approximation of a shape.

Fourier descriptors I

Some shape descriptors see the object as:

- a binary function
  \[
  f(x, y) = \begin{cases} 
  1 & \text{if } (x, y) \text{ belongs to the object} \\
  0 & \text{otherwise} 
  \end{cases} 
  \]  
  \hspace{2cm} (119)

- a series of points in the complex 2D space
  \[
  s[n] = x[n] + jy[n] 
  \]  
  \hspace{2cm} (120)

\(s[n]\) is a one-dimensional discrete collection of complex numbers.
Object description and analysis

Shape description

Fourier descriptors II

Figure: Representation of a shape as a series of points in the complex 2D space.

Fourier descriptors III

Definition (Fourier descriptor)

The discrete Fourier transform of \( s[n] \) is defined as

\[
S[u] = \frac{1}{N} \sum_{n=0}^{N-1} s[n] e^{-2\pi j un / N}
\]  

(121)

The \( S[u] \) coefficients are the Fourier descriptors of the shape.

Approximations.

Figure: Original shape and approximations with a selection of Fourier descriptors.
Approximation of a human silhouette

**Figure:** Approximations of a human silhouette by an increasing number of Fourier descriptors.

The quadtree as a shape descriptor

**Figure:** Quadtree decomposition and quadtree representation.
Introduction to mathematical morphology: elements of set theory I

Sets will be denoted with capital letters, such as \( A, B, \ldots, \) and elements of these sets by lowercase letters \( a, b, \ldots \)

- **Inclusion**
  \( X \) is a *subset* of \( Y \) (that is, \( X \) is included in \( Y \)) if all the elements of \( X \) also belong to \( Y \): \( X \subseteq Y \iff \forall x \in X \Rightarrow x \in Y \).

- **Intersection**
  The *intersection* between \( X \) and \( Y \) is the set composed of the elements that belong to both sets: \( X \cap Y = \{ x \text{ such that } x \in X \text{ and } x \in Y \} \).

- **Union**
  The *union* between two sets is the set that gathers all the elements that belong to at least one set: \( X \cup Y = \{ x \text{ such that } x \in X \text{ or } x \in Y \} \).

- **Translated set**
  The *translate* of \( X \) by \( b \) is given by \( \{ z \in \mathcal{E} | z = x + b, x \in X \} \).

\[ \begin{align*}
\text{Original} & & \text{Translated} \\
X & & X_b \\
\end{align*} \]
Erosion

Definition (Morphological erosion)

\[ X \ominus B = \{ z \in \mathcal{E} | B_z \subseteq X \}. \quad (122) \]

Erosion with a disk

Figure: Erosion of \( X \) with a disk \( B \). The origin of the structuring element is drawn at the center of the disk (with a black dot).
**Dilation I**

**Definition (Dilation)**

From an algebraic perspective, the *dilation* (*dilatation* in French!), is the union of translated version of $X$:

$$X \oplus B = \bigcup_{b \in B} X_b = \{ x + b | x \in X, b \in B \}. \quad (123)$$

---

**Dilation II**

*Figure:* Dilation of $X$ with a disk $B$. 

---

Marc Van Droogenbroeck

Computer Vision

Object description and analysis | Shape description

Academic year: 2019-2020 284 / 600
Algebraic properties

- erosion and dilation are *increasing* operators: if \( X \subseteq Y \), then \((X \ominus B) \subseteq (Y \ominus B)\) and \((X \oplus B) \subseteq (Y \oplus B)\);
- if the structuring element *contains* the origin, then the erosion is *anti-extensive* and the dilation is *extensive*, that is \(X \ominus B \subseteq X\) and \(X \subseteq X \oplus B\).

Morphological opening I

**Definition (Opening)**

The *opening* results from cascading an erosion and a dilation with the same structuring element:

\[
X \circ B = (X \ominus B) \oplus B.
\]
The interpretation of the opening operator (which can be seen as an alternative definition) is based on

\[ X \circ B = \bigcup \{ B_z | z \in \mathcal{E} \text{ and } B_z \subseteq X \} \]  

(125)

In other words, the opening of a set by structuring element \( B \) is the set of all the elements of \( X \) that are covered by a translated copy of \( B \) when it moves inside of \( X \).

---

**Morphological closing**

**Definition (Closing)**

A *closing* is obtained by cascading a dilation and an erosion with a unique structuring element:

\[ X \bullet B = (X \oplus B) \ominus B. \]  

(126)
First attempts to define the skeleton

- Consider a continuous set $X$ and its frontier $\partial X$; an element $x$ of the object $X$ belongs to the skeleton of $X$, denoted by $S(X)$, if there exists a disk centered on $x$, included in $X$, that touches $\partial X$ at least twice (maximal balls with two contact points).

- Locus of the center of all the maximal balls $B$ contained in $X$.

These definitions are not fully equivalent, but they have the same topological closure in $\mathbb{R}^2$. 

The morphological skeleton to describe a shape I
The morphological skeleton to describe a shape II

Figure: Shapes and their skeleton $S(X)$.

Skeletons are sensitive to noise on the object.

Properties of a skeleton transform

1. The skeleton transform is nor increasing, neither non-increasing. Indeed, $X \subseteq Y$ does not imply that $S(X) \subseteq S(Y)$, nor that $S(Y) \subseteq S(X)$.
2. The skeleton transform is anti-extensive: $S(X) \subseteq X$.
3. The skeleton transform is idempotent: $S(S(X)) = S(X)$.

Figure: What is the skeleton of this object $X$? Ideally, it should be located between the two rows. In practice, it is on the upper or lower row.
Properties of a skeleton transform

1. The skeleton transform is **nor increasing, neither non-increasing**. Indeed, $X \subseteq Y$ does not imply that $S(X) \subseteq S(Y)$, nor that $S(Y) \subseteq S(X)$.

2. The skeleton transform is **anti-extensive**: $S(X) \subseteq X$.

3. The skeleton transform is **idempotent**: $S(S(X)) = S(X)$.

Morphological size I

**Definition (Continuous object size)**

Let $B$ be a **continuous set** whose size is arbitrary set to 1. Then, we can define a scale space of continuous sets by

$$rB = \{rb | b \in B\} \quad r \geq 0 \quad (127)$$

where $r$ is a continuous parameter.
Morphological size II

Definition

The *dilation* (dilatation in French!) of $X$ by $B$ is defined as

$$X \oplus B = \{x + b | x \in X, b \in B\} \quad (128)$$

Definition (Discrete object size)

Let $B$ be a discrete, convex, finite, set of $\mathbb{Z}^2$, whose size is arbitrary set to $1$. Then, we build a family of homothetic versions of $B$, for $n = 0, 1, \ldots$, by

$$nB = B \oplus \ldots \oplus B \quad \text{$n-1$ dilations} \quad (129)$$

In this expression, $nB$ is obtained as a cascade of $(n - 1)$ successive dilations. By convention, for $n = 0$, $0B = \{(0,0)\}$. Note that $nB \oplus mB = (n + m) \oplus B$ for every $n, m$.

Morphological size III

Theorem

Let $\partial(B)$ be the frontier of $B$ in $\mathbb{R}^2$. Then

$$\partial(B) \oplus B = B \oplus B \quad (130)$$

Therefore,

$$nB = B \oplus \ldots \oplus B = B \oplus \partial(B) \oplus \ldots \oplus \partial(B) \quad (131)$$
Multiresolution filters

Definition

We define the multiresolution opening and closing respectively as

\[
X \circ nB = (X \ominus nB) \oplus nB \quad (132)
\]

\[
X \bullet nB = (X \oplus nB) \ominus nB \quad (133)
\]

Since \((X \ominus B) \ominus C = X \ominus (B \oplus C)\) and \((X \oplus B) \oplus C = X \oplus (B \oplus C)\),

\[
X \circ nB = \left[ (X \ominus B) \ominus B \ldots \ominus B \right] \oplus B \oplus B \ldots \oplus B \quad (134)
\]

By definition,

\[
X \circ nB = \bigcup_{(nB)_z \subseteq X} (nB)_z \quad (135)
\]

Formal definition of the skeleton I

Definition (Skeleton [Lantuéjoul])

The skeleton of a set \(X\) is the union of a family of subsets \(S_n\), each of them being the set of references to translates of \(nB\) included in \(X\), except all the references of translates of \((n + 1)B\) contained in \(X\). Formally, we have

\[
S_n(X) = (X \ominus nB) \setminus ([X \ominus nB] \circ B), \quad n = 0, 1, \ldots, N \quad (136)
\]

\[
S(X) = \bigcup_{n=0}^{N} S_n(X) \quad (137)
\]

where \(B\) is a 2 × 2 wide pixels set for a square digital grid or an hexagon for an hexagonal grid.
Inverse skeleton

We define

$$\pi(X) = [[[S_N(X) \oplus B] \cup S_{N-1}(X)] \oplus B \cup S_{N-2}(X) ...] \oplus B \cup S_0(X)$$ (138)

It can be shown that

$$\pi(X) = \bigcup_{n=0}^{N} [S_n(X) \oplus nB] = X$$ (139)

⇒ it is possible to reconstruct the shape perfectly based on the skeleton.
Alternative skeleton formulas I

**Definition (Distance function)**

The **distance function** of a set \( X \subseteq \mathcal{E} \), denoted by \( \phi(X) \), is defined as

\[
[\phi(X)](h) = d(X^c, h)
\]  

(140)

with the convention that \( d(\emptyset, h) = +\infty \) for \( h \in \mathcal{E} \).

Alternative skeleton formulas II

(a) Original image

(b) Level sets of the distance function

(c) Local maximum of (b)

(d) Skeleton by maximal balls

**Figure:** Skeleton by maximal balls.
**Alternative skeleton formulas III**

**Figure:** Some skeletons obtained with Vincent’s algorithm.

---

**Surfacic shape descriptors based on a catalog of primitives**

**Figure:** A shape is described as the union of a rectangle, a triangle and a diamond.
Summary of shape descriptors

There are three fundamental measures
- Perimeter
- Area
- Euler-Poincaré number

There is a relationship between the number of connected components $C$ and the number of holes $H$.

**Definition (Euler number)**

The **Euler number** $E$ is defined as

$$E = C - H$$  \hspace{1cm} (141)
Object description and analysis

Shape description

Measures II

Figure: Euler number: $E = 1 - 2 = -1$.

Measures III

Figure: Two letters with a different Euler number.
Shape factors

**Figure:** An object X and two reference sets (circumscribing disk/ellipse and smallest rectangular bounding box R).

<table>
<thead>
<tr>
<th>Shape factor name</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>compactness</td>
<td>$\frac{P(X)^2}{4\pi A(X)}$</td>
</tr>
<tr>
<td>rectangularity</td>
<td>$\frac{A(X)}{A(R)}$</td>
</tr>
<tr>
<td>circularity</td>
<td>$\frac{4A(X)}{\pi D_{\text{max}}^2}$</td>
</tr>
<tr>
<td>anisometry</td>
<td>$\frac{D_{\text{max}}}{D_{\text{min}}}$</td>
</tr>
</tbody>
</table>

Moments

Moments are surfacic measures, used for example in character recognition. Remember that we can see an object as a binary function

$$f(x, y) = \begin{cases} 1 & \text{if } (x, y) \text{ belongs to the object} \\ 0 & \text{otherwise} \end{cases} \quad (142)$$

**Definition (Moment of order $p + q$)**

The $p + q$ moment of a function $f(x, y)$ is defined as

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \quad (143)$$

All these moments can be centered:

$$\mu_{pq} = \sum_x \sum_y (x - m_{10})^p (y - m_{01})^q f(x, y) \quad (144)$$

and they can be normalized:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{p+3}} \quad (145)$$
Granulometry is an approach to compute a size distribution of grains in binary images, using a series of opening operations.

Consider a ball $nB$ whose radius is given by $n$, then we define a granulometric curve by

$$\psi(n) = A(X \circ nB)$$

where $A()$ denotes the area.

Abrupt changes in this curve are interesting for detecting typical sizes.

**Definition (Pattern spectrum)**

The Pattern Spectrum (PS) is defined as

$$PS(n) = -\frac{A(X \circ nB)}{dn}$$

Figure: A binary object and a corresponding granulometric curve.
**Definition (Convexity)**

A set $X \subseteq \mathcal{E}$ is **convex** if $rx + (1 - r)y \in X$ for every $x, y \in X$ where $r \in [0, 1]$.

In other words, a line joining two arbitrary points of $X$ has to be entirely included in $X$.

![Figure: A, B are convex; C is not convex.](image)

**Convex envelope I**

The notion of convexity leads to that of convex envelope

**Definition (Convex envelope)**

The **convex envelope** of a set $X \subseteq \mathcal{E}$, denoted by $co(X)$, is the intersection of all the convex sets comprising $X$.

![Figure: A set $X$ and its convex envelope.](image)
Convex envelope II

**Theorem**

*For* $X, Y \subseteq \mathcal{E}$,

$$\co(X \oplus Y) = \co(X) \oplus \co(Y)$$

(148)

**Definition (Convexity shape factor)**

The **convexity shape factor** is defined as

$$C = \frac{A(X)}{A(\co(X))}$$

(149)

---

**Outline**

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
   - Motivation
   - General view
   - Deep learning building blocks
   - Applications of deep learning with images
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
Deep learning for visual understanding

Anthony Cioppa, Marc Van Droogenbroeck

Academic year: 2019-2020

Outline

1 Motivation
2 General aspects of deep learning
3 Deep learning building blocks for visual understanding
4 Applications of deep learning in computer vision
**Deep learning research timeline**

*Figure: Timeline of the research in deep learning. The first principles, like the perceptron, were invented 60 years ago.*
Why now and not before?

- Lack of processing power
  - No GPUs at the time.
  - Alexnet was published recently (in 2013).

- Lack of data
  - No big, annotated datasets at the time.
  - The Imagenet dataset was only published in 2009.

- Overfitting
  - Because of the above, models could not generalize all that well.
Deep learning today

- Deep Learning is used for many tasks
  - Object classification
  - *Object detection*, segmentation
  - Pose estimation
  - Image captioning
  - Question answering
  - Speech recognition
  - Robotics
- But some tasks remain difficult
  - Action classification
  - Action detection

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
   - Motivation
   - General view
   - Deep learning building blocks
   - Applications of deep learning with images
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
The goal is to output the correct labels by optimizing on the $\theta$ parameters of the model.

This is done by minimizing a **loss function**.

This loss function takes different forms based on the typology of the problem.

Update on the $\theta$ parameters is done using backpropagation of the error.
Mathematical formulation

Over a training set \((X, Y)\), where \(X\) contains the inputs \((x)\) and \(Y\) contains the labels \((y)\), the goal is to find the optimal parameters \(\theta^*\) of the model such that:

\[
\theta^* \leftarrow \text{argmin}_\theta \sum_{(x, y) \in (X, Y)} L(y, \hat{y}(x, \theta))
\]  

(150)

where \(L(y, \hat{y}(x, \theta))\) is the loss function between the true label \(y\) and the predicted output \(\hat{y}\).

Some loss functions

- Mean squared error (MSE) for regression

\[
L(y, \hat{y}) = (y - \hat{y})^2
\]  

(151)

- Cross entropy for binary classification

\[
L(y, \hat{y}) = -y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})
\]  

(152)
Some typical losses

### Classification
- Log Loss
- Focal Loss
- KL Divergence/Relative Entropy
- Exponential Loss
- Hinge Loss

### Regression
- Mean Square Error/Quadratic Loss
- Mean Absolute Error
- Huber Loss/Smooth Mean Absolute Error
- Log cosh Loss
- Quantile Loss

---

**Learning process: loss surface**

- Local Minima
- Saddle Point
- Global Minima
Gradient descend

The weights are updated, at each epoch $k$, according to the following equation:

$$
\theta_{k+1} \leftarrow \theta_k - \eta_k \frac{\delta L}{\delta \theta_k}
$$

(153)

where an *epoch* is when the entire dataset passes both forward and backward through the neural network once and $\eta_k$ is the *learning rate* at epoch $k$.

Mini-batch stochastic gradient descend

The weights are updated, at each mini-batch $j$, according to the following equation:

$$
\theta_{j+1} \leftarrow \theta_j - \frac{\eta_j}{|B_j|} \sum_{i \in B_j} \frac{\delta L_i}{\delta \theta_j}
$$

(154)

where $B_j$ is a “mini-batch” from the train set $X$. 
Gradient descend summary

The data should be split in 3 disjoint sets:

1. **Training set**
   - To optimize the parameters $\theta$ of the network.

2. **Validation set**
   - To optimize the network architecture.
   - To stop the training and avoid **overfitting**.
   - To optimize other hyper-parameters.

3. **Test set**
   - To evaluate the performances on **unseen** data.

---

**Figure**: Comparison between gradient descend and (mini-batch) stochastic gradient descend.

*No guarantee that this is what is going to always happen.
But the noisy SGC gradients can help some times escaping local optima*
Where are the difficulties?

It might seem now that deep learning can virtually solve any problem using big enough networks and computational resources. But:

- **The process of *training is complex***:
  - issues with convergence
  - avoid overfitting
  - usually no knowledge about the data statistics

- **The architecture** and the type of *building blocks* used inside the network play a great role.
  - It is better to have a small architecture with operations adapted to your problem. But no rules of thumb, directly applicable, available.

---

**Outline**

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. **Deep learning for visual understanding**
   - Motivation
   - General view
   - Deep learning building blocks
   - Applications of deep learning with images
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
The single perceptron (neuron)

Figure: Illustration of a single perceptron. The output is a weighted sum of the inputs. The parameters $w_i$ are the ones to be learned.
The single perceptron (neuron)

- **Main idea:**
  - One weight \( w_i \) per input \( x_i \)
  - Compute a weighted sum of the inputs
  - Add a bias term with weight \( w_0 \) and input \( x_0 = 1 \)
  - Apply a non-linear activation function \( \sigma \) on the weighted sum

\[
\hat{y}(x, w_0, ..., n) = \sigma \left( \sum_{i=1}^{n} w_i x_i + w_0 \right)
\] (155)

**Difficulties**

- One perceptron = one decision
- Some functions cannot be represented by such a simple model

The multi-layer perceptron (MLP)

**Main idea:**

- Organize multiple perceptrons in layers and connect each neuron of a layer to all other neurons of the next layer.
- The hidden layers will capture different features of the input data.

\[
\hat{y}(x, \theta_1, ..., L) = h_L \left( h_{L-1} (...h_1 (x, \theta_1), \theta_{L-1}), \theta_L \right)
\] (156)
Some activation functions $\sigma(x)$ can be used at the output of each neurons in order to add non-linearities inside of the model.

![Activation Functions](image.png)

**Figure:** Different types of commonly used activation functions.

### High dimensional data

- Raw data live in huge dimensionalities.
- Semantically meaningful raw data prefer lower dimensional manifolds.
- The goal is to discover this manifold to embed our data.

![PCA and t-SNE Manifolds](image.png)

**Figure:** Pixel vs feature representation of the digits of the MNIST dataset.
Tensorflow playground

It is possible to visualize the training of a neural network thanks to tensorflow playground:

https://playground.tensorflow.org

Figure: Tensorflow playground running in a web browser.

Deep learning with images

What makes images different?

1920x1080x3 = 6,220,800 input variables

Figure: The number of input variables in images explodes.
The image data structure

- Images are 2-dimensional with either:
  - 1 channel (e.g. grey images)
  - 3 channels (e.g. RGB)
  - $k$ channels (e.g. multi-spectral)

- Huge dimensionality
  - A $256 \times 256$ RGB image amounts to $200K$ input variables.
  - A 1-layered NN with 1000 neurons would have 200 millions parameters.

Spatial correlation

- Images are stationary signals, they share features spatially.
  - After modifications (color, rotation, flipping, ...), images are still meaningful.
  - Small visual changes are often invisible to naked eye.
  - Basic natural image statistics are the same.

- Neighboring variables (pixels) are correlated!
  - Information from neighboring pixels is meaningful when grouped.

Figure: The black pixel of the eye of the cat only makes sense in its own context.
Filters to analyze images

- Several, handcrafted filters in computer vision:
  - Canny, Sobel, Gaussian blur, smoothing, morphological filters, Gabor filters
- Are they optimal for feature extraction and recognition?

\[ \text{e.g. Sobel 2-D filter} \]

\[ \begin{bmatrix}
-1 & 0 & 1 \\
-2 & 0 & 2 \\
-1 & 0 & 1
\end{bmatrix} \]

**Figure:** Example of the well-known handcrafted Sobel 2-D filter.

Learning the filters to use - Convolutional layer

**Convolutional layer**

The main task of the convolutional layer is to detect local conjunctions of features from the previous layer and mapping their appearance to a feature map. The goal is to optimize the \( \theta_{ij} \) parameters of the filters.

\[
\text{filter} \equiv \begin{bmatrix}
\theta_{11} & \theta_{12} & \theta_{13} \\
\theta_{21} & \theta_{22} & \theta_{23} \\
\theta_{31} & \theta_{32} & \theta_{33}
\end{bmatrix}
\]  

(157)
**Quick reminder on convolutions I**

Original image

**Quick reminder on convolutions II**

Original image

Convolutional filter 1

```
1 1 1 0 0
0 1 1 1 0
0 0 1 1 1
0 0 1 1 0
0 1 1 0 0
```
**Quick reminder on convolutions III**

Original image

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Convolutional filter 1

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Convolving the image

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x₁</td>
<td>1</td>
<td>x₀</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>x₀</td>
<td>1</td>
<td>x₁</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>x₁</td>
<td>0</td>
<td>x₀</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Result

4

\[ I(x, y) * h = \sum_{i=-a}^{a} \sum_{j=-b}^{b} I(x - i, y - j) \cdot h(i, j) \]

**Quick reminder on convolutions IV**

Original image

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Convolutional filter 1

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Convolving the image

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x₁</td>
<td>1</td>
<td>x₀</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>x₀</td>
<td>1</td>
<td>x₁</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>x₁</td>
<td>0</td>
<td>x₀</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Result

4 \( \text{Inner product} \quad 3 \)

\[ I(x, y) * h = \sum_{i=-a}^{a} \sum_{j=-b}^{b} I(x - i, y - j) \cdot h(i, j) \]
Filter convolution for $k$ channels images

Figure: Representation of a filter for $k$ channels.
Towards the convolutional layer I

How many weights for this neuron?

\[ 7 \cdot 7 \cdot 3 = 147 \]

Towards the convolutional layer II

The activations of a hidden layer form a volume of neurons, not a 1-d "chain". This volume has a depth 5, as we have 5 filters.

How many weights for these 5 neurons?

\[ 5 \cdot 7 \cdot 7 \cdot 3 = 735 \]
Towards the convolutional layer III

Assume the image is 30x30x3.
1 filter every pixel (stride = 1)
How many parameters in total?

24 filters along the x axis
24 filters along the y axis
Depth of 5
\[
\times 7 \times 7 \times 3 \text{ parameters per filter}
\]
\[
423K \text{ parameters in total}
\]

Towards the convolutional layer IV

- So, if we are anyways going to compute the same filters, why not share?
  - Sharing is caring

Assume the image is 30x30x3.
1 column of filters common across the image.
How many parameters in total?

Depth of 5
\[
\times 7 \times 7 \times 3 \text{ parameters per filter}
\]
\[
735 \text{ parameters in total}
\]
Pooling the features

- Aggregate multiple values into a single value. Advantages:
  - is invariant to small transformations.
  - reduces the size of the layer output/input to next layer.
  - keeps most important information for the next layer.

- Examples:
  - Max pooling
  - Average pooling
Max pooling

- Run a sliding window.
- At each location keep the maximum value.
- The most usual choice of pooling.
- The goal is to keep the most intense activations.

<table>
<thead>
<tr>
<th>12</th>
<th>20</th>
<th>30</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>12</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>70</td>
<td>37</td>
<td>4</td>
</tr>
<tr>
<td>112</td>
<td>100</td>
<td>25</td>
<td>12</td>
</tr>
</tbody>
</table>

$2 \times 2$ Max-Pool

20  30
112 37

General structure of a Convolutional Neural Network (CNN)

- Convolution layer followed by activations and max pooling.
- Repeat this configuration several times.
- The end of the network is a MLP.

Figure: General structure of a CNN for image classification.
End-to-end learning of feature hierarchies

- A pipeline of successive modules.
- Each module’s output is the input for the next module.
- Modules produce features of higher and higher abstractions.
  - Initial modules capture low-level features (e.g. edges or corners).
  - Middle modules capture mid-level features (e.g. circles, squares, textures).
  - Last modules capture high level, class specific features (e.g. face detector).

![Feature hierarchy diagram]

Residual blocks

- Residual blocks allow to only learn the residuals that needs to be added to the input rather than the whole transformation.
  - Useful for learning transformations (e.g. semantic segmentation).
  - Allows to train deeper networks.

![Residual block diagram]

**Figure**: Residual block: the weights learn the *residues* to be added to the original features.
 Dropout I

- During training, some activations are randomly set to 0.
  - Neurons sampled at random from a Bernoulli distribution with $p = 0.5$.
- At test time all neurons are used.
- Benefits:
  - Reduce complex co-adaptations or co-dependencies between neurons.
  - Every neuron becomes more robust.
  - Significantly decrease the overfitting.
  - Improve the training speed significantly.
Dropout II

- Effectively, a different architecture is used at every training epoch.

![Figure: The network and thus the path to the output is different at every epoch but the weights are still shared.](image)

Data normalization

- Center data to be roughly 0.
  - Activation functions are usually “centered” around 0.
  - The gradient of most activation functions is maximum around 0.
  - Convergence is usually faster.

![Figure: Most activation functions are centered around zero.](image)
Data normalization

- **Hypothesis:**
  - Input variables follow a Gaussian distribution (roughly)

- **Goal:**
  - Go from $\mathcal{N}(\mu, \sigma^2)$ to $\mathcal{N}(0, 1)$

- **In practice:**
  - First, compute the mean and standard deviation on the training set.
  - Then subtract the mean from training samples.
  - Finally, divide the result by the standard deviation.

\[ \mu, \sigma \]

**Figure:** Normalization of the distribution in the training set.

Data normalization in images

- If dimensions have similar values (e.g., pixels in natural images):
  - Compute one $\left(\mu, \sigma^2\right)$ instead of as many as the input variables (not per pixel).
  - Or the per color channel pixel mean and variance.

\[ \left(\mu_{\text{red}}, \sigma_{\text{red}}^2\right), \left(\mu_{\text{green}}, \sigma_{\text{green}}^2\right), \left(\mu_{\text{blue}}, \sigma_{\text{blue}}^2\right) \] (158)

- When input dimensions have similar ranges centering might be enough.
  - All pixels have the same range so dividing by 255 and centering might be enough normalization in some cases.
Batch normalization

- It is also possible to normalize the inputs of hidden layer over a mini-batch.
  - The weights change the distribution of the features inside of the network and it might be good to re-center them.
  - It has two learnable parameters $\gamma$ and $\beta$ to shift and scale the normalization.

$$
\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{159}
$$

$$
\hat{y}_i = \gamma\hat{x}_i + \beta \tag{160}
$$

Figure: Batch normalization for the features inside of the networks.

Class imbalance

- Classes might not be represented by the same number of samples.
  - Most datasets do not have exactly equal number of instances in each class.
  - The training can converge to simpler sub-optimal solutions, such as always predicting the most frequent class.
  - The evaluation can be biased. For example, in a dataset with 20:1 ratio, the accuracy of a classifier always predicting the first class is of $\approx 95\%$.

Possible solutions

- Weighting the samples in the loss function.
- Resampling the dataset.
- Use data augmentation for the less frequent classes.
Data augmentation

- Data augmentation can be performed in order to increase the number of training samples.
  - Translations, rotations, warping,…
  - Only use transformations that are relevant for your problem.

![Examples of data augmentation in the data space.](image)

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
   - Motivation
   - General view
   - Deep learning building blocks
   - Applications of deep learning with images
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
Classification

- **Goal:**
  - Classify each image in one of the $N$ classes.

- **Datasets:**
  - MNIST
  - ImageNet
  - ...

- **Networks:**
  - LeNet-5
  - ResNet
  - ...

**Figure:** Examples of a 5-class classification problem for images.
Regression

- **Goal:**
  - Regress one or many continuous values.
  - Object detection

- **Datasets:**
  - Pascal VOC
  - COCO
  - ...

- **Networks:**
  - MaskRCNN
  - Yolo
  - ...

**Figure:** Examples of a bounding box regression problem.

Semantic segmentation

- **Goal:**
  - Assign a class to each pixel.

- **Datasets:**
  - Kitti, CityScapes,...

- **Networks:**
  - PSPNet, MaskRCNN,...

**Figure:** Semantic segmentation for player segmentation in soccer.
Interpretability

- It is important to understand what happens inside the network.
  - Medical applications
  - Self-driving cars
- Some tools can help to understand how the network understands its environment.
  - Capsule networks, prototypes, etc

![Figure: Prototypes of the MNIST digit classes conceptualized by a network (HitNet).](image)

Conclusion

Take away messages

- Understand your data.
- Use an appropriate architecture and operations relevant and for your data.
- Use a valid evaluation methodology.
- Try to interpret the results.
The content of this presentation is partly inspired by the slides of Efstratios Gavves from the University of Amsterdam (UVA). Some slides and figures are directly copied from that source. http://uvadlc.github.io/

Figure: YOLO models detection as a regression problem.
Some characteristics of YOLO

**Loss:** YOLO uses sum-squared error between the predictions and the ground truth to calculate loss. The loss function is composed of three terms:

1. the **classification** loss
2. the **localization** loss (errors between the predicted boundary box and the ground truth)
3. the **confidence loss** (the objectness of the box).

**Implementations:** YOLO, YOLOv2, YOLO9000, TinyYOLO, YOLOv3.

**Read:** [https://pjreddie.com/darknet/yolo/](https://pjreddie.com/darknet/yolo/) or other similar resource

Design possibilities

1. **Re-use an existing architecture network,** which you then train on your own annotated data.
2. **Design your own network** (or mimic an existing one)
3. **Skip one of the most computational intensive tasks during the training process** and **simplify the training process by using a pre-trained model** (for example, convolutional weights that are pre-trained on Imagenet). The ideas are the following:
   1. process the image with a pre-trained network to transform the image into a compressed form (also to reduce the amount of input data); this is equivalent to a feature extraction task.
   2. input the features extracted in one of your own network which might include a few layers (fully connected layers, softmax, etc).
   3. **train your network,** which should be faster than training a network starting from the raw image data.
Outline

1 Fundamentals of 2D imaging
2 3D vision
3 Linear filtering
4 Edge detection
5 Introduction to machine learning
6 Performance analysis
7 Segmentation
8 Object description and analysis
9 Deep learning for visual understanding
10 Template Matching & Image Registration
11 Mathematical morphology
12 Non-linear filtering
13 Motion analysis and background subtraction
14 Feature detection and tracking

Template Matching & Image Registration

Philippe Latour

Academic year: 2019-2020
Table of content

- Introduction & Applications
  - Definition & Examples
  - Applications
- Some solutions and their corresponding approach
- Template matching & image registration components
- Implementation speed-up
Template matching - Definition

Template matching is the process of either finding any instance of a template image $T$ within another image $I$ or finding which ones of the templates $T_1, T_2, \ldots, T_N$ correspond in some way to another image $I$.

We search how to transform or warp a template (resp. image) to make it similar to a reference image (resp. template).

The template is also called the pattern or the model.

**Figure:** Find where the template $T$ is located in the observed image $I$.

**Figure:** Find the template $T_k$ which correspond to the observed image $I$. 
Image registration or alignment - Definition

Definition

Image registration is the process of spatially aligning two images of a scene/object so that corresponding points assume the same coordinates.

- Given two images taken, for example,
  - at different times,
  - from different devices
  - or different point of view;

- the goal is to determine a reasonable transformation of the images such that a transformed version of the first image is similar to the second one.
Machine vision (and especially template matching) is pervasive in almost all steps of an industrial production chain.

Machine vision II

- **Finding the image transform or warping** that would be needed to fit or align a source image on another one (Pose estimation & object alignment).
  → Robot arm control, gauging
Machine vision III

- **Measuring and assessing** the matching quality, the component presence/absence/type and/or position and orientation
  → Non destructive testing (NDT), defect detection, model conformance assessment

Machine vision IV

- **Counting the number of instances** that matched the template
  → Detection of given objects, computing their location, type and/or properties
Wafer dicing is the process by which die are separated from a wafer of semiconductor.
Die bonding is the process of attaching the semiconductor die either to its package or to some substrate.

Wire bonding is the process of making interconnections between an integrated circuit or other semiconductor device and its packaging.

Position of picked components

Position of placement area

Control of welding after the process
Multi-view correspondences I

- **3D reconstruction**
  → find the correspondences between the left and right view of the same scene

Multi-view correspondences II

- **Panoramic images**: Image alignment for stitching
  → find correspondences between several views of the same scene
Multi-modality correspondences and fusion

- Image alignment and fusion
- Remote sensing: satellite image fusion

Biomedical (elastic) image registration

Non-rigid (elastic) image registration

From [Modersitzki Jan, “FAIR Flexible Algorithms for Image Registration”, 2009, Figure 1.1]
Naive solution I

We consider:

- the image $I$ with $L_I$ lines and $K_I$ columns, represented by the matrix $I = (I_{lk})$ where $l \in [0, L_I]$, $k \in [0, K_I]$,
- the template $T$ with $L_T$ lines and $K_T$ columns, represented by the matrix $T = (T_{ij})$ where $i \in [0, L_T]$, $j \in [0, K_T]$. 

![Diagram of Template Matrix T and Observed Image Matrix I]
Naive solution II

We define all the admissible (sub-)windows $W^{(l,k)}$ completely included within the image $I$ and of the same size as the template $T$ by the following sub-matrices $W^{(l,k)}$:

$$W^{(l,k)}_{ij} = \begin{cases} 
I_{l+i,k+j} & \text{for } i \in [0, L_T - 1], \ j \in [0, K_T - 1] \\
0 & \text{otherwise}
\end{cases}$$

where $l \in [0, L_I - L_T]$ and $k \in [0, K_I - K_T]$ are the indices, in the image $I$, of the upper left pixel of $W^{(l,k)}$.

Compute the Euclidean distance

$$\text{dist} \left( T, W^{(l,k)} \right) = \sqrt{\sum_{i=0}^{L_T-1} \sum_{j=0}^{K_T-1} \left[ T_{ij} - W^{(l,k)}_{ij} \right]^2}$$

then create the distance map $D$, represented by the matrix $D$:

$$D\left[ \left\lfloor \frac{K_T}{2} \right\rfloor + k, \left\lfloor \frac{L_T}{2} \right\rfloor + l \right] = \begin{cases} 
\text{dist} \left( T, W^{(l,k)} \right) & \text{for } l \in [0, L_I - L_T], \ k \in [0, K_I - K_T] \\
0 & \text{otherwise}
\end{cases}$$

and find the position of the minimum in these map.
Previously, in the naive solution, we matched the template $T$ to a part of the translated version of the image (the sub-image function $W_{k,l}(x, y) = I(x + k, y + l)$).

In pixel coordinates, this may be illustrated by:

It is equivalent to match a translated template to the original image:
The block diagram of this pixel based (naive) solution is:
Question?
Which “elements” are we going to match in the reference and template images?

▶ All pixels of an image/template:

- For all warped templates, compare all pairs of corresponding pixels (\(\equiv\) located at the same place in the images and the warped template).
- Then compute a global score based on the individuals comparisons.
- And choose the warping for which the score is maximum/minimum.

... or Feature-based?

Question?
Which “elements” are we going to match in the reference and template images?

▶ “Interesting” points of an image/template:

- First search for the feature points in the image/template,
- Then find the best matching between feature points in the image and the template,
- And finally compute the warping based on the best feature points matching.
Introduction & Applications.

Some solutions and their corresponding approach.
- Naive solution & Block diagram.
- **Pixel based approach.**
- Feature based approach.
- Comparisons.

Template matching & image registration components.

Implementation speed-up.

---

**Pixel-based approach**

Shift or **warp the images** relatively to each other, then look at how much the pixels agree and find the warping parameters for which the agreement is maximum.

So, we first need to decide which kind of warping is eligible between the template and the image.

In the previous naive solution, it is only translation.

The eligible warping defines the parameter space or the search space:
- Translation (2D) + rotation (1D) + isotropic scaling (1D) → 4D search space
- Affine / Projective transform → 6D / 8D search space

Applying the warping
- to the pattern
- or the image
- or both?

Image re-sampling and sub-pixel accuracy?
Pixel-based approach II

Shift or warp the images relatively to each other, then look at **how much the pixels agree** and find the warping parameters for which the agreement is maximum.

- Then a suitable similarity or dissimilarity measure must be chosen to compare the images.
- In the previous naive solution, it is the Euclidean distance.
- The similarity or dissimilarity measure depend on the image characteristics to which it is necessary to be invariant. The insensitivity properties guide the choice of a score/distance measure.
  - Lighting conditions (linear gain and offset)
  - Noise
  - “Small” rotation or scaling
  - Thinning
  - → Define the similarity/dissimilarity measure

Pixel-based approach III

Shift or warp the images relatively to each other, then look at how much the pixels agree and find the **warping parameters** for which the **agreement is maximum**.

- In the previous naive solution, we try all possible alignment (an exhaustive search).
- But this solution is often impractical and hierarchical coarse-to-fine techniques based on image pyramids are often used.
- The search technique must be devised:
  - Exhaustive search
  - Coarse to fine hierarchical refinement.
  - Steepest descent
  - Conjugate gradient
  - Quasi-Newton method
  - Levenberg-Marquardt
  - Simulated annealing
The block diagram of a more general solution could be
Feature-based approach

In both images, **extract feature points** and compute their **descriptor vector**. Then, **match the corresponding feature points** and compute the **image warping** that transforms at best (**maximum agreement**) each feature points into its corresponding one.

Let's consider two images of the same (or similar/related) scene/object taken

- at different moment, or
- from a different point of view, or
- under different environmental conditions, or
- with different sensor (parameters).
Feature points definition I

Feature-based approach

In both images, extract feature points and compute their descriptor vector. Then, match the corresponding feature points and compute the image warping that transforms at best (maximum agreement) each feature points into its corresponding one.

Feature points in each image carry critical information about (local) scene structure.

- Also called critical points, interest points, key points, extremal points, anchor points, landmarks, control points, tie points.
- For instance corners, vertices, junctions, edges, dark/light blob center, unique patches, moments, . . .

Feature points definition II

Feature-based approach

In both images, extract feature points and compute their descriptor vector. Then, match the corresponding feature points and compute the image warping that transforms at best (maximum agreement) each feature points into its corresponding one.

Feature points are/should be:

- Independent of noise, blurring, contrast, lightning conditions.
- Dependent or Independent of geometric changes (rotation, scaling, affine transform).
- Widely used in image analysis (not only for image registration).
Feature points definition III

Feature-based approach

In both images, extract feature points and compute their descriptor vector. Then, match the corresponding feature points and compute the image warping that transforms at best (maximum agreement) each feature points into its corresponding one.

- Defining and computing a “featureness” function
  - Invariant to some image variation (translation, rotation, scaling, brightness, contrast, ...)
  - For instance; central moments, pixel intensity variances, gradient module, LoG/DoG, entropy, Harris cornerness, ...
- And keeping only the most-significant maxima/minima/zeroes of the “featureness” function.

Feature points representation & matching I

Feature-based approach

In both images, extract feature points and compute their descriptor vector. Then, match the corresponding feature points and compute the image warping that transforms at best (maximum agreement) each feature points into its corresponding one.

Represent each feature point by a descriptor vector

- Vector of characteristic values describing the feature point in an unique and discriminatory way.
- Position, gradient, image moment, scale, orientation, Histogram Of Gradients (HOG), Local Binary Pattern (LPB), ...
Feature points representation & matching II

Feature-based approach

In both images, extract feature points and compute their descriptor vector. Then, match the corresponding feature points and compute the image warping that transforms at best (maximum agreement) each feature points into its corresponding one.

Find/match the corresponding feature points in both images:

- Compare the descriptor vectors to find the best correspondence between feature points in each images.

Compute the transformation/warping such that the feature points in the left image fit their corresponding point in the right image.

- Define which kind of warping is admissible; rigid global warping (homography) or elastic/local warping.
- Use robust fitting methods:
  - RANSAC,
  - ICP (Iterative Closest Point)
Pixel-based vs. feature based approach I

**Pixel-based approach**
Shift or *warp the images* relatively to each other, then look at how much the pixels agree and find the *warping parameters* for which the agreement is maximum.

**Feature-based approach**
In both images, *extract feature points* and compute their descriptor vector. Then, *match the corresponding feature points* and compute the *image warping* that transforms at best (*maximum agreement*) each feature points into its corresponding one.

Template matching and/or image registration ARE optimization problems

Pixel-based vs. feature based approach II

<table>
<thead>
<tr>
<th></th>
<th>Pixel based</th>
<th>Feature based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern information usage</td>
<td>Use all pixels in the pattern in an uniform way. No need to analyze or understand the pattern.</td>
<td>Find and use pattern features (most informative part of the pattern). → Sensitive operation.</td>
</tr>
<tr>
<td>Occlusion or pose variation</td>
<td>Sensitive</td>
<td>Could be design to be insensitive</td>
</tr>
<tr>
<td>Sub-pixel accuracy</td>
<td>Interpolation of the similarity/dissimilarity measure</td>
<td>Naturally accurate at the sub-pixel level.</td>
</tr>
<tr>
<td>Admissible warping</td>
<td>The choice has to be done at the beginning of the process (orientation and scaling)</td>
<td>Mostly insensitive to differences in orientation and scaling</td>
</tr>
<tr>
<td>Noise and lighting conditions</td>
<td>Sensitive</td>
<td>Naturally much more insensitive</td>
</tr>
<tr>
<td></td>
<td>Pixel based</td>
<td>Feature based</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Rigid pattern warping</strong></td>
<td>Mostly limited to rigid pattern warping</td>
<td>Enable non-rigid warping.</td>
</tr>
<tr>
<td><strong>Dimensionality of the search space</strong></td>
<td>Mostly limited to low dimensionality (the search time is exponential in the search space dimensionality)</td>
<td>Higher dimensionality search space are more easily reachable</td>
</tr>
<tr>
<td><strong>Implementation</strong></td>
<td>Easy to implement, natural implementation on GPUs</td>
<td>Much more difficult to implement and/or to optimize</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>Complexity proportional to the image size. Need specific search strategies to reach real-time.</td>
<td>Complexity roughly proportional to the number of feature points (depend more on the content of the scene than on the image size)</td>
</tr>
</tbody>
</table>

Philippe Latour  
Template Matching & Image Registration  
Academic year: 2019-2020  433 / 600

Table of content

- Introduction & Applications
- Some solutions and their corresponding approach
- Template matching & image registration components
  - Pixel-based approach: Similarity/Dissimilarity measures
  - Feature-based approach: Feature detection
- Implementation speed-up
## Pixel-based approach: Similarity/dissimilarity measures

<table>
<thead>
<tr>
<th>Similarity / Correlation / Score</th>
<th>Dissimilarity / Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation coefficient</td>
<td>L1 Norm</td>
</tr>
<tr>
<td>Tanimoto Measure</td>
<td>Median of Absolute Differences</td>
</tr>
<tr>
<td>Stochastic Sign Change</td>
<td>Square L2 Norm</td>
</tr>
<tr>
<td>Deterministic Sign Change</td>
<td>Median of Square Differences</td>
</tr>
<tr>
<td>Minimum Ratio</td>
<td>Normalized Square L2 Norm</td>
</tr>
<tr>
<td>Spearman's Rho</td>
<td>Incremental Sign Distance</td>
</tr>
<tr>
<td>Kendall’s Tau</td>
<td>Intensity-Ratio Variance</td>
</tr>
<tr>
<td>Greatest Deviation</td>
<td>Intensity-Mapping-Ratio Variance</td>
</tr>
<tr>
<td>Ordinal Measure</td>
<td>Rank Distance</td>
</tr>
<tr>
<td>Correlation Ratio</td>
<td>Joint Entropy</td>
</tr>
<tr>
<td>Energy of Joint Probability</td>
<td>Exclusive F–Information</td>
</tr>
<tr>
<td>Distribution Material</td>
<td></td>
</tr>
<tr>
<td>Similarity Shannon Mutual Information</td>
<td></td>
</tr>
<tr>
<td>Rényi Mutual Information</td>
<td></td>
</tr>
<tr>
<td>Tsallis Mutual Information</td>
<td></td>
</tr>
<tr>
<td>F-Information Measures</td>
<td></td>
</tr>
</tbody>
</table>

**Philippe Latour**

Template Matching & Image Registration

Academic year: 2019-2020  435 / 600
Pixel-based approach: Similarity/dissimilarity measures

Given two corresponding sequences of measurement \( \{a_i \mid i = 1, \cdots, n\} \) and \( \{b_i \mid i = 1, \cdots, n\} \), we will represent them by the following (column) vectors:

- \( A = (a_i)_{i=1,\cdots,n} \in \mathbb{R}^n \)
- \( B = (b_i)_{i=1,\cdots,n} \in \mathbb{R}^n \)

\( A \) and \( B \) might represent measurements from two objects or phenomena. Here, in our case, we assume they represent images and \( a_i \) and \( b_i \) are the intensities of the corresponding pixels in the images.

- The similarity (dissimilarity) between them is a measure that quantifies the dependency (independency) between the sequences.
- In the naive solution, we used the Euclidian distance (the square L2 norm).

Pearson correlation coefficient

Definition

The Pearson correlation coefficient of two vectors \( A \) and \( B \) is:

\[
 r(A, B) = \frac{\text{covar}(A, B)}{\sqrt{\text{var}(A) \text{var}(B)}} \tag{164}
\]

With the usual notations:

- \( \mu_A = \frac{1}{n} \sum_{i=1}^{n} a_i \)
- \( \mu_B = \frac{1}{n} \sum_{i=1}^{n} b_i \)
- \( \text{var}(A) = \sigma_A^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu_A)^2 = \frac{1}{n} \sum_{i=1}^{n} a_i^2 - \mu_A^2 \)
- \( \text{var}(B) = \sigma_B^2 = \frac{1}{n} \sum_{i=1}^{n} (b_i - \mu_B)^2 = \frac{1}{n} \sum_{i=1}^{n} b_i^2 - \mu_B^2 \)
- \( \text{covar}(A, B) = \frac{1}{n} \sum_{i=1}^{n} (a_i - \mu_A)(b_i - \mu_B) = \frac{1}{n} \sum_{i=1}^{n} a_i b_i - \mu_A \mu_B \)

We may write:

\[
 r(A, B) = \frac{\frac{1}{n} \sum_{i=1}^{n} (a_i - \mu_A)(b_i - \mu_B)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - \mu_A)^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (b_i - \mu_B)^2}}
\]

\[
 r(A, B) = \frac{\frac{1}{n} \sum_{i=1}^{n} a_i b_i - \mu_A \mu_B}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} a_i^2 - \mu_A^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} b_i^2 - \mu_B^2}}
\]

The Pearson coefficient \( r(A, B) \) is then easily computed with one pass on the images \( A \) and \( B \).
Pearson correlation coefficient properties I

If we introduce the reduced or normalized vectors:

$$\tilde{A} = \frac{1}{\sigma_A} (A - \mu_A 1) \text{ and } \tilde{B} = \frac{1}{\sigma_B} (B - \mu_B 1)$$

We have the following relation:

**Theorem**

$$r(A, B) = \frac{1}{n} \tilde{A}^T \tilde{B} = r(\tilde{A}, \tilde{B})$$  \hspace{1cm} (165)

Which is obvious from

$$r(A, B) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{a_i - \mu_A}{\sigma_A} \right) \left( \frac{b_i - \mu_B}{\sigma_B} \right) = \frac{1}{n} \tilde{A}^T \tilde{B} = r(\tilde{A}, \tilde{B})$$

Pearson correlation coefficient properties II

The Pearson coefficient doesn’t depend on the image’s gain and offset.

**Theorem (Invariance to affine transformation of pixel values)**

*For any values $\alpha, \beta, \gamma, \delta$ such that $\alpha \neq 0$ and $\gamma \neq 0$, we have:*

$$r(\alpha A + \beta 1, \gamma B + \delta 1) = \text{sign}(\alpha \gamma) r(A, B)$$

This result comes immediately from

$$(\alpha \tilde{A} + \beta 1) = \frac{1}{|\alpha| \sigma_A} [(\alpha A + \beta 1) - (\alpha \mu_A + \beta) 1] = \frac{\text{sign}(\alpha)}{\sigma_A} (A - \mu_A 1) = \text{sign}(\alpha) \tilde{A}$$

Then

$$r(\alpha A + \beta 1, \gamma B + \delta 1) = r(\alpha \tilde{A} + \beta 1, \gamma \tilde{B} + \delta 1)$$

$$r(\alpha A + \beta 1, \gamma B + \delta 1) = r(\text{sign}(\alpha) \tilde{A}, \text{sign}(\gamma) \tilde{B}) = \text{sign}(\alpha \gamma) r(A, B)$$
Due to the usual properties of variance and covariance;

**Theorem**

*The range of values of \( r(A, B) \) is \([-1, +1]\):*

- \( r = +1 \) if and only if \( B = \alpha A + \beta 1 \) with \( \alpha > 0 \). It is a perfect direct matching between \( A \) and \( B \).
- \( r = -1 \) if and only if \( B = \alpha A + \beta 1 \) with \( \alpha < 0 \). It is a perfect inverse matching between \( A \) and \( B \).
- \( r = 0 \) if and only if there is no linear correlation between \( A \) and \( B \).

**Pearson correlation coefficient map**

(a) A template \( T \).
(b) An image \( I \) containing the template \( T \).
(c) The correlation image \( C[T, I] \) with intensity at a pixel showing the correlation coefficient between the template and the window centered at the pixel in the image.
(d) The real part of image \( C_p[T, I] \), showing the phase correlation result with the location of the spike encircled.
The Spearman rank correlation or Spearman’s Rho ($\rho$) between vectors $A = (a_i)_{i=1,\ldots,n}$ and $B = (b_i)_{i=1,\ldots,n}$ is given by

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} [R(a_i) - R(b_i)]^2}{n(n^2 - 1)}$$  \hspace{1cm} (166)

where $R(a_i)$ and $R(b_i)$ represent ranks of $a_i$ and $b_i$ in images $A$ and $B$.

Remark: to eliminate possible ties among discrete intensities in images, the images are smoothed with a Gaussian of a small standard deviation, such as 1 pixel, to produce unique floating-point intensities.

Comparison with the Pearson correlation coefficient:

- $\rho$ is less sensitive to outliers and, thus, less sensitive to impulse noise and occlusion.
- $\rho$ is less sensitive to nonlinear intensity difference between images than Pearson correlation coefficient.
- Spearman’s $\rho$ consistently produced a higher discrimination power than Pearson correlation coefficient.
- Computationally, $\rho$ is much slower than $r$ primarily due to the need for ordering intensities in $I$ and $J$. 
Kendall’s tau I

- If \( a_i \) and \( b_i \), for \( i = 0, \ldots, n \), show intensities of corresponding pixels in \( A \) and \( B \), then for \( i \neq j \), two possibilities exist:
  - Either concordance: \( \text{sign}(a_j - a_i) = \text{sign}(b_j - b_i) \)
  - Or discordance: \( \text{sign}(a_j - a_i) = -\text{sign}(b_j - b_i) \)

- Assuming that out of the \( C_n^2 \) possible combinations, \( N_c \) pairs are concordant and \( N_d \) pairs are discordant, Kendall’s \( \tau \) is defined by:
  \[
  \tau = \frac{N_c - N_d}{\frac{n(n-1)}{2}}
  \] (167)

- If bivariate \((A, B)\) is normally distributed, Kendall’s \( \tau \) is related to Pearson correlation coefficient \( r \) by:
  \[
  r = \sin\left(\frac{\pi \tau}{2}\right)
  \] (168)

Kendall’s tau II

- Comparison with other similarity measures:
  - Pearson correlation coefficient can more finely distinguish images that represent different scenes than Kendall’s \( \tau \).
  - Conversely, Kendall’s \( \tau \) can more finely distinguish similar images from each other when compared to Pearson correlation coefficient.
  - Spearman’s \( \rho \) and Kendall’s \( \tau \) have the same discrimination power when comparing images of different scenes.
  - Kendall’s \( \tau \) is one of the costliest similarity measures.
Spearman’s rho and Kendall’s tau maps

Spearman’s Rho

Kendall’s Tau.

Table of content

- Introduction & Applications
- Some solutions and their corresponding approach
- Template matching & image registration components
  - Pixel-based approach: Similarity/Dissimilarity measures
  - Feature-based approach: Feature points detection
- Implementation speed-up
A large number of point detectors have been developed throughout the years:

- Corner-based detectors
- Edge-based detectors
- Model-based detectors
- Uniqueness-based detectors
- Curvature-based detectors
- Laplacian-based detectors
- Gradient-based detectors
- Hough Transform-based detectors
- Symmetry-based detectors
- Filtering-based detectors
- Transform Domain detectors
- Pattern Recognition-based detectors
- Moment-based detectors
- Entropy-based detectors
The angle between the line connecting pixel \((x, y)\) to the \(i\)th pixel on the smallest circle and the \(x\)-axis is \(\theta_i\), and the intensity at the \(i\)th pixel is \(I_1(\theta_i)\).

If \(\tilde{I}_j(\theta_i)\) represents the normalized intensity at \(\theta_i\) in the \(j\)th circle, then

\[
C(x, y) = \sum_{i=1}^{n} \prod_{j=1}^{m} \tilde{I}_j(\theta_i)
\]

is used to measure the strength of a vertex or a junction at \((x, y)\).

In the following formula, if \(m = 2\) (2 circles), \(C(x, y)\) is the Pearson coefficient of the two vectors \(\tilde{I}_1(\theta_i)\) and \(\tilde{I}_2(\theta_i)\).

Pixel \((x, y)\) is then considered a **corner** if \(C(x, y)\) is locally maximum.
A number of detectors use either the Laplacian of Gaussian (LoG) or the difference of Gaussians (DoG) to detect points in an image.

For the following development, we consider a “continuous” image

\[ I : \Omega \subset \mathbb{R}^2 \to \mathbb{R} : (x, y) \in \Omega \to I(x, y), \tag{170} \]

and we define a scaled version of this image

\[ L(x, y; t) = G(x, y; t) \otimes I(x, y), \tag{171} \]

where \( G(x, y; t) = \frac{1}{2\pi t} \exp\left(-\frac{x^2+y^2}{2t}\right) \) is the Gaussian kernel of variance \( \sigma^2 = t \).

The Laplacian of Gaussian \( \text{LoG} = \triangle L(x, y; t) \) is defined as

\[ \triangle L(x, y; t) = \triangle [G(x, y; t) \otimes I(x, y)] = [\triangle G(x, y; t)] \otimes I(x, y). \tag{172} \]

So, we compute the laplacian of the Gaussian

\[ \triangle G(x, y; t) = \frac{1}{2\pi t^3} \left[x^2 + y^2 - 2t\right] \exp\left(-\frac{x^2 + y^2}{2t}\right) \tag{173} \]

and its partial derivative relatively to \( t \)

\[ \frac{\partial}{\partial t} G(x, y; t) = \frac{1}{4\pi t^3} \left[x^2 + y^2 - 2t\right] \exp\left(-\frac{x^2 + y^2}{2t}\right). \tag{174} \]

Then, we deduce

\[ \triangle L(x, y; t) = 2 \frac{\partial}{\partial t} L(x, y; t), \tag{175} \]

which is the diffusion equation of the well-known Gaussian scaling process.
The previous equation demonstrates that the DoG operator is an approximation to the LoG operator.

The best approximation to the LoG operator of standard deviation $\sigma$ is the difference of Gaussians of standard deviations $\sigma$ and $1.6\sigma$. That is

$$\triangle L(x, y; \sigma) = \frac{1.6[L(x, y; 1.6\sigma) - L(x, y; \sigma)]}{\sigma^2}. \quad (176)$$

Local extrema of LoG or its approximation DoG detect **centers of bright or dark blobs** in an image.

- So, they are not as much influenced by noise as points representing corners and junctions and points detected by the LoG operator are generally more resistant to noise than points detected by vertex and junction detectors.

SIFT (Scale Invariant Feature Transform) used the difference of Gaussians (DoG) to find points in an image.
Introduction & Applications

Some solutions and their corresponding approach

Template matching & image registration components

Implementation speed-up
- FFT
- Multiresolution: Coarse to fine
- Hybrid approach: feature extraction in one image only

The Pearson coefficient may also be computed thanks to the 2D Discrete Fourier Transform (2D-DFT) and the cross-correlation of the images.

Definition
A discrete image of \( L \) lines and \( K \) columns is first represented by:
the matrix of its pixel values \( I = (I_{lk}) \)
where \( l = 0, \cdots, L - 1 \) and \( k = 0, \cdots, K - 1 \),
and then, we extend the image by defining \( a_{lk} = 0 \)
when \( l \notin [0, L - 1] \) or \( k \notin [0, K - 1] \)
Definition (2D-DFT on a grid of $L$ lines by $K$ columns)

The 2D-DFT of $I$ is the matrix

\[
I = \mathcal{F}[I] = (\alpha_{uv})
\]

where $u = 0, \cdots, L - 1$ and $v = 0, \cdots, K - 1$, and

\[
\alpha_{uv} = \sum_{l=0}^{L-1} \sum_{k=0}^{K-1} I_{lk} e^{-j2\pi(\frac{lu}{L} + \frac{kv}{K})} = \sum_{l,k \in \mathbb{Z}} I_{lk} e^{-j2\pi(\frac{lu}{L} + \frac{kv}{K})}
\]

And the inverse formulas are

\[
I_{lk} = \frac{1}{LK} \sum_{u=0}^{L-1} \sum_{v=0}^{K-1} \alpha_{uv} e^{j2\pi(\frac{lu}{L} + \frac{kv}{K})}
\]

In the previous definition, we extend the range of values of $u$ and $v$ and obtain a periodic “infinite” array, thanks to the following relation:

\[
\alpha_{u+L,v+K} = \sum_{l,k \in \mathbb{Z}} I_{lk} e^{-j2\pi(\frac{(l+u)L}{L} + \frac{(k+v)K}{K})}
\]

\[
\alpha_{u+L,v+K} = \sum_{l,k \in \mathbb{Z}} I_{lk} e^{-j2\pi(\frac{lu}{L} + \frac{kv}{K})} e^{-j2\pi(l+k)} = \alpha_{uv}
\]

So, we define $\alpha_{uv}$ for all integral values of $u$ and $v$ by $\alpha_{u+nL,v+mK} = \alpha_{uv}$. 
Definition

The cross-correlation of two “extended” images \( I \) (initially \( L_I \) lines by \( K_I \) columns) and \( T \) (initially \( L_T \) lines by \( K_T \) columns) is

\[
\Gamma \left( I, T \right) = \mathbf{C} = \left( c_{lk} \right)
\]

where

\[
c_{lk} = \sum_{n=0}^{L_I-1} \sum_{m=0}^{K_I-1} I_{nm} T_{n+l,m+k} = \sum_{n,m \in \mathbb{Z}} I_{nm} T_{n+l,m+k}
\]

with \( c_{lk} \neq 0 \) if \( l = -L_I + 1, \ldots, L_T - 1 \) and \( k = -K_I + 1, \ldots, K_T - 1 \), and \( L_C = L_I + L_T - 1 \) and \( K_C = K_I + K_T - 1 \).

The 2D-DFT of the cross-correlation of these two images is

\[
\gamma_{uv} = \sum_{l,k \in \mathbb{Z}} c_{lk} e^{-j2\pi \left( \frac{lu}{L_C} + \frac{kv}{K_C} \right)}
\]

\[
\gamma_{uv} = \sum_{l,k \in \mathbb{Z}} \left( \sum_{n,m \in \mathbb{Z}} a_{nm} b_{n+l,m+k} \right) e^{-j2\pi \left( \frac{lu}{L_C} + \frac{kv}{K_C} \right)}
\]

\[
\gamma_{uv} = \sum_{n,m \in \mathbb{Z}} a_{nm} \sum_{l,k \in \mathbb{Z}} b_{n+l,m+k} e^{-j2\pi \left( \frac{lu}{L_C} + \frac{kv}{K_C} \right)}
\]

\[
\gamma_{uv} = \sum_{n,m \in \mathbb{Z}} a_{nm} \sum_{l',k' \in \mathbb{Z}} b_{l',k'} e^{-j2\pi \left( \frac{(l' - n)u}{L_C} + \frac{(k' - m)v}{K_C} \right)}
\]

\[
\gamma_{uv} = \sum_{n,m \in \mathbb{Z}} a_{nm} e^{+j2\pi \left( \frac{nu}{L_C} + \frac{mv}{K_C} \right)} \sum_{l',k' \in \mathbb{Z}} b_{l',k'} e^{-j2\pi \left( \frac{lu}{L_C} + \frac{k'v}{K_C} \right)}
\]

\[
\gamma_{uv} = \alpha_{uv}^* \beta_{uv}
\]
To speed-up the search step we can use the DFT:

- $\tilde{T}$ represents the “extended” normalized template.
- $\tilde{I}$ represents the “extended” normalized image inside which $\tilde{T}$ is to be found.
- The best-matching template window in the image is located at the peak of

$$
C[T, I] = \frac{1}{n} \Gamma(\tilde{T}, \tilde{I}) = \frac{1}{n} F^{-1} \left\{ F\{\tilde{T}\}^* F\{\tilde{I}\} \right\}
$$

Phase correlation: the information about the displacement of one image with respect to another is included in the phase component of the cross-power spectrum of the images:

$$
C_p[T, I] = F^{-1} \left\{ \frac{F\{\tilde{T}\}^* F\{\tilde{I}\}}{\|F\{\tilde{T}\}^* F\{\tilde{I}\}\|} \right\}
$$

Multiresolution - Coarse-to-fine approach

- Compute image and pattern down-scaled pyramids.
- Proceed to a full search of the most reduced (coarser) pattern within the most reduced image.
- Find a number of possible candidates at the coarsest scale by an exhaustive search.
- For each candidates at a given scale:
  - Upscale the image and the candidate and look for the best matching pattern location in a neighborhood of the candidate.
  - Reduce the number of candidates
  - If the finer scale has not yet been reached, proceed to the next scale level
Hybrid approach: feature extraction in one image only

- Search for some feature points in the template.
- Consider very simple feature points as
  - the local maxima of the gradient,
  - regularly spaced patches.
- Scan the warping parameter space following a given strategy:
  - Transform the feature points of the template following the current eligible warping parameters.
  - Superimpose the warped feature points (of the template) on the observed image.
  - At each warped feature points location in the observed image, check if a compatible feature point exists in the observed image and measure its similarity/dissimilarity score.
  - Compute a global measure of similarity/dissimilarity by adding all the individual scores of the feature points.
  - Find the optimum of this measure on the search space.
- “Detect and track” instead of “detect and match”.

Bibliography

A. Goshtasby.  

R. Brunelli.  

R. Szeliski.  
Mathematical morphology

- Reminders of the set theory
- Basic morphological transforms
- Neighboring transformations
- Geodesy and reconstruction
- Grayscale morphology
Sets will be denoted with capital letters, such as $A$, $B$, $\ldots$, and elements of these sets by lowercase letters $a$, $b$, $\ldots$.

- **Set equality**
  Two sets are equal if they contain the same elements: $X = Y \iff (x \in X \Rightarrow x \in Y$ and $x \in Y \Rightarrow x \in X)$. The empty set is denoted as $\emptyset$.

- **Inclusion**
  $X$ is a subset of $Y$ (that is, $X$ is included in $Y$) if all the elements of $X$ also belong to $Y$: $X \subseteq Y \iff (x \in X \Rightarrow x \in Y)$.

- **Intersection**
  The intersection between $X$ and $Y$ is the set composed of the elements that belong to both sets: $X \cap Y = \{x \text{ such that } x \in X \text{ and } x \in Y\}$.

- **Union**
  The union between two sets is the set that gathers all the elements that belong to at least one set: $X \cup Y = \{x \text{ such that } x \in X \text{ or } x \in Y\}$.

- **Difference**
  The set difference between $X$ and $Y$, denoted by $X - Y$ or $X \setminus Y$ is the set that contains the elements of $X$ that are not in $Y$:
  $X - Y = \{x | x \in X \text{ and } x \notin Y\}$.
Complementary
Assume that $X$ is a subset of a $\mathcal{E}$ space, the complementary set of $X$ with respect to $\mathcal{E}$ is the set, denoted $X^c$, given by $X^c = \{x \text{ such that } x \in \mathcal{E} \text{ and } x \notin X\}$.

Symmetric
The symmetric set, $\tilde{X}$, of $X$ is defined as $\tilde{X} = \{-x | x \in X\}$.

Translated set
The translate of $X$ by $b$ is given by $\{z \in \mathcal{E} | z = x + b, x \in X\}$. 
Basic morphological operators I

Erosion

Definition (Morphological erosion)

\[ X \ominus B = \{ z \in \mathcal{E} | B_z \subseteq X \} \]

The following algebraic expression is equivalent to the previous definition:

Definition (Alternative definition for the morphological erosion)

\[ X \ominus B = \bigcap_{b \in B} X_{-b} \]

\[ (180) \]

\( B \) is named “structuring element”.

\[ (179) \]
Erosion with a disk

Figure: Erosion of $X$ with a disk $B$. The origin of the structuring element is drawn at the center of the disk (with a black dot).
Definition (Dilation)

From an algebraic perspective, the *dilation* (*dilatation* in French!), is the union of translated version of $X$:

$$X \oplus B = \bigcup_{b \in B} X_{b} = \bigcup_{x \in X} B_{x} = \{x + b | x \in X, b \in B\}. \quad (181)$$
Properties of the erosion and the dilation

Duality
Erosion and dilation are two dual operators with respect to complementation:

\[ X \ominus \tilde{B} = (X^c \oplus B)^c \]  
(182)

\[ X \ominus B = (X^c \oplus \tilde{B})^c \]  
(183)

Erosion and dilation obey the principles of “ideal” morphological operators:

1. erosion and dilation are invariant to translations: \( X_z \ominus B = (X \ominus B)_z \). Likewise, \( X_z \oplus B = (X \oplus B)_z \);

2. erosion and dilation are compatible with scaling: \( \lambda X \ominus \lambda B = \lambda(X \ominus B) \) and \( \lambda X \oplus \lambda B = \lambda(X \oplus B) \);

3. erosion and dilation are local operators (if \( B \) is bounded);

4. it can be shown that erosion and dilation are continuous transforms.

Algebraic properties

- erosion and dilation are \textit{increasing} operators: if \( X \subseteq Y \), then \( (X \ominus B) \subseteq (Y \ominus B) \) and \( (X \oplus B) \subseteq (Y \oplus B) \);

- if the structuring element \textit{contains} the origin, then the erosion is \textit{anti-extensive} and the dilation is \textit{extensive}, that is \( X \ominus B \subseteq X \) and \( X \subseteq X \oplus B \).
Definition (Opening)

The opening results from cascading an erosion and a dilation with the same structuring element:

\[ X \circ B = (X \ominus B) \oplus B. \]  

(184)

Interpretation of openings (alternative definition)

The interpretation of the opening operator (which can be seen as an alternative definition) is based on

\[ X \circ B = \bigcup \{B_z \mid z \in \mathcal{E} \text{ and } B_z \subseteq X\}. \]  

(185)

In other words, the opening of a set by structuring element \( B \) is the set of all the elements of \( X \) that are covered by a translated copy of \( B \) when it moves inside of \( X \).
### Definition (Closing)

A **closing** is obtained by cascading a dilation and an erosion with a unique structuring element:

\[ X \bullet B = (X \oplus B) \ominus B. \]  

(186)

Opening and closing are dual operators with respect to set complementation: indeed,

\[ (X \circ B)^c = X^c \bullet \breve{B} \]  

(187)

and

\[ (X \bullet B)^c = X^c \circ \breve{B}. \]  

(188)

---

### Opening and closing of \( X \) with a disk \( B \)

![Diagram of opening and closing of X with a disk B](image-url)
Opening and closing properties

By construction, the opening and closing follow the “ideal” principles of morphological operators.

The most important algebraic properties of $X \circ B$ and $X \bullet B$ are

1. opening and closing are increasing. If $X \subseteq Y$, then

   $$(X \circ B) \subseteq (Y \circ B) \text{ and } (X \bullet B) \subseteq (Y \bullet B)$$  \hspace{1cm} (189)

2. opening is anti-extensive, and closing is extensive

   $$X \circ B \subseteq X, \quad X \subseteq X \bullet B$$  \hspace{1cm} (190)

3. opening and closing are idempotent operators (projective operators). This means that

   $$(A \circ B) \circ B = A \circ B \text{ and } (A \bullet B) \bullet B = A \bullet B$$  \hspace{1cm} (191)

General properties I

- Dilation is commutative and associative

  $$X \oplus B = B \oplus X$$  \hspace{1cm} (192)

  $$ (X \oplus Y) \oplus C = X \oplus (Y \oplus C)$$  \hspace{1cm} (193)

- Dilation distributes the union

  $$ (\bigcup_j X_j) \oplus B = \bigcup_j (X_j \oplus B)$$  \hspace{1cm} (194)

- The erosion distributes the intersection

  $$ (\bigcap_j X_j) \ominus B = \bigcap_j (X_j \ominus B)$$  \hspace{1cm} (195)
General properties II

- **Chain rule** (≡ cascading rule):

\[ X \ominus (B \oplus C) = (X \ominus B) \ominus C \]  \hspace{1cm} (196)

- The opening and closing are not related to the exact location of the origin (so they do not depend on the location of the origin when defining \( B \)). Let \( z \in \mathcal{E} \):

\[ X \circ B_z = X \circ B \]  \hspace{1cm} (197)

\[ X \bullet B_z = X \bullet B \]  \hspace{1cm} (198)

---

**A practical problem: dealing with borders I**

*Figure:* Two possible physical assumptions for borders.
A practical problem: dealing with borders II

Some pixels added to $X$

$X \ominus B$

Physical assumption 1

Physical assumption 2

Figure: Comparison of the effects of two physical assumptions on the computation of the erosion of $X$.

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
   - Morphology on sets/objects
   - Geodesic operators
   - Grayscale morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
The Hit or Miss transform is defined such as
\[ X \uparrow (B, C) = \{ x | B_x \subseteq X, C_x \subseteq X^c \} \]  \hspace{1cm} (199)

If \( C = \emptyset \) the transform reduces to an erosion of \( X \) by \( B \).

Geodesy and reconstruction I

Geodesic dilation
A geodesic dilation is always based on two sets (images).

Definition

The geodesic dilation of size 1 of \( X \) conditionally to \( Y \), denoted \( D_Y^{(1)}(X) \), is defined as the intersection of the dilation of \( X \) and \( Y \):

\[ \forall X \subseteq Y, \quad D_Y^{(1)}(X) = (X \oplus B) \cap Y \]  \hspace{1cm} (200)

where \( B \) is usually chosen according to the frame connectivity (a 3 \( \times \) 3 square for a 8-connected grid).
**Definition**

The *reconstruction* of $X$ conditionally to $Y$ is the geodesic dilation of $X$ until idempotence. Let $i$ be the iteration during which idempotence is reached, then the reconstruction of $X$ is given by

$$R_Y(X) = D_Y^{(i)}(X) \text{ with } D_Y^{(i+1)}(X) = D_Y^{(i)}(X).$$  \hfill (202)

---

**Figure:** Blob extraction by marking and reconstruction.

(a) Blobs \hspace{1cm} (b) Marking blobs \hspace{1cm} (c) Reconstructed blobs
Grayscale morphology I

Notion of a function

Let \( \mathcal{G} \) be the range of possible grayscale values. An image is represented by a function \( f : \mathcal{E} \rightarrow \mathcal{G} \), which projects a location of a value of \( \mathcal{G} \). In practice, an image is not defined over the entire space \( \mathcal{E} \), but on a limited portion of it, a compact \( D \).

We need to define an order between functions.

**Definition (Partial ordering between functions)**

Let \( f \) and \( g \) be functions. \( f \) is inferior to \( g \),

\[
f \leq g \text{ if } f(x) \leq g(x), \; \forall x \in \mathcal{E}
\]  

(203)
Grayscale morphology II

Definition (Infimum and supremum)
Let $f_i$ be a family of functions, $i \in I$. The \textit{infimum} (respectively the \textit{supremum}) of this family, denoted $\wedge_{i \in I} f_i$ (resp. $\vee_{i \in I} f_i$) is the largest lower bound (resp. the lowest upper bound).

In the practical case of a finite family $I$, the supremum and the infimum correspond to the maximum and the minimum respectively. In that case,

$$\forall x \in \mathcal{E}, \begin{cases} (f \vee g)(x) = \max(f(x), g(x)) \\ (f \wedge g)(x) = \min(f(x), g(x)) \end{cases}$$ (204)

Definition (Translate of a function)
The \textit{translate} of a function $f$ by $b$, denoted by $f_b$, is defined as

$$\forall x \in \mathcal{E}, \quad f_b(x) = f(x - b).$$ (205)

Additional definitions related to operators I

Definition (Idempotence)
An operator $\psi$ is \textit{idempotent} if, for each function, a further application of it does not change the final result. That is, if

$$\forall f, \quad \psi(\psi(f)) = \psi(f)$$ (206)

Definition (Extensivity)
An operator is \textit{extensive} if the result of applying the operator is larger than the original function

$$\forall f, \quad f \leq \psi(f)$$ (207)
**Definition (Anti-extensivity)**

An operator is *anti-extensive* if the result of applying the operator is lower than the original function

\[ \forall f, \ f \geq \psi(f) \]  

(208)

**Definition (Increasingness)**

An increasing operator is such that it does not modify the ordering between functions:

\[ \forall f, g, \ f \leq g \Rightarrow \psi(f) \leq \psi(g) \]  

(209)

By extension, an operator \( \psi_1 \) is lower than an operator \( \psi_2 \) if, for every function \( f \), \( \psi_1(f) \) is lower to \( \psi_2(f) \):

\[ \psi_1 \leq \psi_2 \iff \forall f, \ \psi_1(f) \leq \psi_2(f) \]  

(210)

**Erosion and dilation**

**Definition (Grayscale dilation and erosion)**

Let \( B \) be the domain of definition of a structuring element. The *grayscale dilation* and *erosion* (with a flat structuring element) are defined, respectively as,

\[ f \oplus B = \bigvee_{b \in B} f_b(x) \]  

(211)

\[ f \ominus B = \bigwedge_{b \in B} f_{-b}(x) \]  

(212)
Numerical example ($B = \{-1, 0, 1\}$)

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>25</td>
<td>27</td>
<td>30</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>$f(x - 1)$</td>
<td>25</td>
<td>27</td>
<td>30</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>$f(x)$</td>
<td>25</td>
<td>27</td>
<td>30</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>$f(x + 1)$</td>
<td>27</td>
<td>30</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>18</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>$f \ominus B(x) = \min$</td>
<td>25</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>15</td>
<td>22</td>
<td>18</td>
<td>18</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>$f \oplus B(x) = \max$</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>24</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

Typical questions:

▶ best algorithms? (note that there is some redundancy between neighboring pixels)
▶ how do proceed around borders?

 Algorithms

▶ Based on the decomposition of the structuring element:
  
  * $f \ominus (H \oplus V) = (f \ominus H) \ominus V$
  * $f \ominus (B \oplus B) = (f \ominus B) \ominus \partial(B)$

▶ Appropriate structure for storing and propagating the local min and max
  
  * queues
  * histogram
Illustration I

\[ f \ominus B \]

Figure: Erosion of a function.

Illustration II

Figure: Erosions with squares of increasing sizes.
Figure: Dilations with squares of increasing sizes.

The opening $f \circ B$ is obtained by cascading an erosion followed by a dilation. The closing $f \bullet B$ is the result of a dilation followed by an erosion.

Definition (Morphological opening and closing)

$$f \circ B = (f \ominus B) \oplus B$$  \hspace{1cm} (213)

$$f \bullet B = (f \oplus B) \ominus B$$  \hspace{1cm} (214)
Properties of grayscale morphological operators

Erosion and dilation are increasing operators

\[ f \leq g \Rightarrow \left\{ \begin{array}{l} f \ominus B \leq g \ominus B \\ f \oplus B \leq g \oplus B \end{array} \right. \] \hspace{1cm} (215)

Erosion distributes the infimum and dilation distributes the supremum

\[ (f \land g) \ominus B = (f \ominus B) \land (g \ominus B) \] \hspace{1cm} (216)
\[ (f \lor g) \oplus B = (f \oplus B) \lor (g \oplus B) \] \hspace{1cm} (217)

Opening and closing are idempotent operators

\[ (f \circ B) \circ B = f \circ B \] \hspace{1cm} (218)
\[ (f \bullet B) \bullet B = f \bullet B \] \hspace{1cm} (219)

Opening and closing are anti-extensive and extensive operators respectively

\[ f \circ B \leq f \] \hspace{1cm} (220)
\[ f \leq f \bullet B \] \hspace{1cm} (221)
Definition

The **reconstruction** of $f$, conditionally to $g$, is the geodesic dilation of $f$ until idempotence is reached. Let $i$, be the index at which idempotence is reached, the reconstruction of $f$ is then defined as

$$R_g(f) = D_{g}^{(i)}(f) \text{ with } D_{g}^{(i+1)}(f) = D_{g}^{(i)}(f).$$

(222)

**Figure:** Original image, eroded image, and several successive geodesic dilations.
Reconstruction of grayscale images III

**Figure:** Original image, eroded image, reconstructed image starting from the eroded image, and difference image (reverse video).

Reconstruction of grayscale images IV

**Figure:** Original image, dilated image, reconstructed image starting from the dilated image (dual reconstruction), and difference image (reverse video).
Non-linear filtering

- Rank filters
  - Median
- Morphological filters
  - Algebraic definition
  - How to build a filter?
  - Examples of filters
    - Alternate sequential filters
    - Morphological filter
Introduction to rank filters

Let $k \in \mathbb{N}$ be a threshold.

**Definition (Rank filter)**

The operator or $k$-order rank filter, denoted as $\rho_{B,k}(f)(x)$, defined with respect to the $B$ structuring element, is

$$\rho_{B,k}(f)(x) = \sqrt{\{t \in \mathcal{G} | \sum_{b \in B} [f(x + b) \geq t] \geq k\}}$$

(223)

The simplest interpretation is that $\rho_{B,k}(f)(x)$ is the $k$-est value when all the $f(x + b)$ values are ranked in decreasing order.

Rank filters are ordered. Let $\sharp(B)$, be the surface of $B$, then

$$\rho_{B,\sharp(B)}(f)(x) \leq \rho_{B,\sharp(B)-1}(f)(x) \leq \ldots \leq \rho_{B,1}(f)(x)$$

(224)
If $n$ is odd, the $k = \frac{1}{2}(\#(B) + 1)$ choice leads to the definition of a self-dual operator, that is a filter that produces the same result as if applied on the dual function. This operator, denoted $med_B$, is the median filter.

<table>
<thead>
<tr>
<th>$f(x)$</th>
<th>25</th>
<th>27</th>
<th>30</th>
<th>24</th>
<th>17</th>
<th>15</th>
<th>22</th>
<th>23</th>
<th>25</th>
<th>18</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>24</td>
<td>17</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>22</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>$med_B$</td>
<td>25</td>
<td>27</td>
<td>27</td>
<td>24</td>
<td>17</td>
<td>17</td>
<td>22</td>
<td>23</td>
<td>23</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>24</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

$m_f \ominus B(x) = \min$  
25 24 17 15 15 15 22 18 18

$m_f \oplus B(x) = \max$  
30 30 30 24 22 23 25 25 25

---

(a) Original image $f +$ noise  
(b) Opening with a $5 \times 5$ square

(c) Low-pass Butterworth ($f_c = 50$)  
(d) Median with a $5 \times 5$ square
Effect of the size of the median filter

Notes about the implementation

The median filter is not idempotent. Successive applications can result in oscillations (theoretically if the domain of the function is infinite)

\[
\cdots \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \cdots \\
\downarrow
\cdots \quad \circ \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \cdots \\
\downarrow
\cdots \quad \circ \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \bullet \quad \circ \quad \cdots
\]

*Figure:* Repeated application of a median filter.

Also,

\[
\text{med}_{5 \times 5}(f) \neq \text{med}_{1 \times 5}(\text{med}_{5 \times 1}(f))
\]  

(225)

but it’s an acceptable *approximation.*
Morphological filters

Definition (Algebraic filter)
By definition, a filter is an algebraic filter if and only if the operator is increasing and idempotent:

\[
\psi \text{ is an algebraic filter } \iff \forall f, g \left\{ f \leq g \Rightarrow \psi(f) \leq \psi(g) \quad (226) \right. \\
\left. \psi(\psi(f)) = \psi(f) \right\}
\]

Definition (Algebraic opening)
An algebraic opening is an operator that is increasing, idempotent, and anti-extensive. Formally,

\[
\forall f, g, f \leq g \Rightarrow \psi(f) \leq \psi(g) \quad (227) \\
\forall f, \psi(\psi(f)) = \psi(f) \quad (228) \\
\forall f, \psi(f) \leq f \quad (229)
\]

An algebraic closing is defined similarly, except that the operator is extensive.

How to build a filter? I
By combining know filters!

Figure: The composition of two openings is not an opening.
New filters can be built starting from openings, denoted $\alpha_i$, and closings, denoted $\phi_i$. The rules to follow are:

1. the supremum of openings is an opening: $(\bigvee_i \alpha_i)$ is an opening;
2. the infimum of closings is a closing: $(\bigwedge_i \phi_i)$ is a closing.

Supremum of two openings

Openings: $\gamma_{mH \oplus nV}(f)$, $\gamma_{mH}(f)$, $\gamma_{nV}(f)$, and $\gamma_{mH}(f) \lor \gamma_{nV}(f)$
Let ψ₁ and ψ₂ be two filters such that ψ₁ ≥ I ≥ ψ₂ (for example, ψ₁ is a closing and ψ₂ an opening).

**Theorem (Structural theorem)**

Let ψ₁ and ψ₂ be two filters such that ψ₁ ≥ I ≥ ψ₂, then

\[
\psi_1 \geq \psi_1 \psi_2 \psi_1 \geq (\psi_2 \psi_1 \lor \psi_1 \psi_2) \geq (\psi_2 \psi_1 \land \psi_1 \psi_2) \geq \psi_2 \psi_1 \psi_2 \geq \psi_2
\]

(230)

\[
\psi_1 \psi_2, \psi_2 \psi_1, \psi_1 \psi_2 \psi_1, \psi_2 \psi_1 \psi_2 \text{ are all filters}
\]

(231)

*Note that there is no ordering between ψ₁ψ₂ and ψ₂ψ₁.*

---

**Alternate Sequential Filters (ASF)**

Let γᵢ (φᵢ) be an opening (resp. a closing) of size i and I be the identity operator (i.e. \(I(f) = f\)). We assume that there is the following order:

\[
\forall i, j \in \mathbb{N}, \quad i \leq j, \quad \gamma_j \leq \gamma_i \leq I \leq \phi_i \leq \phi_j
\]

(232)

For each index i, we define these operators:

\[
m_i = \gamma_i \phi_i, \quad r_i = \phi_i \gamma_i \phi_i, \\
n_i = \phi_i \gamma_i, \quad s_i = \gamma_i \phi_i \gamma_i.
\]
Examples of filters II

Definition (Alternate Sequential Filters (ASF))

For each index $i \in \mathbb{N}$, the following operators are the alternate sequential filters of index $i$

$$
M_i = m_i m_{i-1} \ldots m_2 m_1 \quad R_i = r_i r_{i-1} \ldots r_2 r_1 \quad (233)
$$

$$
N_i = n_i n_{i-1} \ldots n_2 n_1 \quad S_i = s_i s_{i-1} \ldots s_2 s_1 \quad (234)
$$

Theorem (Absorption law)

$$
i \leq j \Rightarrow M_j M_i = M_j \quad \text{but} \quad M_i M_j \leq M_j \quad (235)
$$

Examples of filters III

![Images](a) Image $f$ (b) $M_1(f)$ (c) $M_2(f)$ (d) $M_3(f)$

(e) $5 \times 5$ median (f) $N_1(f)$ (g) $N_2(f)$ (h) $N_3(f)$

Figure: Use of alternate sequential filters to remove some noise.
The *morphological center* is a typical example of *toggle mapping*.

**Definition (Morphological center)**

Let \( \psi_i \) be a family of operators. The *morphological center* \( \beta \) of a function \( f \) with respect to the \( \psi_i \) family is defined, for each location \( x \) of the domain of \( f \) as follows:

\[
\beta(f)(x) = (f(x) \lor (\bigwedge_i \psi_i(x))) \land (\bigvee_i \psi_i(x))
\]  

(236)
Motion analysis and background subtraction

Detection of motion

There are basically two “pure” approaches for motion analysis/detection in a video sequence:

1. **Motion analysis by tracking** (= motion estimation based techniques):
   - detects some particular points in a video frame.
   - find the corresponding points/objects in the next frame.
   - based on a model, interpret the trajectories of the points/objects (usually at the object level).

2. **Motion analysis by background subtraction**:
   - build a reference frame or model with no foreground in it.
   - compare a next frame to the reference.
   - update the reference.
Motion analysis by tracking: principles

There are several techniques but, usually, they involve the following steps:

1. detect features in successive frames.
2. make some correspondences between the features detected in consecutive frames.
3. based on a model, regroup some features to facilitate the tracking of objects.
Feature detection

Some known feature detectors:
- Harris’s corner detector
- Scale Invariant Feature Transform (SIFT)
- Speeded Up Robust Features from an image (SURF)
- Features from Accelerated Segment Test (FAST)
- ...

Original image  Features detected by SURF

Feature correspondence
Difficulties for feature correspondence

Typical questions/difficulties for tracking approaches (targeting motion estimation):

▶ How to filter the features? (remove some useless features)
▶ How do we regroup features?
  - we need a model. But this introduces a bias towards the model.
▶ How do we ensure continuity over time?
▶ What happens when occlusions occur?
▶ How to solve ambiguities (one feature in one image is identical to several in the next frame)?
▶ …

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
  - Motion analysis by tracking
  - Background subtraction
  - Technical details of background subtraction
Motion analysis and background subtraction

Background subtraction

Motion analysis by background subtraction

Original image

Features detected by ViBe

Figure: Segmentation (pixels that are “in motion”) by background subtraction.

Example of a pixel-based background subtraction technique

- Any application with moving objects
- Video-surveillance
**Classical processing chain**

(pre-processing) of an input frame

⇓

*segmentation* map after background processing

⇓

(post-processing) of the segmentation map

Typical post-processing operations are:

▶ morphological filtering (cleaning): erosion, dilation, opening, area opening
▶ median filtering
▶ analysis of connected components
▶ shadow removal
▶ semantic post-processing
▶ ...

Marc Van Droogenbroeck  
Template Matching & Image Registration  
Academic year: 2019-2020  
Motion analysis and background subtraction  
Background subtraction
Challenges

- **Input related issues**
  - lighting/illumination changes
    - slow (day/night cycles)
    - fast (light switch, clouds in the sky, ...)
  - unwanted motions
    - camera shaking (wind)
    - in the background (tree leaves, waving grass, water)
  - appearance changes of foreground objects (reflections), shadows, camouflage, ...

- **Implementation related issues**
  - robustness
  - real time

Applications

- **Motion detection** (you don’t need a precise segmentation map):
  - raise an alarm
  - start/stop recording

- **Foreground segmentation**:
  - counting
  - traffic analysis
  - human motion interpretation
  - sport analysis
  - coding
  - aerial imaging
  - detection of exoplanets
  - ...
Background subtraction

Common tasks:
- background subtraction $\Rightarrow$ segmentation
- change detection $\Rightarrow$ generate alarms (and reduce rate of false alarms!)

Different contexts:
- pixel based
- region based
- tracking

Plethora of methods (2016):
- 'motion detection' on IEEE Xplore: 16,471 papers
- 'background subtraction video' on IEEE Xplore: 960 papers
Objective of pixel-based background subtraction techniques

separate the *foreground* (pixels “in motion”) from the *background* (“static” pixels).

Implementation perspective

A video sequence is like a data cube whose dimension is only fixed in 2 dimensions.

- Cube extends with time.
- But the use of “memory” should be kept constant.

Challenges:

1. need to find a way to **accumulate knowledge of increasing size** inside of a constant sized memory block
2. this knowledge should be **updated** regularly to deal with changes

Steps

[**Initialization**] build a *reference frame* or a *statistical model* for the background.

[**Subtraction or segmentation**] *compare* the current frame to the reference frame or model, and “*subtract*” the frame to get a binary image indicating pixels who have changed.

[**Updating**] *update* the reference frame or model.

When we develop a technique, we have to detail these three steps!
**Principles**

- Major assumption: fixed camera
- Otherwise,
  - we have to compensate for the camera motion
  - what for new areas (for which we have no past information)?

**Reference frame or a model?**

*Figure:* Building a reference frame or a model.
Naive approach (static background)

Foreground is detected, pixel by pixel, as the difference between the current frame and a static reference image (background):

$$|I_t - B| > \text{threshold} \quad (237)$$

where

- $I_t$ is the current pixel value (at time $t$),
- $B$ is the reference background value for that pixel.

Problems:

- How do we choose the reference image?
- The reference image should change over time.
- What’s the best threshold?
Simple techniques

- One "good" image is chosen (usually a frame empty of foreground objects).
- Exponentially updated reference image (blending)

\[ B_t = \alpha I_t + (1 - \alpha)B_{t-1} \]  \hspace{1cm} (238)

Typical value for \( \alpha \): 0.05 (in that case, \( B_t = 0.05 I_t + 0.95 B_{t-1} \))
- Median of the last \( N \) frames (for each pixel separately).

Intermediate summary: remember the components

- **Initialization**
  - \( \Rightarrow \) fill in the first reference image or initialize the model
  - **difficulty**: problem with the presence of *ghosts*.

- **Segmentation**
  - \( \Rightarrow \) rules + parameters …
  - **difficulties**:
    - homogeneous rule for the whole image?
    - how to set the values of the parameters?

- **Updating**
  - \( \Rightarrow \) we need strategies
  - **difficulty**: how to adapt the scene dynamics?
There are two main strategies for the updating

**Blind update** (the most common)

update the reference image or the model for all pixels, regardless of the segmentation result (that is if the pixel is a foreground or a background pixel)

▶ does it make sense?

**Conservative update**

only update when the pixel belongs to the background

▶ what if the segmentation was wrong (deadlock)?
The background is modeled as a probability density function to be estimated

- One Gaussian distribution per pixel
- *Mixture of Gaussians for each pixel*
- *Kernel based estimation of the probability density function*

**One Gaussian per pixel**

For each pixel, the probability density function of observed values is modeled by a **single** Gaussian.

Once the model is built (here it means that we need to estimate the mean and variance), we evaluate the distance to the mean. If

\[ |l_t - \mu| \leq \text{parameter} \times \sigma \quad (239) \]

then the pixel belongs to the background.
**Mixture of Gaussians Model**

**Motivation:** the probability density function of the background is multi-modal

For each pixel:

\[ P(X) = \sum_{i=1}^{N} \alpha_i N(\mu_i, \sigma_i) \]  

Typical values for \( n \): 3 or 5

**Fundamental assumptions**

- The background has a low variance.
- The background is more frequently visible than the foreground.
Kernel Density Estimation (KDE) methods

For each pixel:

\[
P(X) = \sum_{i=1}^{N} \alpha_i K_\sigma(X - X_i)
\]  

(241)

where \(\{X_i\}_{i=1,...,N}\) are the \(N\) last values observed (samples) for that pixel, and \(K_\sigma()\) is a kernel probability function centered at \(X_i\).

Decision rule

The pixel belongs to the background if \(P(X) > \text{threshold}\).

[Elgammal, 2000] [Zivkovic, 2006]

Parameters of KDE techniques

Typical values \(P(X) = \sum_{i=1}^{N} \alpha_i K_\sigma(X - X_i)\) with

- Number of samples: \(N = 100\)
- Weight: \(\alpha_i = \alpha = \frac{1}{N}\)
- Spreading factor: \(\sigma = \text{Variance}(X_i)\)
- Probability density function chosen to be Gaussian: 
  \(K_\sigma(X - X_i) = N(X_i, \sigma^2)\)
GMM techniques vs KDE techniques

Other advanced techniques

- Codebook
- Principal Components Analysis (PCA)
- Relaxation techniques (mainly post-processing of binary segmentation maps)
- Local Binary Patterns
- ViBe: no model, all based on samples
- Deep learning based methods
More details about ViBe: design rules

The design of ViBe was motivated by:

1. *information at the pixel level* should be preferred to aggregated features (for speed and efficiency).
2. *collect samples*, rather than build a statistical model:
   1. sample values have been observed in the past.
   2. no bias towards a model.
3. *minimize the number* of collected samples.
4. *no “planned obsolescence”* notion for samples.
   1. don’t use the commonly (and exclusively) adopted substitution rule that consists to replace oldest samples first or to reduce their associated weight.
   2. all samples are equally valid.
5. foresee a mechanism to ensure *spatial consistency*.
6. keep the *decision* process *simple*.

Requirements while designing ViBe

1. No pre- or post-processing!
2. Be real-time.
3. A unique set of parameters.
4. Working as soon as you get the second frame.
Keys of a background subtraction technique

1. **Model.** What is a good model for the background?
2. How to **classify** pixels in the background/foreground? Need for a classification criterion.
3. How to **update** the model?
4. **Initialization**?

---

**Model**

- Model each background pixel with a *set of samples* (non-parametric method), instead of with an explicit pixel model. Each background pixel $x$ is modeled by a collection of $N$ background sample values: $M(x) = \{v_1, v_2, \ldots, v_N\}$

  - Avoid the difficult task of estimating the probability density function.
  - No statistical notion. Assume a binary image, what is the meaning of the mean? It is even a value that might never be observed ...
  - How could a model consider values of its immediate neighborhood?
Pixel classification

- A pixel belongs to the background if there are at least two matches:
  \[ \# \{ S_R(v(x)) \cap \{ v_1, v_2, ..., v_N \} \} \geq \#_{\text{min}} \]
- \( N = 20 \) and \( \#_{\text{min}} = 2 \)

![Pixel classification diagram](image)

Fig. 1. Comparison of a pixel value with a set of samples in a two-dimensional Euclidean color space \((C_1, C_2)\). To classify \( v(x) \), we count the number of samples of \( M(v) \) intersecting the sphere of radius \( R \) centered on \( v(x) \).

Updating the model over time

**Principles:**
- **Conservative update:** update the model only if a pixel value is declared as a background. If it is foreground, do not update at all.

**Mechanisms:**
- The background model of a pixel is **updated randomly**! One sample in the model is chosen randomly and replaced. The lifetime of the replaced value is ignored.
- Random **time subsampling**. Not every model is updated. Which one is decided randomly (typically 1 out of 16, on average).
- Randomly select a neighbor and modify the model of it with the value of the local pixel and according to the 1 and 2 mechanisms (**spatial diffusion**).
Model Initialization

From a single frame, populate the pixel models with values found in the spatial neighborhood of each pixel.

- Values randomly taken in their neighborhood
- Values chosen in the close 8-connected neighborhood

Results

(a) Input image  (b) Ground-truth

(c) ViBe (RGB)  
(d) ViBe (gray)  
(e) Bayesian histogram

(f) Codebook  
(g) EGMM [Zivkovic]  
(h) GMM [Li et al.]

(i) Gaussian model  
(j) 1st order filter  
(k) Sigma-Delta Z
Traditionally

We choose the best feature set and optimize the background subtraction algorithm. But there are many (sub-optimal) possibilities.


What if we automate the search (quest) for the best features?

What’s next? I

More and more machine learning for background subtraction.

Figure: Deep learning for extracting the background in video scenes (M. Braham and M. Van Droogenbroeck. Deep Background Subtraction with Scene-Specific Convolutional Neural Networks. In IEEE IWSSIP, May 2016).
**What’s next? II**

<table>
<thead>
<tr>
<th>Method</th>
<th>$F_{overall}$</th>
<th>$F_{Baseline}$</th>
<th>$F_{Jitter}$</th>
<th>$F_{Shadows}$</th>
<th>$F_{LowFramerate}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvNet-GT</td>
<td>0.9046</td>
<td>0.9813</td>
<td>0.9020</td>
<td>0.9454</td>
<td>0.9612</td>
</tr>
<tr>
<td>IUTIS-5</td>
<td>0.8093</td>
<td>0.9683</td>
<td>0.8022</td>
<td>0.8807</td>
<td>0.8515</td>
</tr>
<tr>
<td>SuBSENSE</td>
<td>0.8018</td>
<td>0.9603</td>
<td>0.7675</td>
<td>0.8732</td>
<td>0.8441</td>
</tr>
<tr>
<td>PAWCS</td>
<td>0.7984</td>
<td>0.9500</td>
<td>0.8473</td>
<td>0.8750</td>
<td>0.8988</td>
</tr>
<tr>
<td>PSP-MRF</td>
<td>0.7927</td>
<td>0.9566</td>
<td>0.7690</td>
<td>0.8735</td>
<td>0.8109</td>
</tr>
<tr>
<td>ConvNet-IUTIS</td>
<td>0.7897</td>
<td>0.9647</td>
<td>0.8013</td>
<td>0.8590</td>
<td>0.8273</td>
</tr>
<tr>
<td>EFIC</td>
<td>0.7883</td>
<td>0.9231</td>
<td>0.8050</td>
<td>0.8270</td>
<td>0.9336</td>
</tr>
<tr>
<td>Spectral-360</td>
<td>0.7867</td>
<td>0.9477</td>
<td>0.7511</td>
<td>0.7156</td>
<td>0.8797</td>
</tr>
<tr>
<td>SC_SOBS</td>
<td>0.7450</td>
<td>0.9491</td>
<td>0.7073</td>
<td>0.8602</td>
<td>0.7985</td>
</tr>
<tr>
<td>GMM</td>
<td>0.7444</td>
<td>0.9478</td>
<td>0.6103</td>
<td>0.8396</td>
<td>0.8182</td>
</tr>
<tr>
<td>GraphCut</td>
<td>0.7394</td>
<td>0.9304</td>
<td>0.5183</td>
<td>0.7543</td>
<td>0.8208</td>
</tr>
</tbody>
</table>

**Table:** Overall and per-category $F$ scores for different methods.

---

**Outline**

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. **Motion analysis and background subtraction**
   - Motion analysis by tracking
   - Background subtraction
   - Technical details of background subtraction
Evaluation

Elements

- Metric
- Ground truth data
- Interpretation

Evaluation tasks

- Determine the intrinsic performance of a method
- Compare methods
- Rank methods

Outline

1. Fundamentals of 2D imaging
2. 3D vision
3. Linear filtering
4. Edge detection
5. Introduction to machine learning
6. Performance analysis
7. Segmentation
8. Object description and analysis
9. Deep learning for visual understanding
10. Template Matching & Image Registration
11. Mathematical morphology
12. Non-linear filtering
13. Motion analysis and background subtraction
14. Feature detection and tracking
   - Introduction
   - Feature detection
Introduction

Features are edges, corners, etc.
Most of them are based on derivatives.

The choice of features is application-dependent!

For tracking:
  - We need matching measures and metrics to find the best matching locations between frames.
  - Features need to be “strong”: resilient to illumination changes, to local deformations, etc.
**Feature detection and tracking**

**Introduction**

**Historical example: the Laplacian of Gaussians I**

**Principle:**
- first apply a Gaussian filter (with the purpose of reducing noise).
- compute the Laplacian of the filtered image.
- detect the locations of zero-crossings in the image

**Parameters:**
- variance of the Gaussian filter

**Result:**
- collection of images with features detected at different scales.
- not suitable for tracking.

---

**Historical example: the Laplacian of Gaussians II**

---

Marc Van Droogenbroeck  
Template Matching & Image Registration  
Academic year: 2019-2020  
579 / 600
Feature detection and tracking

Introduction

Feature-based matching

Steps:
1. Select a set of robust features in the reference frame (or sub-frame/Region of Interest [ROI])
2. Compute the same set of features in the target image
3. Establish the relationship between the two sets of features. Many strategies are possible:
   - Global optimization
   - Local optimization
   - RANSAC algorithm

Feature detectors I

Classification [source: wikipedia]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Edge</th>
<th>Corner</th>
<th>Blob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sobel</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canny</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harris</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Laplacian of Gaussians</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>FAST</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MSER</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Feature detection and tracking

Introduction

Feature detectors II

MSER: 5 matches
ASIFT: 202 matches

Outline

1 Fundamentals of 2D imaging
2 3D vision
3 Linear filtering
4 Edge detection
5 Introduction to machine learning
6 Performance analysis
7 Segmentation
8 Object description and analysis
9 Deep learning for visual understanding
10 Template Matching & Image Registration
11 Mathematical morphology
12 Non-linear filtering
13 Motion analysis and background subtraction
14 Feature detection and tracking
   - Introduction
   - Feature detection
Let $p$ be a pixel with the $(x, y)$ coordinates. We compute

$$M = \sigma^2_D g(\sigma_I) \otimes \begin{bmatrix}
    l^2_x(p, \sigma_D) & l_x(p, \sigma_D)l_y(p, \sigma_D) \\
    l_x(p, \sigma_D)l_y(p, \sigma_D) & l^2_y(p, \sigma_D)
\end{bmatrix} \quad (242)$$

where

$$l_x(p, \sigma_D) = \frac{\partial}{\partial x} g(\sigma_D) \otimes I(p) \quad (243)$$

$$g(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (244)$$

**Definition (Cornerness)**

*Cornerness* is defined as

$$\det(M) - \lambda \text{trace}^2(M) \quad (245)$$

where $\lambda$ is a tunable sensitivity parameter.
Feature detection and tracking

Other feature detectors I

The calculation of features combines several steps and usually includes a scale space and derivates.

Some variants:

- MSER: Maximally Stable Extremal Region extractor (2002)
- SURF: Speeded Up Robust Features (2006)
- FAST (2006)
- BRISK: Binary Robust Invariant Scalable Keypoints (2011)
- ORB: Oriented BRIEF (2011)
- FREAK: Fast Retina Keypoint (2012)
- ...

Don’t forget that these methods have some parameters.

Other feature detectors II
### Approach

- There is an **object**, characterized by some quantities put in a vector $X$, that cannot be observed directly.
- We assume that we can trust a **dynamic model** of the state that allows us to **predict** how it evolves over time.
- Periodically, we observe the scene and produce a **measure** $Y$, that is a function of the state of the system.
- With this observation, we **correct** our estimate of the system state.
The tracking loop in practice (for one object)

1. Wait for a new image.
2. **Predict** the current state of the tracked object.
3. Use that prediction to delineate a region of interest.
4. Extract the tracked object in the new image.
5. Derive the object state from the previous step and use it to **correct** the estimate of the object state.

Independence assumptions

- The system state and the observation are considered as *random* vectors.
- We limit the model to a Markov model. If $X_i$ defines the object state in frame $i$, then

  $$p(X_i|X_{i-1}, X_{i-2}, \ldots) = p(X_i|X_{i-1})$$

  (246)

  In other words, only the immediate past matters.
- The measurement only depends on the current state (no memory):

  $$p(Y_i|X_i, \ldots) = p(Y_i|X_i)$$

  (247)
Assumptions

1. The dynamic model is linear.
2. Errors in the modeling of state variables are considered as additive Gaussian noise.
3. The measurement process is linear.
4. Noise on the measurement is characterized by an additive Gaussian noise.

Prediction and correction

Prediction
Multiply the state by a matrix $D$, and add Gaussian noise:

$$x_t \sim p(X_t|X_{t-\Delta t}) = \mathcal{N}(Dtx_{t-\Delta t}; \Sigma_D)$$  \hspace{1cm} (248)

where $\Sigma_D$ is the covariance matrix.

Measurement
Multiply the state by a matrix $M$, and add Gaussian noise:

$$y_t \sim p(Y_t|X_t) = \mathcal{N}(Mtx_t; \Sigma_M)$$  \hspace{1cm} (249)
Example: first order model for the displacement of an object

- **State vector:**
  \[
  x = [x, y, \dot{x}, \dot{y}]^T
  \]  
  (250)
  where \((\dot{x}, \dot{y})\) are the components of the speed vector.

- **Measurement:**
  \[
  y = [x, y]^T
  \]  
  (251)

- **Observation matrix:**
  \[
  M = \begin{bmatrix}
  1 & 0 & 0 & 0 \\
  0 & 1 & 0 & 0 
  \end{bmatrix}
  \]

- **Matrix of the dynamic system (state-transition model):**
  \[
  D = \begin{bmatrix}
  1 & 0 & \Delta t & 0 \\
  0 & 1 & 0 & \Delta t \\
  0 & 0 & 1 & 0 \\
  0 & 0 & 0 & 1 
  \end{bmatrix}
  \]  
  (252)

A linear, recursive estimator

Compute: \(x_t\) given \(x_{t-\Delta t}\) and \(y_t\)

Dynamic and measurement models:

\[
\begin{align*}
  x_t & \sim \mathcal{N}(D_t x_{t-\Delta t}; \Sigma_{Dt}) \\
  y_t & \sim \mathcal{N}(M_t x_t; \Sigma_{Mt})
\end{align*}
\]  
(253) (254)

Starting assumptions: \(\bar{x}_0\) and \(\Sigma_0\) are known.

Under these assumptions (linear dynamics, linear measurement function, Gaussian noise), the Kalman filter is the optimal recursive estimator.
Kalman filtering equations

**Prediction**

Predicted state estimate
\[ \bar{x}_t^- = D_t \bar{x}_{t-\Delta t}^+ \]

Predicted estimate covariance
\[ \Sigma_t^- = \Sigma_{t-\Delta t}^+ D_t \]

**Correction/update**

Measurement residual
\[ r_t = y_t - M_t \bar{x}_t^- \]

Residual covariance
\[ R_t = M_t \Sigma_t^- M_t^T + \Sigma_{Mt} \]

Optimal Kalman gain
\[ K_t = \Sigma_t^- M_t^T (R_t)^{-1} \]

Updated state estimate
\[ \bar{x}_t^+ = \bar{x}_t^- + K_t r_t \]

Updated estimate covariance
\[ \Sigma_t^+ = (I - K_t M_t) \Sigma_t^- \]

**Limitations**

- We are not using future measurements to refine past state estimate. **Solution**: forwards-backwards Kalman filtering.

- Non-linear dynamic and measurement models. **Solutions**:
  - Extended Kalman Filter (EKF): local linearization representation.
  - Unscented Kalman filter: non-linear propagation of the mean and covariance using the unscented transform.
How to deal with missing or multiple measurements?

**Missing measurements.** Skip the correction step.

**Multiple measurements.** Choose the closest measurement.

**Multiple measurements and multiple targets.** Find the best global target assignment by computing a distance matrix.

The choice of the (generally probabilistic) distance metric is crucial in this data association problem.

Alternative to Kalman filter: particle filter