Fundamental Concepts of Computer Vision

Alina Trifan
António J. R. Neves

Electronics, Telecomunications and Informatics Department

University of Aveiro

alina.trifan@ua.pt
an@ua.pt
http://sweet.ua.pt/an

Contents

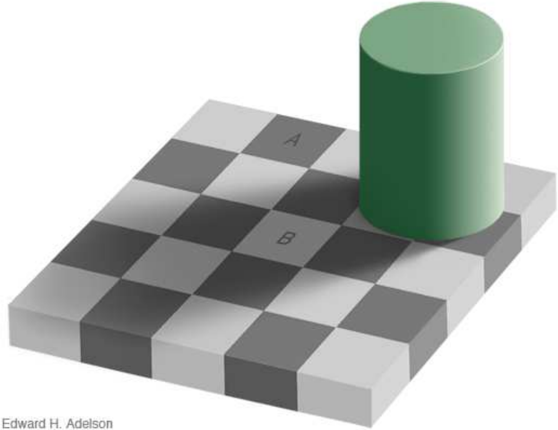
Contents

Introduction (1)

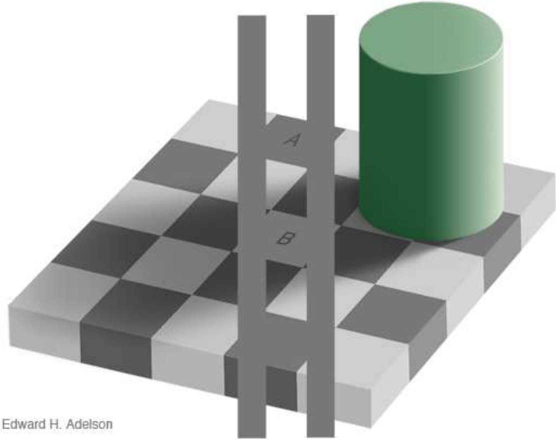
- Vision is a complex physical and intellectual human task that stands as a primary interaction tool with the world.
- It is a complex process not completely understood, even after hundreds of years of research.
- The visualization of a physical process involves an almost simultaneous interaction of the eyes and the brain.
- This interaction is performed by a network of neurons, receptors and other specialized cells.

Introduction (2)

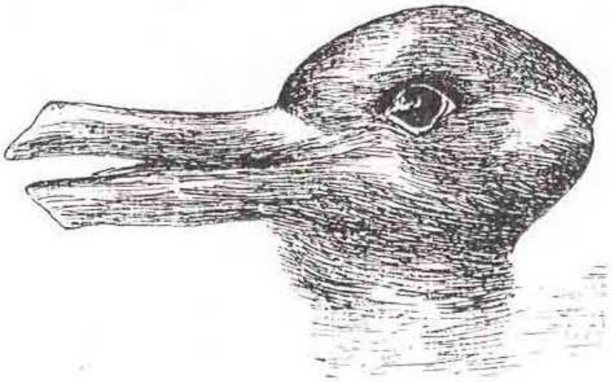
- The human eye is equipped with a variety of optical elements, including the cornea, iris, pupil, a variable lens and the retina.
- Can do amazing things like:
  - Recognize people and objects
  - Navigate through obstacles
  - Understand mood in the scene
  - Imagine stories
- But:
  - Suffers from illusions
  - Ignores many details
  - Ambiguous description of the world
  - Doesn’t care about accuracy of world


A
B
Edward H. Adelson

Illusions (2)

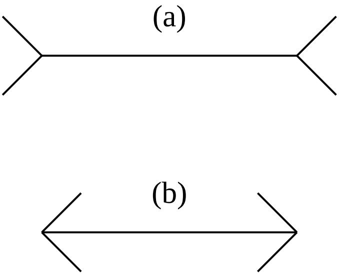


Illusions (3)

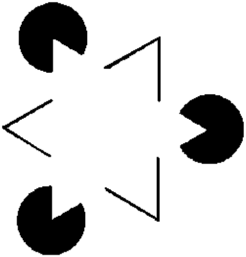


Other illusions…

- The human visual system exhibits a considerable cognitive component, influenced by memory, context, and intention:



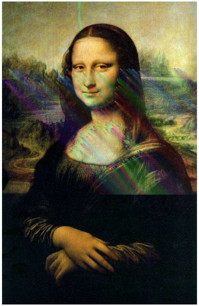
Which is the longer one?

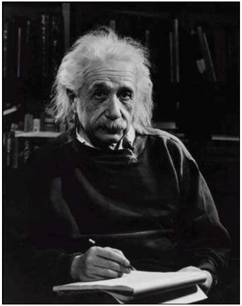


A triangle?

A picture is worth 1000 words

A picture is worth millions of words





Motivation

- Computer vision is a field that includes methods for acquiring, processing, analyzing, and understanding images . . .
- Computer vision applications are increasing:
  - surveillance;
  - machine inspection;
  - medicine;
  - robotics;
  - entertainment;
  - media.
- The utopic goal: make computer vision converge towards human vision. Can we ever accomplish that?

Objectives

- Computer vision seeks to develop algorithms that replicate one of the most amazing capabilities of the human brain - inferring properties of the external world purely by means of the light reflected from various objects to the eyes.
- With vision, it is possible to determine how far away objects are, how they are oriented with respect to the subject, and in relationship to various other objects.
- It is possible to guess their colors and textures and recognize them.
- It is possible to segment regions of space corresponding to particular objects and track them over time.
- In this class, we will see an overview of the concepts and algorithms used in Computer Vision to achieve the referred tasks . . .

Some definitions (1)

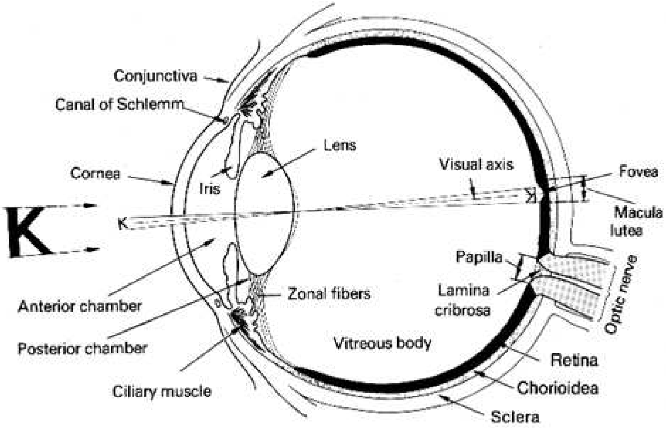
- Image Processing
  - Signal processing where the input signal is an image (2 dimensional signal) and the output can be a transformation of this image or a set of characteristics associated to the image.
- Computer Vision
  - Techniques for image acquisition, extraction, caracterization and interpretation of the information gathered from images of the 3D world.
- Machine Vision
  - Computer Vision for automation and robotic applications.
- Artificial Vision
  - Wide research field that includes all sciences and techniques that allow the study and application of all activities related to the use and interpretation of an image.

Some definitions (2)

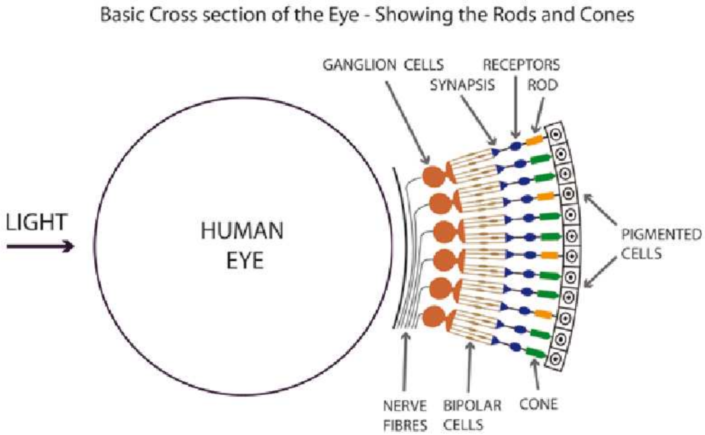
- Perception
  - The process of acquiring an image.
- Pre-processing
  - Image correction(distortion, noise) or enhancement(filtering).
- Segmentation
  - Breaking an image into segments (areas of interest).
- Descriptors
  - Features of the image (shape, size ...).
- Object Detection
  - **Where** is **this** object in the image?
- Object Recognition
  - **Which** object is depicted in the image?
- Image understanding
  - ...

Contents

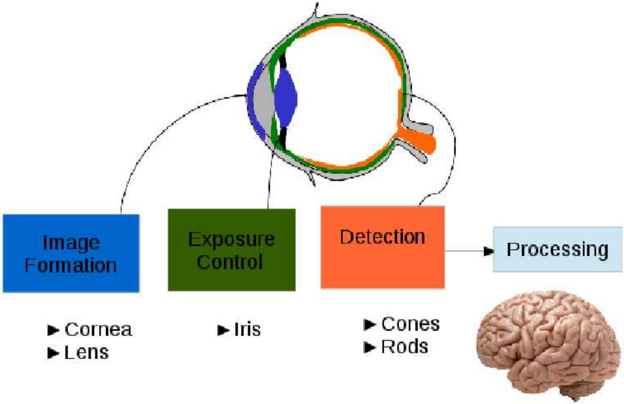
Human vision



Human vision

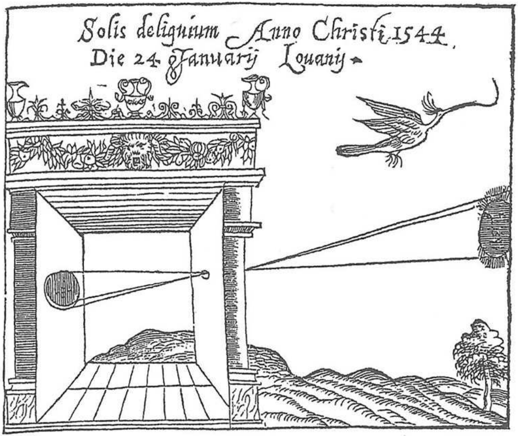


Human vision

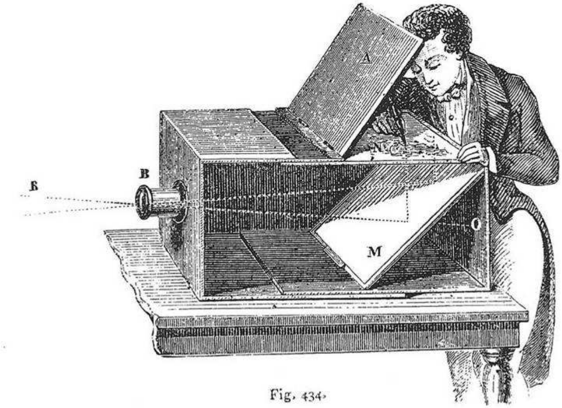


Contents

History of cameras (1544)



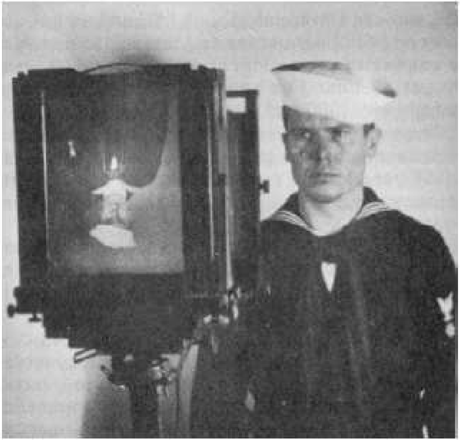
Camera Obscura, Gemma Frisius, 1544



Lens Based Camera Obscura, 1568

Still Life, Louis Jaques Mande Daguerre, 1837

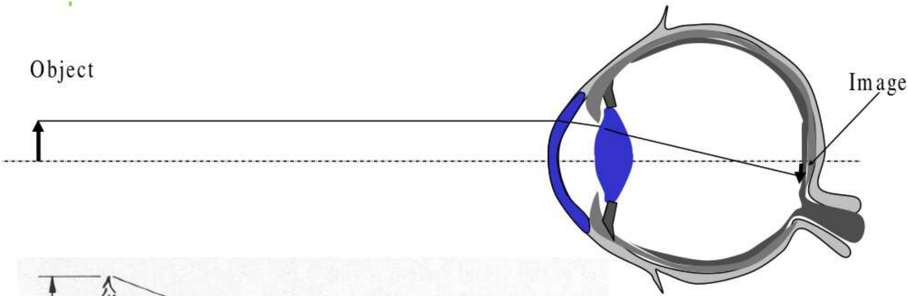
History of cameras (1930)

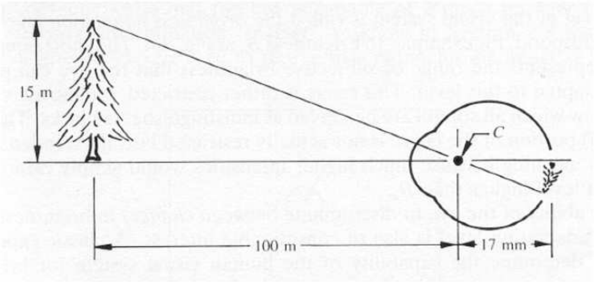


History of cameras (1970 - nowadays)

Silicon Image Detector, 1970 - digital cameras

Human eye





Pinhole Camera Model

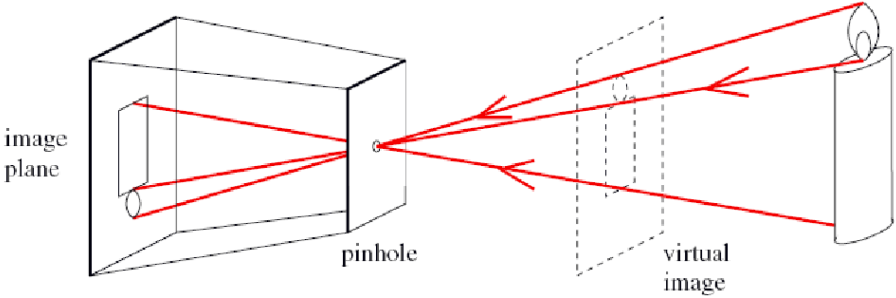
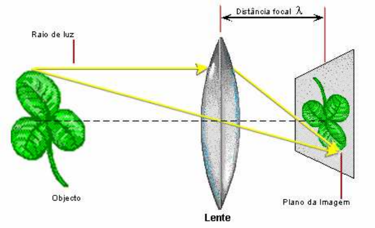


Image through a lens



- All the rays of light that come from an object in direction to the lens converge, on the other side, in another point at a certain distance from the lens. This distance is called focal distance.
- f smaller – wide-angle camera; f gets larger – more telescopic.
- All the points that verify this fact are denoted the focal plane.
- There are some other important parameters related to lens: Field of View, Depth of Field, ...

Basic Camera Geometry

- Far objects appear smaller.
- Lines project to lines.
- Lines in 3D project to lines in 2D.
- Distances and angles are not preserved.
- These geometric properties are “common sense”. Other properties can be inferred if we formalize the model using . . . Mathematics, of course. . .

Contents

Digital camera

- Image acquisition using a digital camera:
  (IEEE SP Magazine, Jan 2005)

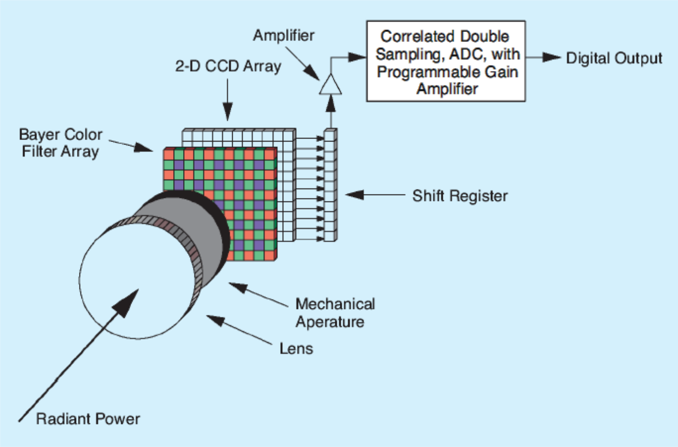
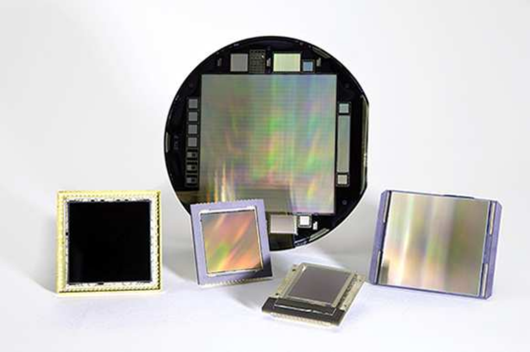
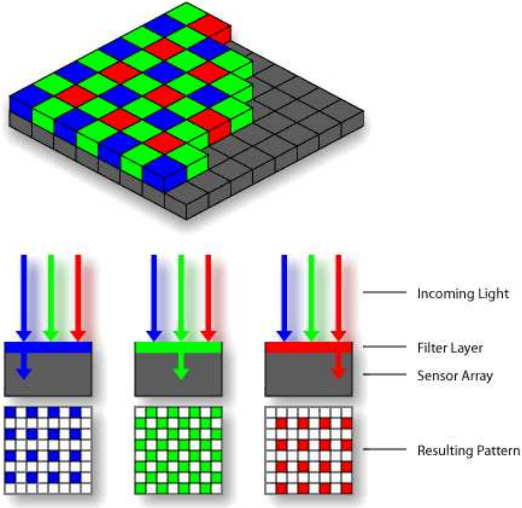


Image sensors



- Some considerations: speed, resolution, cost, signal/noise ratio, . . .
- CCD - charge coupled device - Higher dynamic range, High uniformity, Lower noise.
- CMOS - Complementary Metal Oxide Semiconductor - Lower voltage, Higher speed, Lower system complexity.

The Bayer matrix



Digital cameras - several solutions

Digital cameras - several solutions

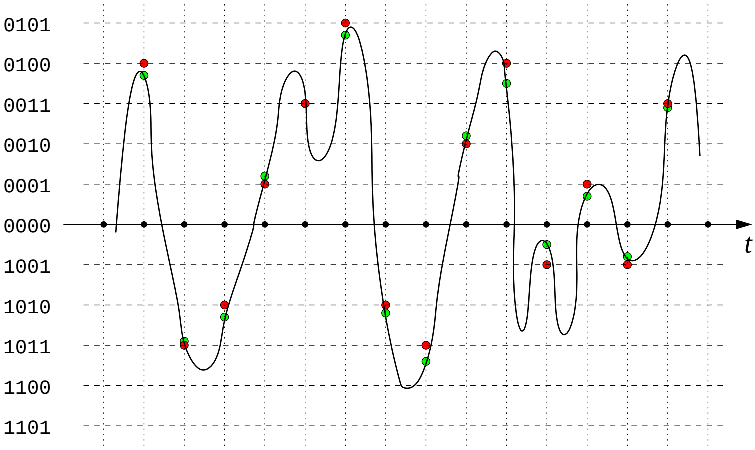
- Several interfaces (Firewire, GigE, CameraLink, USB, . . .).
- Scientific usage (high resolution, long exposure time, . . .).
- High speed (ex. 1000 fps).
- Linear (ex. 10000 lines per second).
- 3D
- Infrared (ex. 8 to 14 μm).
- High dynamic range (ex. using a prism and two sensors).
- Multispectral

Contents

Sampling and quantization

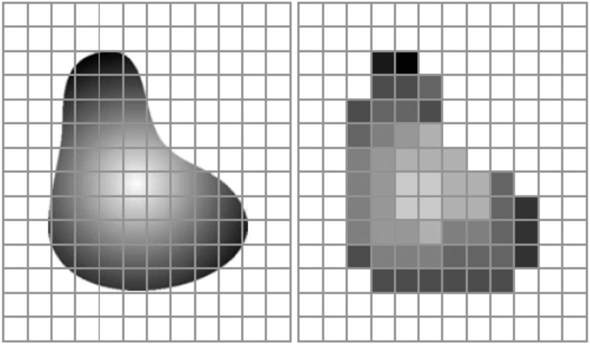
- Generally, an image can be represented by a two-dimensional function, $f(x, y)$, where x and y are spatial coordinates.
- The meaning of f in a given point in space, (x, y), depends on the source that generated the image (visible light, x-rays, ultrasound, radar, ...).
- Nevertheless, we generally assume that $f(x, y) \geq 0$.
- Moreover, both the spatial coordinates and the function values are continuous quantities.
- Therefore, to convert $f(x, y)$ into a digital image, it is necessary to perform spatial sampling and amplitude quantization.

Digitalization: sampling + quantization

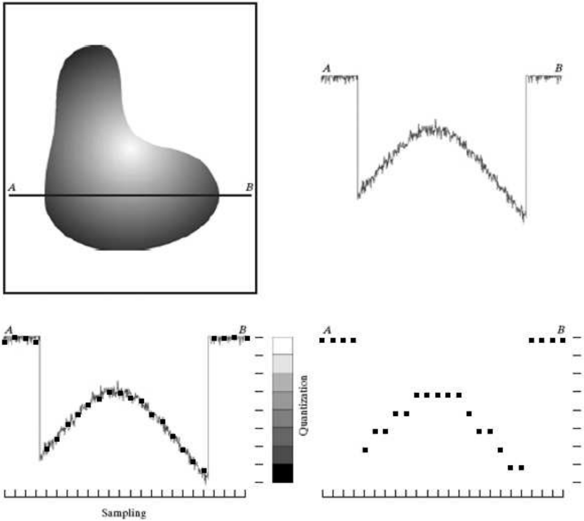


0100 1011 1010 0001 0011 0101 1010...

- Sampling and quantization — example:
(Gonzalez & Woods)

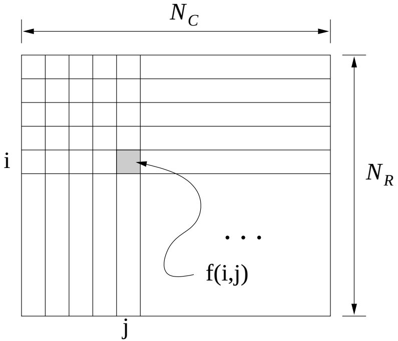


- Sampling and quantization — example:
  (Gonzalez & Woods)



Digital images

- Typically, a digital image is represented by a rectangular matrix of scalars or vectors.



- The $f(i,j)$ are named *pixels* and, usually, $f(i,j) \in \mathcal{I} \subset \mathbb{N}_0^n$.

Digital images

- We will consider digital images of the following types:
  - Black and white (binary images).
    $f(i,j) \in \{0,1\}$
  - Grayscale images.
    $f(i,j) \in \{0,1,\ldots,2^b-1\}$
  - Color-indexed images.
    $f(i,j) \in \{0,1,\ldots,2^b-1\} \xrightarrow{\alpha} \mathcal{I} \subset \{0,1,\ldots,2^{b'}-1\}^3$
  - Color images (for example, RGB images).
    $f(i,j) \in \{0,1,\ldots,2^b-1\}^3$

Examples

Color

Color-indexed (256)

Grayscale (256)

Black and white

Contents

The visible spectrum

Spectral colors (pure colors)

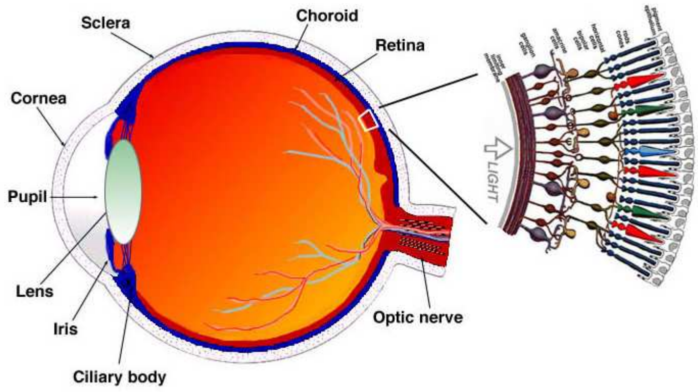
Cor	Wavelength
Violet	$\approx$ 380–440 nm
Blue	$\approx$ 440–485 nm
Cyan	$\approx$ 485–500 nm
Green	$\approx$ 500–565 nm
Yellow	$\approx$ 565–590 nm
Orange	$\approx$ 590–625 nm
Red	$\approx$ 625–740 nm

The human perception of color

- Normally, the characteristics that allow colors to be distinguished are:
  - The brightness (how bright is the color).
  - The hue (the dominant color).
  - The saturation (how pure is the color).
- Together, the hue and the saturation define the chromaticity.
- Therefore, a color can be characterized by the brightness and the chromaticity.

The human perception of color

- The human eye has photoreceptors that are sensitive to short wavelengths (*S*), medium wavelengths (*M*) and long wavelengths (*L*), also known as the blue, green and red photoreceptors.



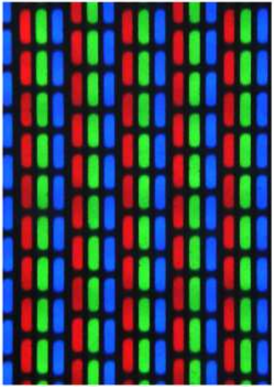
Additive primaries

- The red, green and blue are the three additive primary colors.

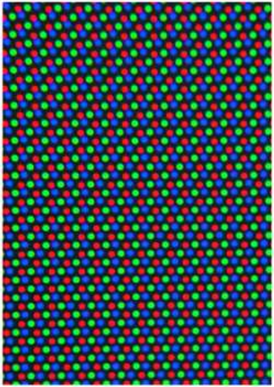
- Adding these three colors produces white.

The *RGB* color space

- Besides the use in acquisition on digital cameras, for example, the displays have pigments of these three colors…



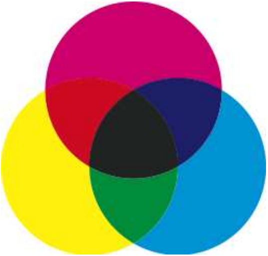
21" TV CRT Display



17" PC CRT Display

The *CMY* color space

- The *CMY* color space is based on the subtractive properties of inks.
- The cyan, magenta and yellow are the subtractive primaries. They are the complements, respectively, of the red, green and blue. For example, the cyan subtracts the red from the white.



- Conversion from *RGB* to *CMY*: $C = 1 - R$, $M = 1 - G$, $Y = 1 - B$.

The *CMY* color space

C component

M component

Y component

The *CMYK* color space

- Due to technological difficulties regarding the reproduction of black, the *CMYK* color space is generally used for printing.

C component

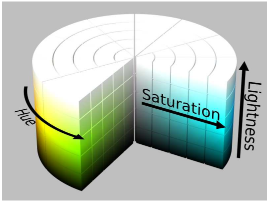
M component

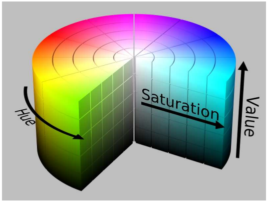
Y component

K component

The *HLS* and *HSV* color spaces

- The HSL and HSV are the two most common cylindrical coordinate representations of colors.
- They rearrange the geometry of RGB colors in an attempt to be more intuitive and perceptually relevant than the cartesian (cube) representation.
- They were developed in the 1970s for computer graphics applications, and are used for color pickers, in color-modification tools in image editing software, and commonly for image analysis and computer vision.





- RGB to HSV:

$$
\begin{aligned}
V &= \max(R, G, B) \\
S &= \begin{cases} \frac{V - \min R,G,B}{V} & if\_V \neq 0 \\ 0 & otherwise \end{cases} \\
H &= \begin{cases} 60(G-B)/S & if\ V = R \\ 120 + 60(B-R)/S & if\ V = G \\ 240 + 60(R-G)/S & if\ V = B \end{cases}
\end{aligned}
$$

- RGB to HSL:

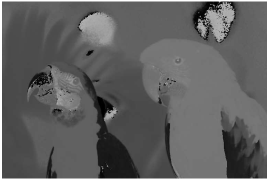
$$
\begin{aligned}
V_{max} &= \max R, G, B \\
V_{min} &= \min R, G, B
\end{aligned}
$$

$$
L = \frac{V_{max} + V_{min}}{2}
$$

$$
S = \begin{cases} \frac{V_{max} - V_{min}}{V_{max} + V_{min}} & \text{if } L < 0.5 \\ \frac{V_{max} - V_{min}}{2 - (V_{max} + V_{min})} & L \geq 0.5 \end{cases}
$$

$$
H = \begin{cases} 60(G - B)/S & \text{if } V_{max} = R \\ 120 + 60(B - R)/S & \text{if } V_{max} = G \\ 240 + 60(R - G)/S & \text{if } V_{max} = B \end{cases}
$$

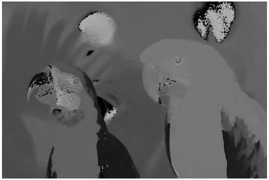
The *HLS* and *HSV* color spaces



H component

S component

V component



H component

S component

L component

The *YUV* color space

- The *YUV* color space is used in some television standards.
- Y is the luminance component:

$$Y = 0.299R + 0.587G + 0.114B$$

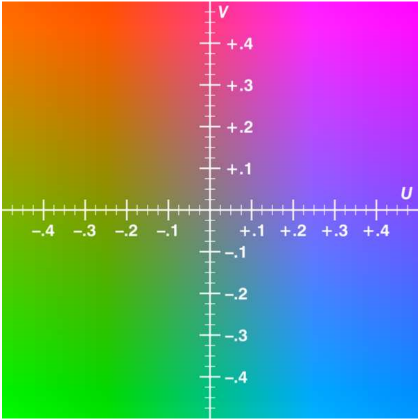
- Components U and V represent the chrominance:

$$\begin{aligned} U &= -0.147R - 0.289G + 0.436B = 0.492(B - Y) \\ V &= 0.615R - 0.515G - 0.100B = 0.877(R - Y) \end{aligned}$$

- For $R, G, B \in [0, 1]$, we have $Y \in [0, 1]$, $U \in [-0.436, 0.436]$ and $V \in [-0.615, 0.615]$.

The *YUV* color space

- $U - V$ plane, for a constant value of Y, equal to 0.5:



Advantages of the *YUV* color space

- The *YUV* color space allowed to maintain the compatibility with the old “black and white” television receivers.
- The human eye is more sensitive to the green color, which is represented mainly by the *Y* component.
- The *U* and *V* components are related to the blue and red.
- Since the human eye is less sensitive to the blue and red, it is possible to reduce the bandwidth used to represent the *U* and *V* components, without introducing significant perceptual degradation.

The YC_bC_r color space

- This is usually designated the digital version of YUV.
- The JPEG standard, as well as some other MPEG video standards, allows all 256 values in an 8 bits per component representation.
- In this case, considering $R, G, B \in \{0, \ldots, 255\}$, we have:

$$
\begin{aligned}
Y &= 0.299R + 0.587G + 0.114B \\
C_b &= 128 - 0.168736R - 0.331264G + 0.5B \\
C_r &= 128 + 0.5R - 0.418688G - 0.081312B
\end{aligned}
$$

- After the conversion, $Y, C_b, C_r \in \{0, \ldots, 255\}$.
- Besides its use in image and video coding, this color space is also used in some computer vision applications.

The YC_bC_r color space

Y component

C_b component

C_r component

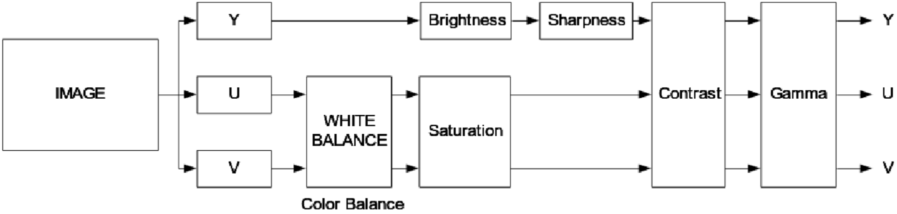
Contents

Camera calibration

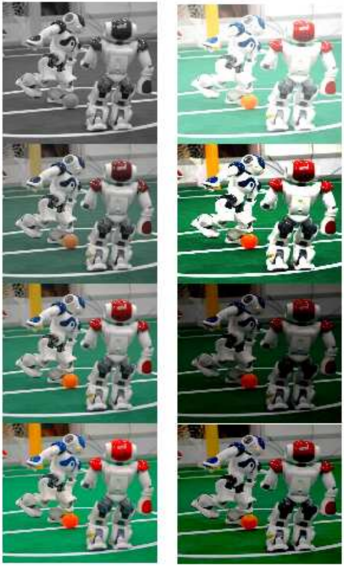
- In order to use a digital camera in some applications, it is necessary to calibrate some parameters.
- Colormetric parameters: the parameters that are related to color and intensity of the acquired image (gain, white-balance, brightness, sharpeness, . . .). The available parameters depends on the image processing pipeline of each camera.
- Extrinsic parameters: the parameters that define the location and orientation of the camera reference frame with respect to a known world reference frame.
- Intrinsic parameters: the parameters necessary to link the pixel coordinates of an image point with the corresponding coordinates in the camera reference frame.

Colormetric parameters (1)

A typical image processing pipeline (inside the image device) for a tri-stimulus system is shown bellow. This processing can be performed on the YUV or RGB components depending on the system. This should be understood as a mere example.



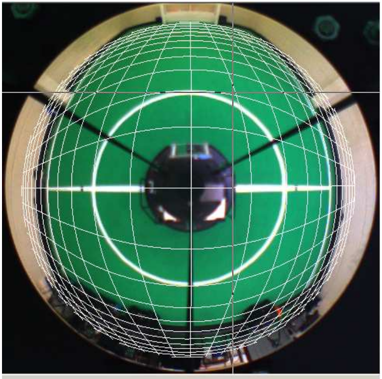
Colormetric parameters (2)

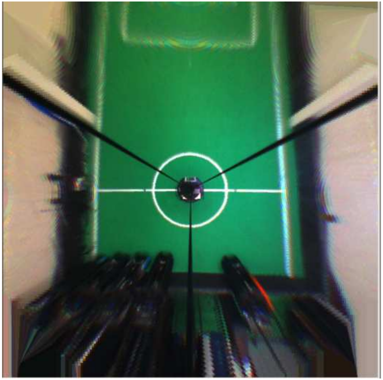


Extrinsic and Intrinsic parameters

- **Extrinsic parameters** denote the coordinate system transformations from 3D world coordinates to 3D camera coordinates.
- They reflect the projection view of the camera taking into consideration its relative position to the world coordinates (translation operation) and its rotations when we consider the camera coordinate system three axis in relation to the wolrd coordinate system three axis.
- **Intrinsic parameters** include focal length, image format, principal point, skew.

Example of intrinsic and extrinsic calibration

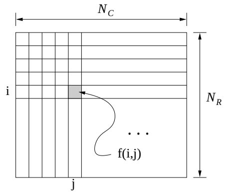


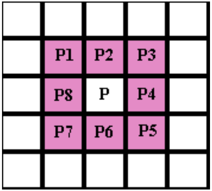


Contents

Pixel Neighbours

- Many image processing operations make use of spatial relationships between pixels.
- A number of methods have been devised to specify pixel neighbors and calculate distance.
- The 4-neighbors of a pixel (x,y) are the closest pixels in horizontal and vertical directions (D4).
- The 8-neighbors are the 4-neighbors plus the four closest pixels in diagonal direction (D8).
- Diagonal only (DN).





Pixel connections

- A group of pixels is said to be 4-connected if every pixel is 4-connected to the group.
- A group of pixels is said to be 8-connected every pixel is 8-connected to the group.

Distances

- The distance between pixels (x,y) and (u,v) can be calculated in several ways:
  - Euclidean (L2): $D = [(x - u)^2 + (y - v)^2]^{1/2}$
  - City-block (L1): $D = |x - u| + |y - v|$
  - Chessboard (Linf): $D = max(|x - u|, |y - v|)$
- Although Euclidean distance is more accurate, the sqrt makes it expensive to calculate.

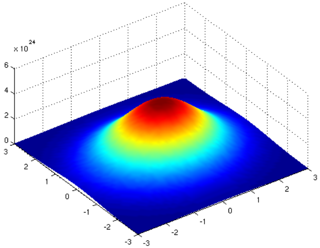
Spacial filtering

- Spatial filters make use of a fixed sized neighborhood in an input image to calculate output intensities.
- Linear filters use a weighted sum of pixels in the input image $f(i,j)$ to calculate the output pixel $g(i,j)$. In most cases, the sum of weights is one, so the output brightness = input brightness.
- Nonlinear filters can not be calculated using just a weighted sum (sqrt, log, sorting, selection).
- We can formalize the phrase “weighted sum of pixels” using correlation and convolution.
- The mathematical model is the discrete convolution operator based on the kernel h:

$$g(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} h(i-m, j-n) f(i,j)$$

Examples of filters (1)

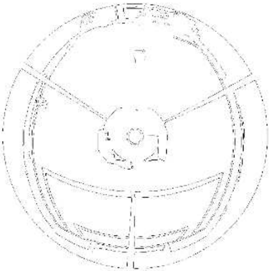
- Average - the easiest spatial filter to implement. The kernel is a matrix with all the values equals to one (the pixel is replaced by an average of the $N \times M$ neighbors). This filter smooths an image and removes noise and small details.
- Binomial - uses Binomial coefficients as weights to give more emphasis to pixels near the center of the $N \times M$ neighborhood.
- Gaussian - uses the Gaussian function to define the neighborhood weights.



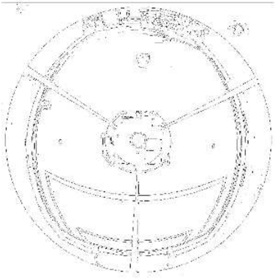
Example of filters (2)

Edge detection

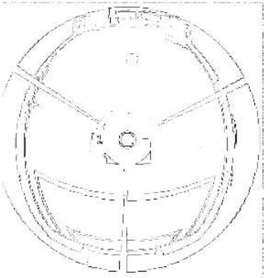
Canny

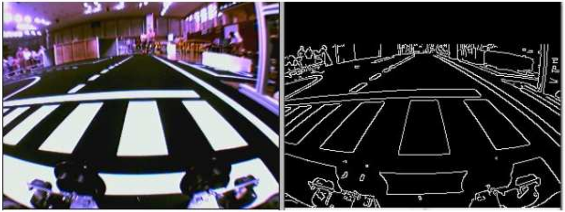


Laplace

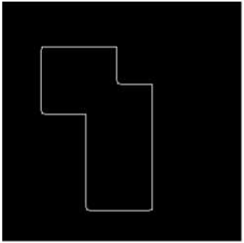


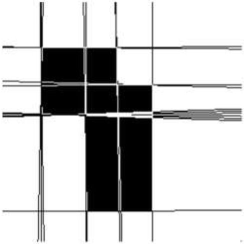
Sobel

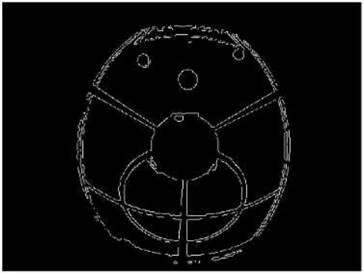


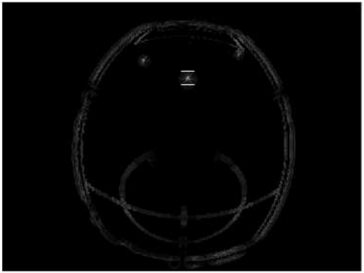


Hough Transform









Contents

Histograms: definition

- In statistics, a histogram is a graphical display of tabulated frequencies.
- Typically represented as a bar chart.

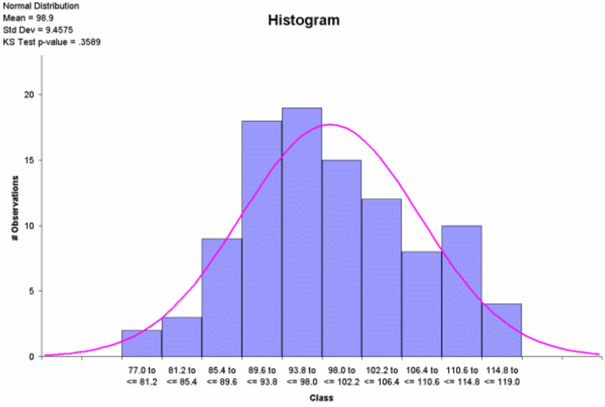
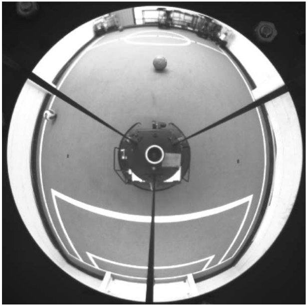


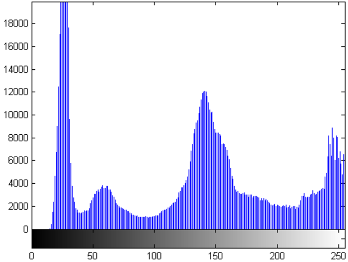
Image histograms

- In images, allow us to see the color or intensity distribution.
- The collected counts of data can be organized into a set of predefined bins.
- It is also possible to count image features that we want to measure (i.e. gradients, directions, etc).
- Some important parts of an histogram:
  - dims: The number of parameters you want to collect data.
  - bins: The number of subdivisions in each dim.
  - range: The limits for the values to be measured.
- If we want to count two features, the resulting histogram would be a 3D plot (in which x and y would be bin_x and bin_y for each feature and z would be the number of counts for each combination of (bin_x, bin_y)).

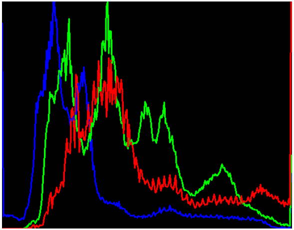
Histograms: example (1)

- Example of an histogram obtained from a grayscale image.
- Each bin shows the number of times each one of the gray values are present in the image.





- Example of an histogram showing the distribution of the colors on an image.



Histograms: operations

- Histogram operations are designed to enhance the visibility of objects of interest in an image.
- Histogram Equalization - improves the contrast in an image, in order to stretch out the intensity range.
- Local Histogram Equalization - increase the amount of enhancement by looking at local intensity properties (dividing an image into regions and perform histogram equalization on each sub-image or using local statistics).
- Histogram Comparison - get a numerical parameter that expresses how well two histograms match each other (ex. Correlation, Chi-Square, Intersection, . . .).
- Sum, subtract, . . .

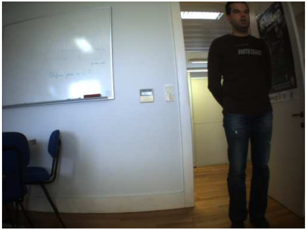
- Goal of histogram equalization is to reshape the image histogram to make it flat and wide.
- One of the solutions is to use the cumulative histogram (integral of intensity histogram) as the intensity mapping function.

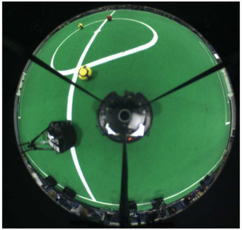
Contents

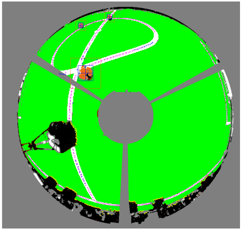
Segmentation: concept

- Intermediate processing towards object recognition.
- Localize regions with common properties.
- Make a partition over the pixel ensemble.
- Usual grouping properties (Gray level, Color, Texture).
- Often requires preprocessing.
- Segmentation of non-trivial images is a difficult task.
- Segmentation accuracy determines the eventual success/failure of computerized image analysis.

Applications of segmentation







Thresholding

- The basis of many region based segmentation algorithms.
- The most immediate and computationally appealing step.
- Direct image partition based on intensity properties.
- Several approaches:
  - Global Thresholding
  - Variable Thresholding
    - Local - $T(x, y)$ depends on properties of the neighborhood of (x, y).
    - Adaptive - $T(x, y)$ depends on the spatial coordinates, x and y.
    - The Otsu's method - Optimal global thresholding based on probabilistic estimates obtained from the histogram.

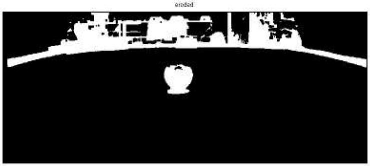
Region Growing

- Region growing is a procedure that groups pixels or subregions into larger regions based on a predefined criteria.
  - Start with a set of "seed" points and from these, grow regions by appending to each seed those neighboring pixels that have properties similar to the seed (intensity, color, . . .).
- Selection of seeds
  - Often interactive
  - Automated
- Centroids of pixel clusters
- Additional criteria: size and shape of region grown so far
- Stopping rules
  - Ideally, growing a region should stop when no more pixels satisfy the criteria for inclusion in that region.

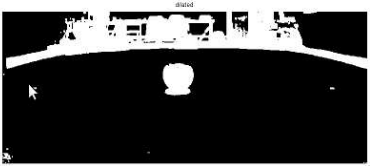
Contents

Morphological operators

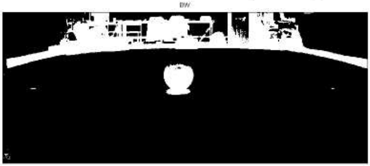
Green channel of the RGB image



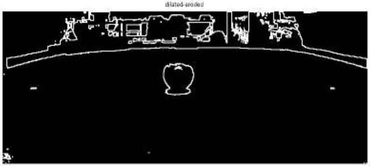
Binary image (Otsu method)



Dilation of the binary image



Erosion of the binary image



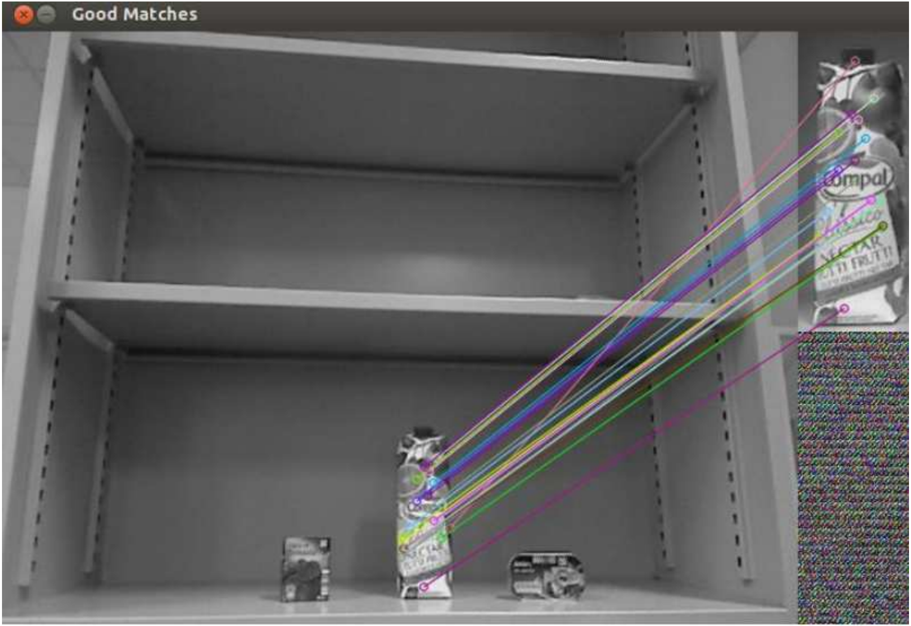
Diference between dilation and erosion

Contents

Image Descriptors

- **SIFT** - Scale Invariant Features
  - SIFT features are generated by transforming an image into a collection of feature vectors, similar to neurons in inferior temporal cortex used in primate vision.
  - SIFT keypoints of objects are extracted from a set of reference images and stored in a database.
  - Recognition of an object in a new image is done by comparing each feature from this new image to the database.
  - Best candidate is found based on the Euclidean distance of the feature vectors.
- **SURF** - Speeded Up Robust Features
  - Inspired by SIFT . . . faster, more robust to image transformations.
  - Relies on integral images for image convolutions and uses a Hessian matrix-based measure for the detector and a distribution-based descriptor, simplified to the essential.

Image Descriptors



Contents

Digital video

- A video signal can be represented by a 3D-function, $v(x, y, t)$, where x and y are spatial coordinates and t denotes time.
- The process of converting analog video into digital video requires both spatial and temporal sampling, besides amplitude quantization.
- Therefore, a digital video is a temporal sequence of digital images which we represent by $v(i, j, k)$, with $k = t/T, k \in \mathbb{N}_0$.
- $T \in \mathbb{R}$ indicates the period of time between two consecutive images (we call them frames). Therefore, $1/T$ (Hz) is the frame rate.
- Sometimes we will refer to video fields. They occur in interlaced video and are made of the even (odd) lines of a frame.

Motion analysis

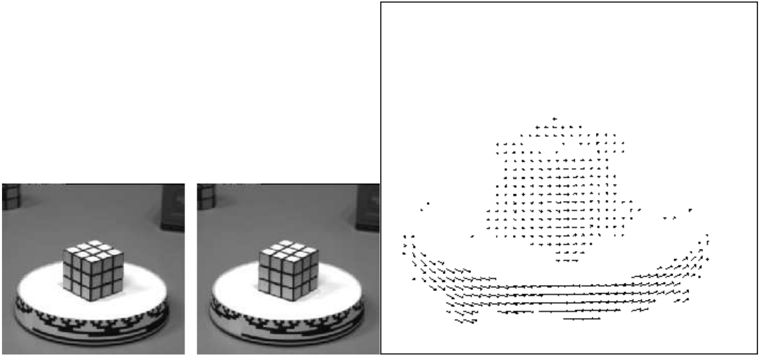
- Several information can be extracted from time varying sequences of images:
  - Camouflaged objects are only easily seen when they move
  - The relative sizes and position of objects are more easily determined when the objects move
  - Even simple image differencing provides an edge detector for the silhouettes of texture-free objects moving over any static background.

Motion analysis

- The analysis of visual motion can be divided into two stages:
  - the measurement of the motion
  - the use of motion data to segment the scene into distinct objects and to extract three dimensional information about the shape and motion of the objects.
- There are two types of motion to consider:
  - movement in the scene with a static camera,
  - and movement of the camera, or ego motion.
- Since motion is relative, these types of motion should be the same. However, this is not always the case, since if the scene moves relative to the illumination, shadow and specularities effects need to be dealt with.

Motion field

- The motion field is the projection of the 3D scene motion into the image.

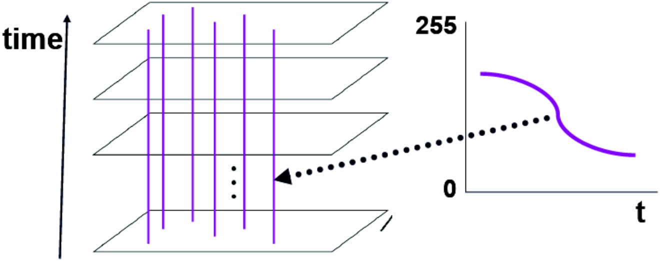


Optical flow

- Definition: optical flow is the apparent motion of brightness patterns (or colors) in the image.
- Ideally, optical flow would be the same as the motion field.
- Have to be careful: apparent motion can be caused by lighting changes without any actual motion.
- To estimate pixel motion from image we have to solve the pixel correspondence problem.
- Given a pixel in frame t, look for nearby pixels with same characteristics (color, brightness, ...) in frame $t - 1$.

Background subtraction

- It is possible to look at video data as a spatio-temporal volume.
- If camera is stationary, each line through time corresponds to a single ray in space.



Background subtraction

- Background subtraction is a commonly used class of techniques for segmenting out objects of interest in a scene for applications such as:
  - Surveillance
  - Robot vision
  - Object tracking
  - Traffic applications
  - Human motion capture
  - Augmented reality

Background subtraction

- It involves comparing an observed image with an estimate of the image if it contained no objects of interest.
- The areas of the image plane where there is a significant difference between the observed and estimated images indicate the location of the objects of interest.
- The name background subtraction comes from the simple technique of subtracting the observed image from the estimated image and thresholding the result to generate the objects of interest.

Important issues

- foreground detection – how the object areas are distinguished from the background;
- background maintenance – how the background is maintained over time;
- post-processing – how the segmented object areas are postprocessed to reject false positives.

$\Rightarrow$