Machine Perception

Lecture 2: Three-dimensional computer vision

Piotr Skrzypczyński
Institute of Robotics and Machine Intelligence
Poznań University of Technology

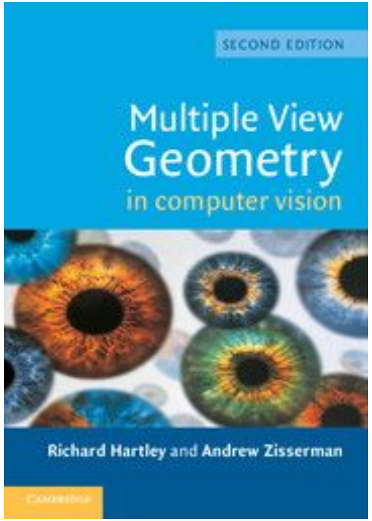
Lecture outline

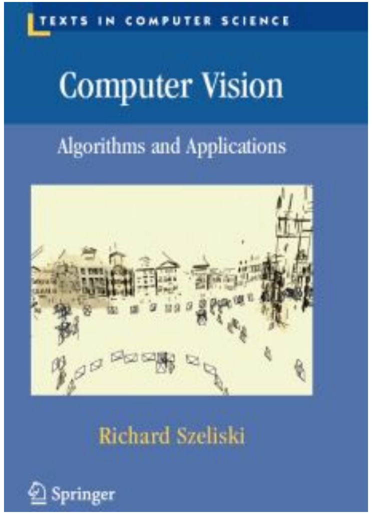
- Image formation and image models.
- Projective geometry.
- Modeling cameras, projection matrix, camera distortions and artifacts.
- Camera calibration (intrinsic, extrinsic).
- The geometry of multiple views: epipolar constraints, disparity, correspondence.
- Sparse and dense stereo in applications.

Literature

1. R. Hartley, A. Zisserman, *Multiple View Geometry in Computer Vision*. 2nd edition, Cambridge University Press, 2004.

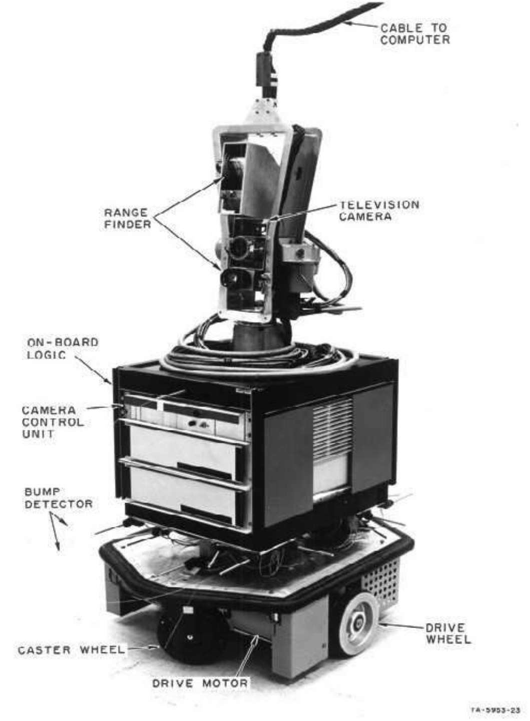
2. R. Szeliski, Computer Vision, Algorithms and Applications, 2nd edition, Spronger, 2022.





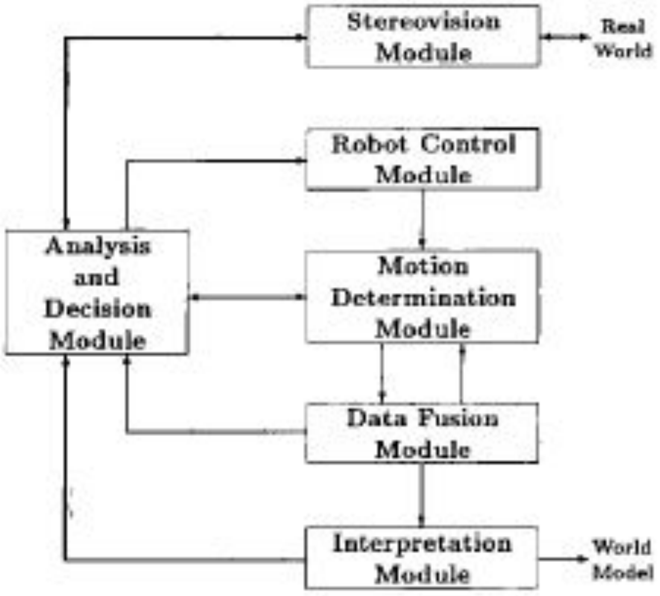
The importance of visual sensing

- The organ of sight is the primary sense in many animals, especially the higher organised and more active ones (e.g. mammals and birds of prey). It is also the primary sense in humans, providing the brain with most information about the environment. Robots, too, can obtain the most information about the environment through perception in the visible light range (electromagnetic waves from about 380 nm to 780 nm) using vision sensors (cameras). The cameras used in robots are passive matrix sensors that record the perspective projection of an illuminated portion of the environment onto the spatially discretised plane of an image matrix.
- Television cameras have been used in autonomous robots essentially since the advent of this type of robot. Early attempts to use cameras as primary sensors to support the navigation of a mobile robot showed that machine vision was difficult to realise, mainly due to the complexity of the image processing methods, requiring a high computational power of the computer system analysing the image.



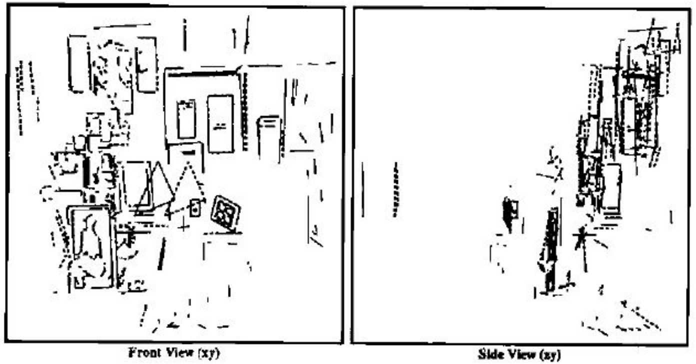
The importance of visual sensing

- The classical paradigm for solving machine vision problems refers to a layered information processing scheme and to generating a description of the scene (from the detail to the whole). It leads to a complex model of the environment built in stages, characterised by a strong reduction in the amount of data and an increase in the abstraction of representations at each stage



The importance of visual sensing

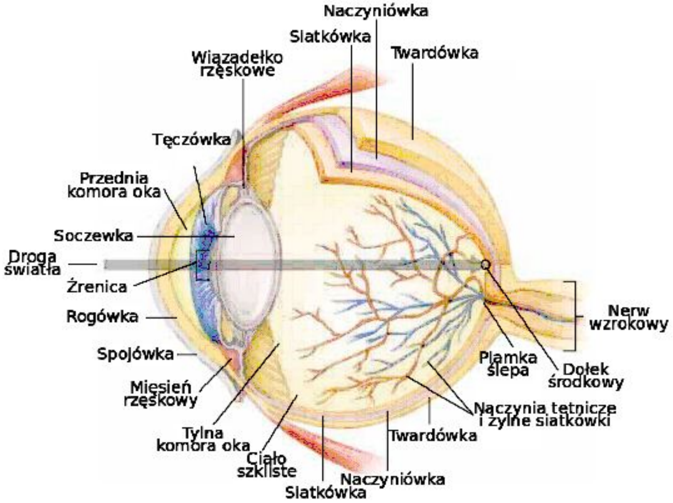
- The result of the image analysis is an abstract, geometric and/or semantic model of the environment or part of it. To reconstruct this model, it is necessary to completely analyse the scene as seen by the camera. In this situation, it can take a long time to obtain simple and basic information for the robot control system.
- The outlines of a new approach to image processing, better adapted to the specificities of autonomous robots, have been outlined on the basis of observations of the perceptual processes of living organisms, in which vision perception is based on the attribution of appropriate processing patterns to individual behaviours. This approach assumes a specialisation of image processing due to the supported functions of the robot, such as obstacle avoidance or self-localisation.

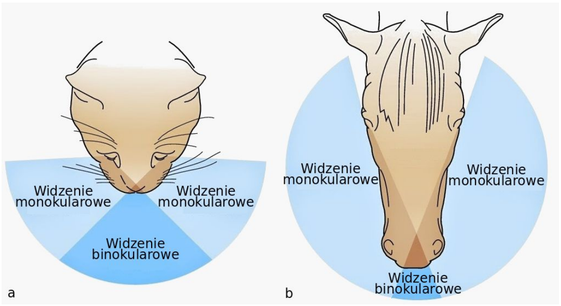

Front View (xy)

Side View (zy)

Biological inspirations of visual sensing

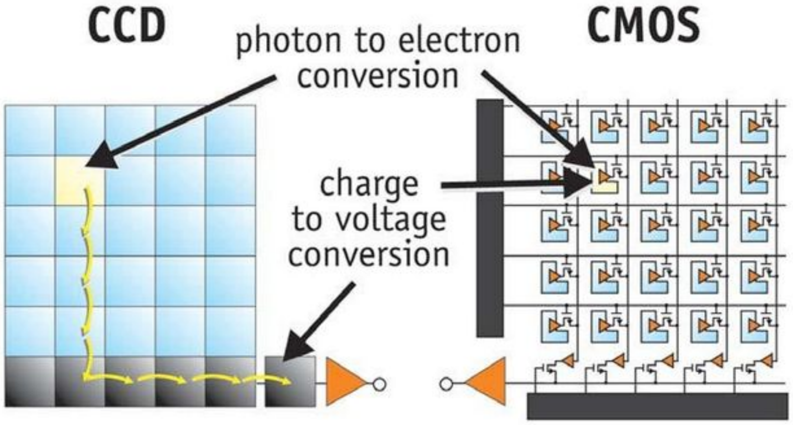
The organ of sight uses photoreceptors to register light stimuli. There are different organs of sight depending on the species and genus of living beings and their evolutionary level.The simplest organ of sight is the maculae otherwise known as ocelli or simple eyes. They are found in invertebrates and some arthropods. Simple eyes are made up of two visual cells and a pigment (pigment) cell, located under the clear epidermis. Simple eyes are unable to register images and colours. Compound eyes otherwise known as mosaic or facet eyes are found in crustaceans and insects. The most advanced type of visual organ is the chambered eyes, found in vertebrates,. The eyes of all vertebrates are structured in a similar way.

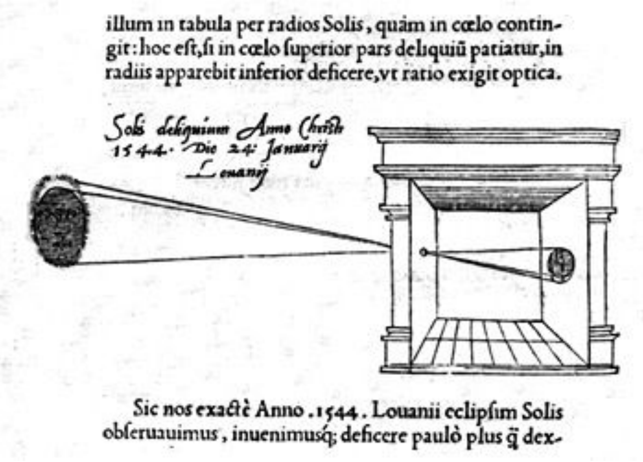




Passive vision sensor

Digital cameras are usually equipped with one of two types of matrix: CCD or CMOS. In the case of a CCD (Charge Coupled Device) matrix, practically the entire area to be registered in the image is searched for as detailed a colour reproduction as possible. This is done in a serial fashion, with individual lines of adjacent sensors transmitting registered data to each other. At the end of each line of sensors, nodes are formed, which in turn also transfer the information in series for the final averaging analysis. CMOS (Complementary Metal Oxide Semiconductor) matrices record the image in a slightly different way. Each element records the image separately.





illum in tabula per radios Solis, quàm in cœlo contingit: hoc eſt, ſi in cœlo ſuperior pars deliquiũ patiatur, in radiis apparebit inferior deficere, vt ratio exigit optica.

Solis deliquium Anno Christi
1544. Die 24. Januarij
Louanij

Sic nos exactè Anno .1544. Louanii eclipſim Solis obſeruauimus, inuenimusq; deficere paulò plus q̃ dex-

Simple camera model

The perspective camera model is complex due to the use of lens of various types, which introduce numerous distortions. However, the simplest is the pinhole camera model. It is a structure consisting of of an impermeable plane with a hole in the centre and a screen. A single ray of light passes through this hole from a specific point in the scene and it forms a rotated image of the object on the light-sensitive surface. Each ray reflected from the object moves towards the towards the centre of the projection, forming an image of the object at the point of intersection of the ray and the projection screen. of the ray and the projection screen. This point is located at the projection distance from the centre of the projection. From the similarity of triangles, the relationship can be obtained

$$\frac{x}{f} = \frac{X}{Z} \Longrightarrow x = f * \frac{X}{Z},$$

where: f - projection distance, Z - distance of the camera from the object, X - size of the object, x - image of the object.

Simple camera model

The next formula allows us to determine the position of a point p(x, y), which is the image of a real point P (X, Y, Z),:

$$x = f_x * \frac{X}{Z} + c_x, y = f_y * \frac{Y}{Z} + c_y,$$

where: fx, fy - the projection distance along the x and y axes, cx, cy - the distance between the centre of the projection screen and the optical axis along the x and y axes. The relationship between a physical point and its image in the camera can also be represented by the relationship:

$$\overrightarrow{p} = M * \overrightarrow{P},$$

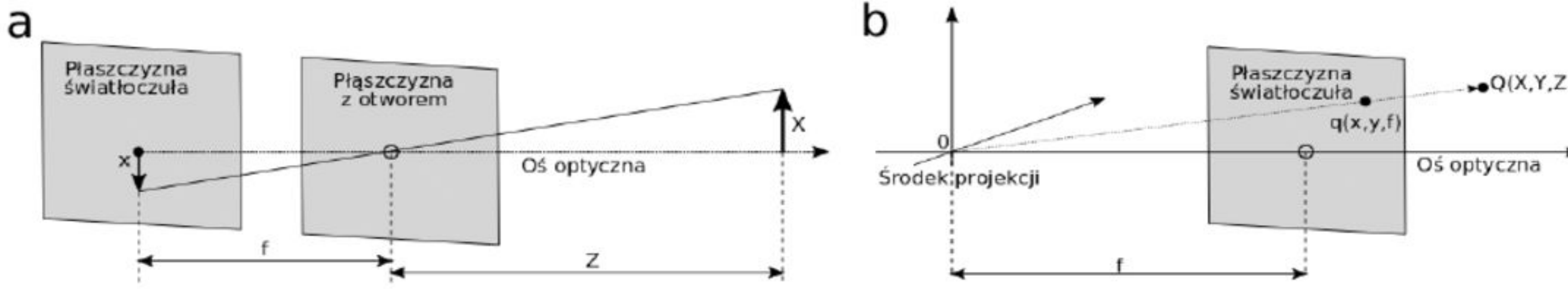
$$\overrightarrow{p} = \begin{bmatrix} x \\ y \\ w \end{bmatrix}, M = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \overrightarrow{P} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix},$$

where: M is the camera intrinsics matrix (camera intrinsics matrix).

Simple camera model

Using the camera intrinsics matrix, three-dimensional coordinates of an object can be converted into two-dimensional image coordinates. The relationship between the object and its image is described by means of linear transformations. On the basis of these relationships, it is also possible to carry out the reverse operation, i.e. determining the position of a physical object, however, the use of a a single-camera vision system does not allow the distance to be determined.

$$x = f_x * \frac{X}{Z} + c_x, y = f_y * \frac{Y}{Z} + c_y,$$

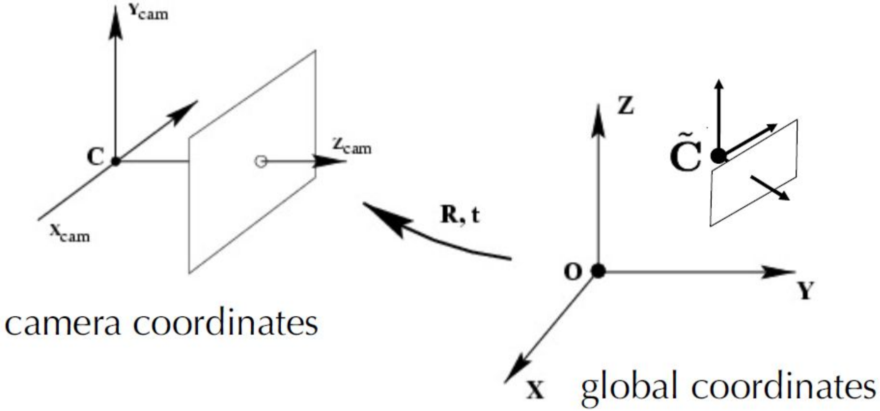


Camera calibration

- Intrinsic parameters
- Extrinsic parameters

$$\mathtt{K} = \begin{pmatrix} f & s & p_x \\ 0 & \alpha f & p_y \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathtt{K} = \begin{pmatrix} f & 0 & w/2 \\ 0 & f & h/2 \\ 0 & 0 & 1 \end{pmatrix}$$



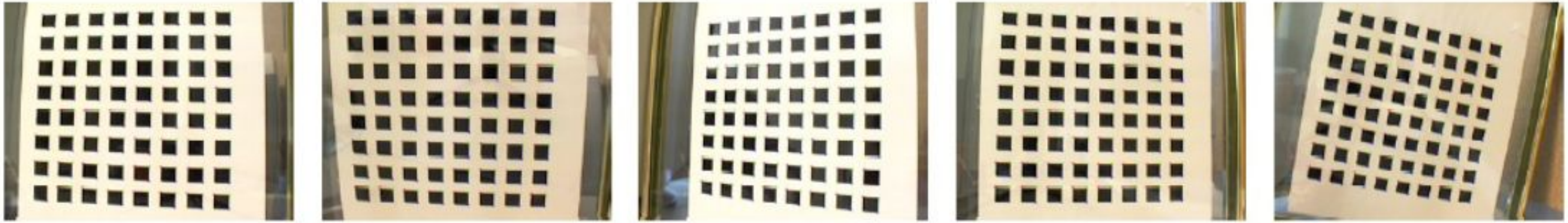
Transformation from global to camera coordinates:

$$\mathbf{X}_{\text{cam}} = \mathtt{R}\left(\mathbf{X}_{\text{global}} - \tilde{\mathbf{C}}\right)$$

Projection from 3D global coordinates to pixels:

$$\lambda \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \mathtt{K}\left(\mathtt{R}\mathbf{X}_{\text{global}} + \mathbf{t}\right)$$

Camera calibration



Unknown: constant camera intrinsics **K**
(varying) camera poses **R,t**

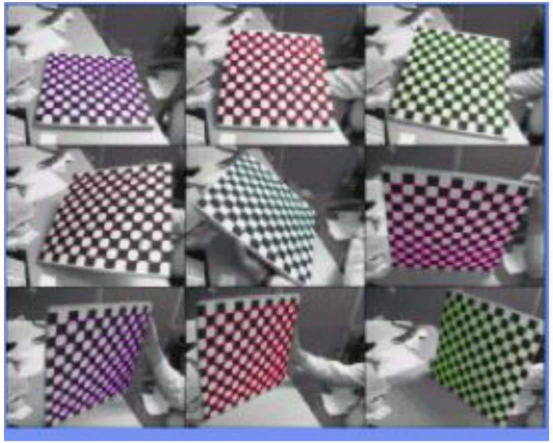
Known: 3D coordinates of chessboard corners
=> Define to be the **z=0** plane ($\mathbf{X}=[X_1\ X_2\ 0\ 1]^T$)

Point is mapped as $\lambda\mathbf{x} = \mathbf{K}\,(\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3\ \mathbf{t})\ \mathbf{X}$

$$\lambda\mathbf{x} = \boxed{\mathbf{K}\,(\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{t})}\,[X_1\ X_2\ 1]'$$

$$\mathbf{K} = \begin{bmatrix} f_x & & p_x \\ & f_y & p_y \\ & & 1 \end{bmatrix}$$

Homography **H** between image and chess coordinates, estimate from known $\mathbf{X}_i$ and measured $\mathbf{x}_i$



Camera calibration

Radial distortion due to imperfect lenses.

One of the possible models:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \sim \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \mathsf{R} \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R}^\top & \text{-}\mathbf{R}^\top \mathbf{t} \\ \mathbf{0}_3^\top & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \right)$$

$$\mathsf{R}: \quad (x, y) = (1 + K_1(x^2 + y^2) + K_2(x^2 + y^2)^2 + \ldots) \begin{bmatrix} x \\ y \end{bmatrix}$$

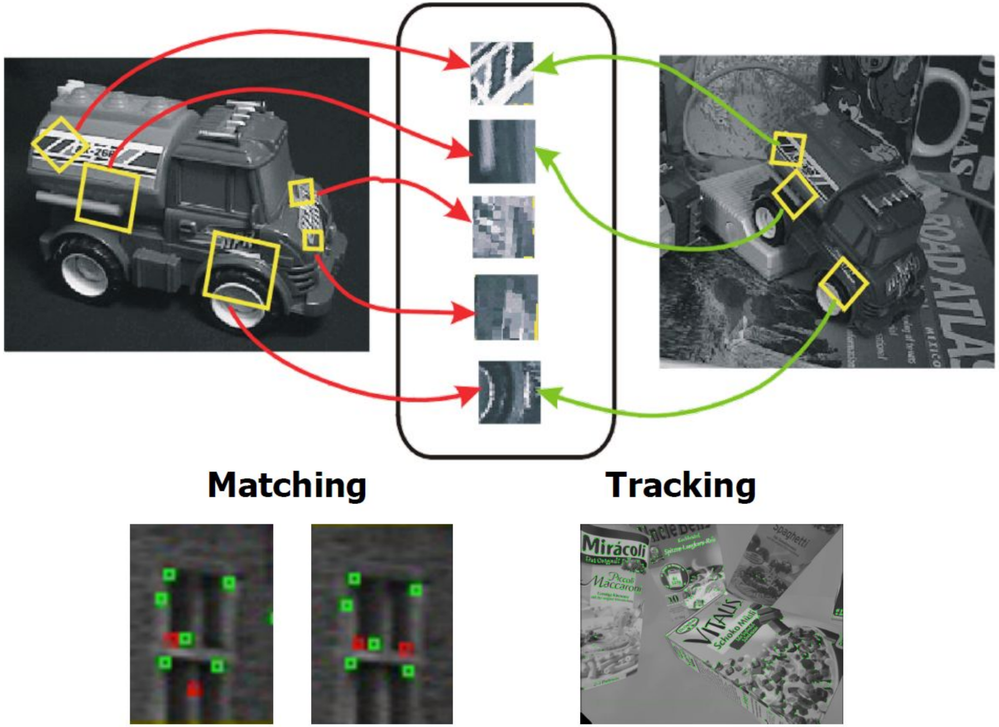
Feature detectors and descriptors

Detector: Find salient structures

- Corners, blob-like structures.
- Keypoints should be repeatable

Descriptor: Compact representation of image region around keypoint

- Describes patch around keypoints
- Establish matches between images by comparing descriptors



Matching

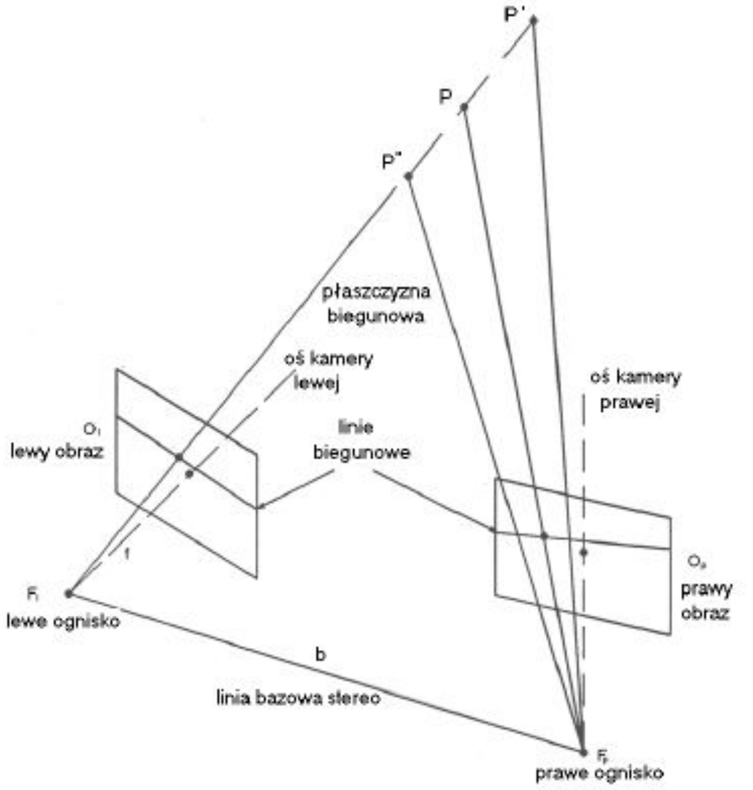
- Extract features independently
- Match by comparing descriptors

Tracking

- Extract features in first image
- Find same feature in next view

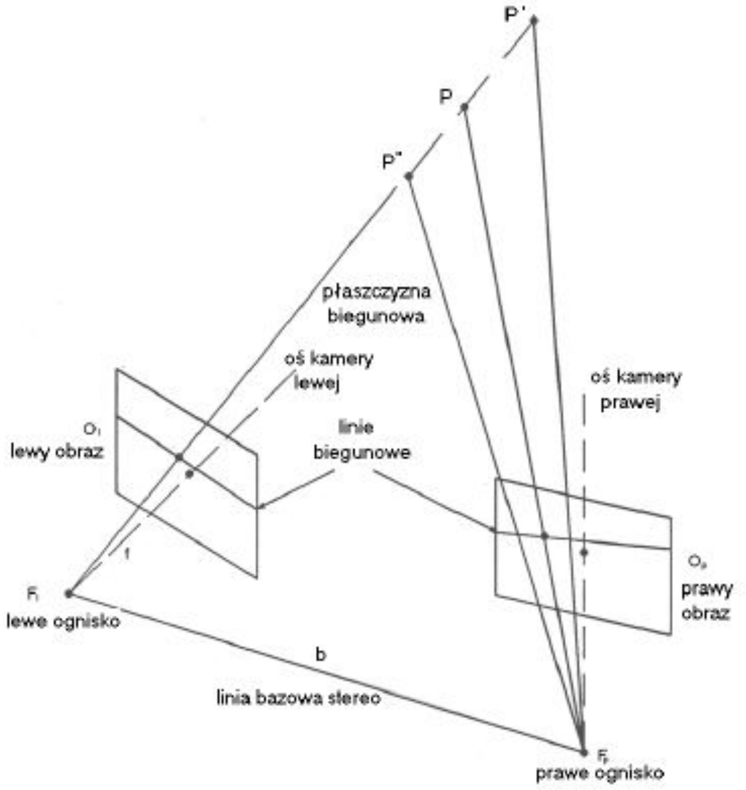
Passive stereo vision

Passive stereo vision: involves measuring the distance to points in a scene by using two cameras and the principle of triangulation. The main computational problem is to find the correspondence between points from the two images. This requires separating certain features (points or lines) in both images and matching corresponding features in both images. Both of these steps are complex and lengthy, and the resulting depth map is generally sparse.



Passive stereo vision

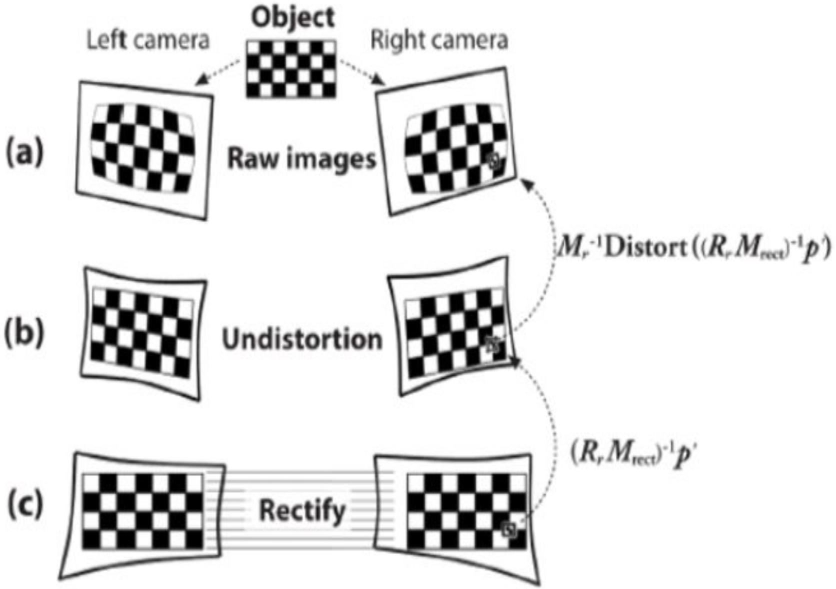
The epipolarity of the geometry of the stereo vision system is that the projections of the point P(X,Y,Z) of the scene onto the image planes of the stereoscopic pair lie on corresponding epipolar lines each point on the left image line has a counterpart on the right image line. In this way, the search to match a point from the first image is restricted to the epipolar line in the second image (instead of the plane), resulting in an important simplification of the calculation.



Passive stereo vision

It is very important that both cameras are perfectly frontally parallel, which is not possible in practice.

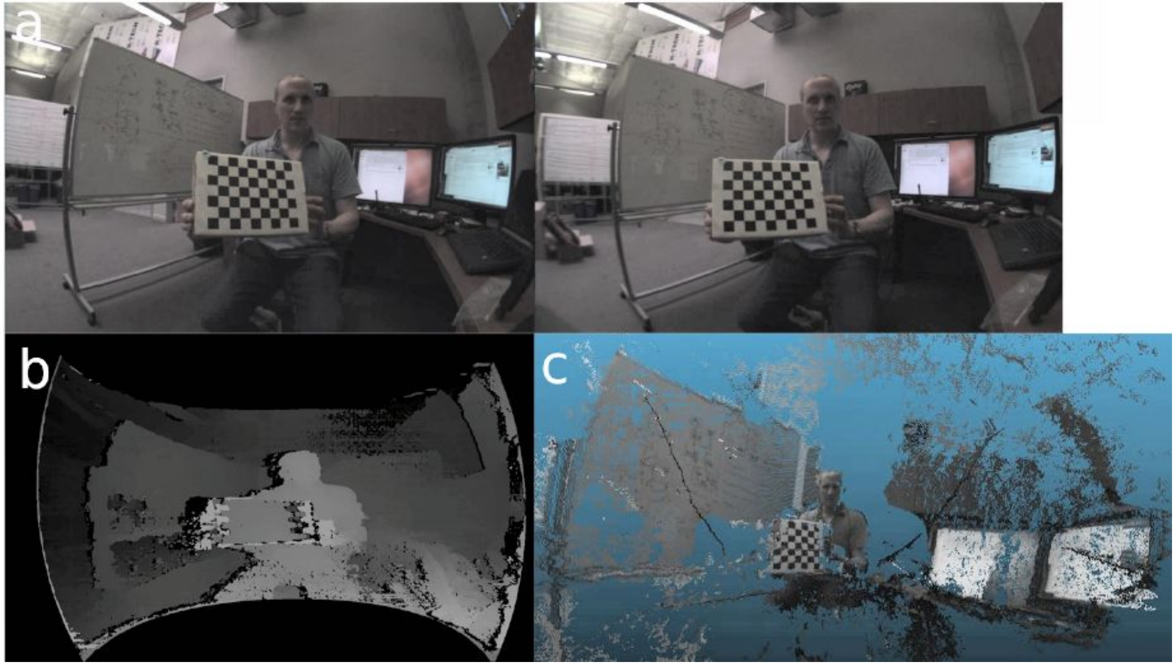
The images are therefore subjected to a rectification process, during which matrices are determined matrices are determined that will allow straighten the images so that the the surfaces of the images belong to the same plane and the corresponding pixel rows on both matrices are collinear.


Object
Left camera
Right camera
(a)
Raw images
$M_r^{-1}\text{Distort}((R_r M_{\text{rect}})^{-1}p')$
(b)
Undistortion
$(R_r M_{\text{rect}})^{-1}p'$
(c)
Rectify

Passive stereo vision

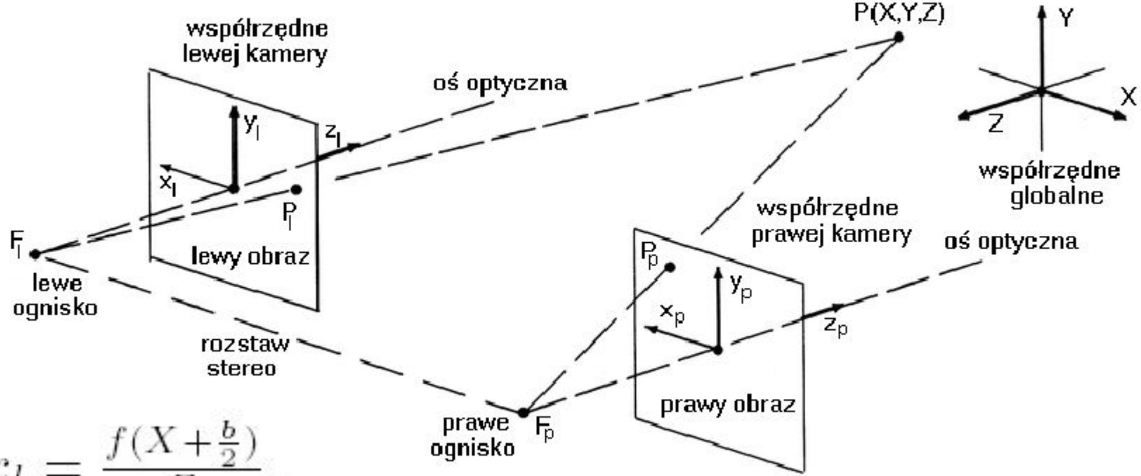
- a - example of disparity determination: stereo image pair
- b - disparity matrix
- c - depth image

Note that the determined depth map, despite the fact by definition dense (dense depth map) has many areas without a specific depth values.



Passive stereo vision

- The formal relationships between the coordinates of the scene points and the image points are given in the formulas below the figure.
- These formulae are valid when the axes of the systems ar parallel and equally oriente projecting each point of th observation field onto the le and right image (through th focal points).



$$x_l = \frac{f(X+\frac{b}{2})}{Z},$$

$$x_p = \frac{f(X-\frac{b}{2})}{Z},$$

$$y_l = y_p = \frac{fY}{Z}. \qquad d \overset{\Delta}{=} x_l - x_p,$$

$$X = \frac{b(x_l + x_p)}{2d}, \qquad Y = \frac{by_l}{d}, \qquad Z = \frac{bf}{d}.$$

Epipolar geometry

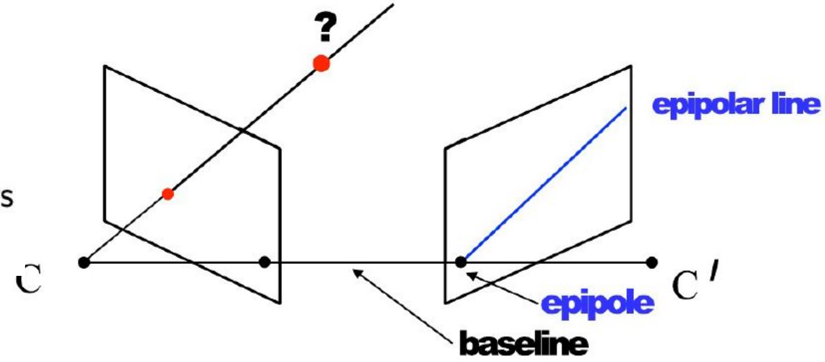
Lets assume the camera parameters and geometry is known!

Given a projection of a 3D point in the left image

Where is it located in 3D?

On the epipolar line defined by this point and the camera centers

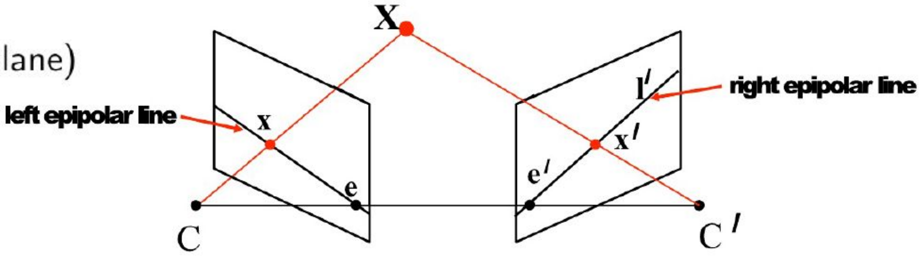
Reduces the search problem to 1D!



$\overline{\mathbf{CC}'}$: Baseline (translation between cameras)

$\mathbf{e}, \mathbf{e}'$: Epipole (intersection of image plane with baseline)

$\mathbf{l}, \mathbf{l}'$: Epipolar line (intersection of image plane with epipolar plane)



Fundamental matrix

- Algebraic representation of epipolar geometry
  - 3x3 Matrix
  - Maps points to epipolar lines

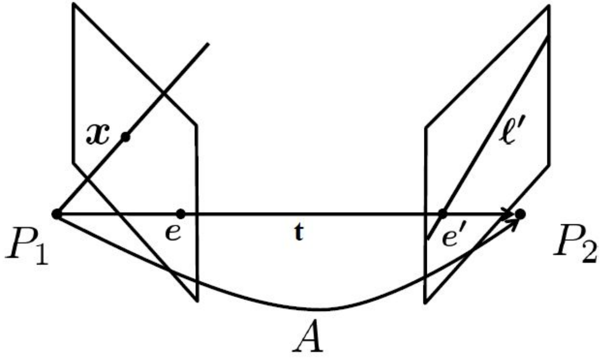
$$\boldsymbol{\ell}' = F\boldsymbol{x} \qquad \boldsymbol{\ell} = F^T\boldsymbol{x}'$$

  - Epipolar constraint $\boldsymbol{x}'^T F \boldsymbol{x} = 0$

F is the unique 3x3 rank 2 matrix that satisfies x’ ^{T}**F**x=0 for all x↔x’

Fundamental matrix encodes relative pose

$$F \quad \Leftrightarrow^{*} \quad \begin{array}{l} P_1 = [I\ \mathbf{0}] \\ P_2 = [A\ \boldsymbol{t}] \end{array}$$



Essential matrix

Relationship to **F**?

$$x_2^T F x_1 = 0 \qquad \hat{x}_2^T E \hat{x}_1 = 0 \qquad \hat{x}_i = K_i^{-1} x_i$$

$$x_2^T \underbrace{K_2^{-T} E K_1^{-1}}_{F} x_1 = 0$$

Linear equations from 5 points

$$\begin{bmatrix} x_1'x_1 & x_1'y_1 & x_{11}' & y_1'x_1 & y_1'y_1 & y_1' & x_1 & y_1 & 1 \\ x_2'x_2 & x_2'y_2 & x_{21}' & y_2'x_2 & y_2'y_2 & y_2' & x_2 & y_2 & 1 \\ x_3'x_3 & x_3'y_3 & x_{31}' & y_3'x_3 & y_3'y_3 & y_3' & x_3 & y_3 & 1 \\ x_4'x_4 & x_4'y_4 & x_{41}' & y_4'x_4 & y_4'y_4 & y_4' & x_4 & y_4 & 1 \\ x_5'x_5 & x_5'y_5 & x_{51}' & y_5'x_5 & y_5'y_5 & y_5' & x_5 & y_5 & 1 \end{bmatrix} \begin{bmatrix} E_{11} \\ E_{12} \\ E_{13} \\ E_{21} \\ E_{22} \\ E_{23} \\ E_{31} \\ E_{32} \\ E_{33} \end{bmatrix} = 0$$

4D linear solution space:

$\mathrm{E} = \mathrm{xX} + \mathrm{yY} + \mathrm{zZ} + \mathrm{wW}$ scale does not matter, choose $\mathrm{w} = 1$

IRIM

POZNAN UNIVERSITY OF TECHNOLOGY