Computer Vision

Module 7
Stereovision and depth estimation

Krzysztof Krawiec
http://www.cs.put.poznan.pl/kkrawiec/

Poznan University of Technology, 2021/2022

Module outline

1. Introduction
2. Perspective camera model
3. Canonical arrangement of epipolar cameras
4. Overview of types of depth estimation methods
5. Depth extraction from single images, including
    a. Knowledge-based methods
    b. Depth from ...' methods
    c. Methods based on active illumination
    d. New imaging techniques, including plenoptic cameras

Introduction
Depth estimation as one of the central problems in computer vision

Depth estimation

The goal: acquiring information about depth, understood (usually) as the distance of particular points of a scene from an observer (more precisely: the focal point of a camera).

Expected result: depth map

- Raster of the same size as the input image, where the pixel value is the distance* to the nearest point of the scene.
- Sometimes called "2.5-dimensional image".

Depth estimation is crucial for scene interpretation, and for this reason it is one of the most important issues within computer vision.

*A precise definition will follow.

Examples

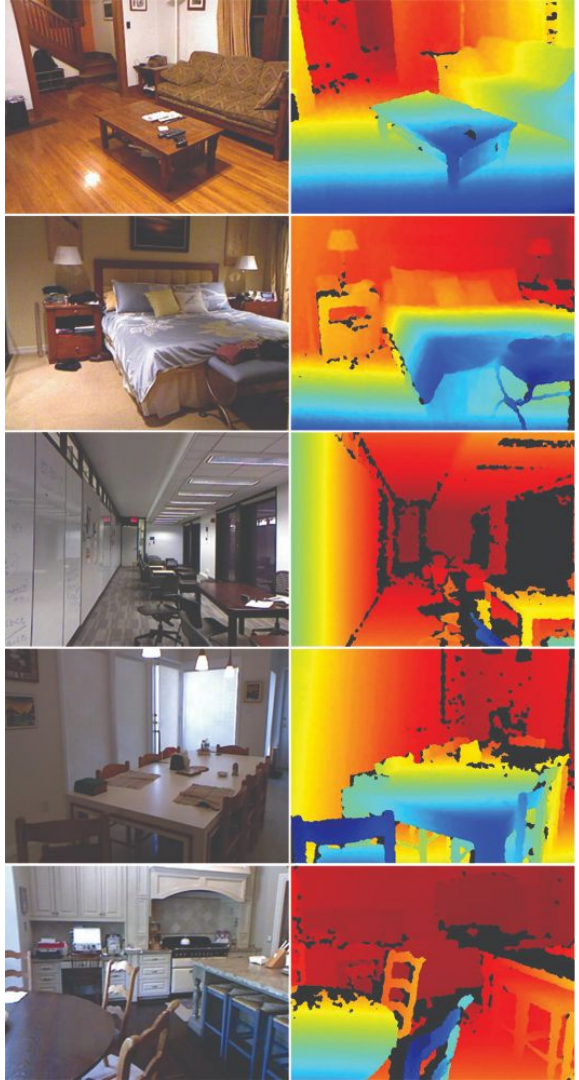
An excerpt from the NYU Depth Dataset V2.

Each example consists of:

- An image of the scene acquired with a traditional camera (RGB)
- A depth map acquired with a Kinect sensor
- [Also: scene segmentation]

Includes missing observations (black pixels due to Kinect sensor deficiencies)

Source: https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html



Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from RGBD images. In Proceedings of the 12th European conference on Computer Vision - Volume Part V (ECCV'12). Springer-Verlag, Berlin, Heidelberg, 746–760. DOI:https://doi.org/10.1007/978-3-642-33715-4_54

Depth estimation: challenges

The main challenge: mapping a three-dimensional (3D) scene onto a two-dimensional (2D) camera sensor results in an inevitable loss of information.

- For example: in perspective (perspective projection) cameras, all points on the ray coming out of the camera focus are mapped onto the same point (pixel) of the sensor.
- In general, there is no way of telling which of those points has been registered by the sensor.

Manifestations of this property:

- occlusion of some objects by others,
- incomplete visibility of three-dimensional objects.

Additional challenges: shiny surfaces (glare), shadow, reflections, …

The estimation problem is ill-posed: full recovery of 3D structure is *in general* impossible.

Depth perception in primates' visual system

Major sources of depth information:

- retinal disparity: the differences in the image of a scene as perceived by the left and right eye (to be defined more precisely soon),
- occlusion of objects.

These two sources of information complement each other:

- disparity enables depth perception for the parts of the scene that is visible to both eyes,
- occlusion assists in segmenting the scene into objects and edges.

Other sources of information:

- eye movement.

What else we know:

- color information has little effect on depth perception.

Retinal disparity

Retinal disparity is zero only for image points projected onto the so-called yellow spot (fovea centralis, fovea, the "central" element of the retina).

- For other points, the disparity is always non-zero.
- In the central nervous system (area V1 of the visual cortex) there are neurons that act as disparity detectors.

Which image features have major influence on the perception of disparity?

- zero-crossings of the second derivative of luminance (brightness) [Marr & Poggio, 1979].
- frequency components of the luminance signal [Frisby & Mayhew, 1980].
- more global features [Mallot] - it has been shown that we can successfully reconstruct depth even when luminance changes very slowly.

Perspective camera model

Types of cameras

A camera model is a specific model of projecting the 3D structure of the scene on the 2D sensor.

The main aspect differentiating cameras: the type of projection:

- Perspective camera
  - Perspective projection
  - Arguably most common, and well approximated by contemporary optics.
  - The ideal ‘implementation’: pinhole camera (camera obscura).
- Simplified perspective camera
- Orthographic camera
  - Orthogonal projection

Perspective camera

Key concepts:

O_c: focal point

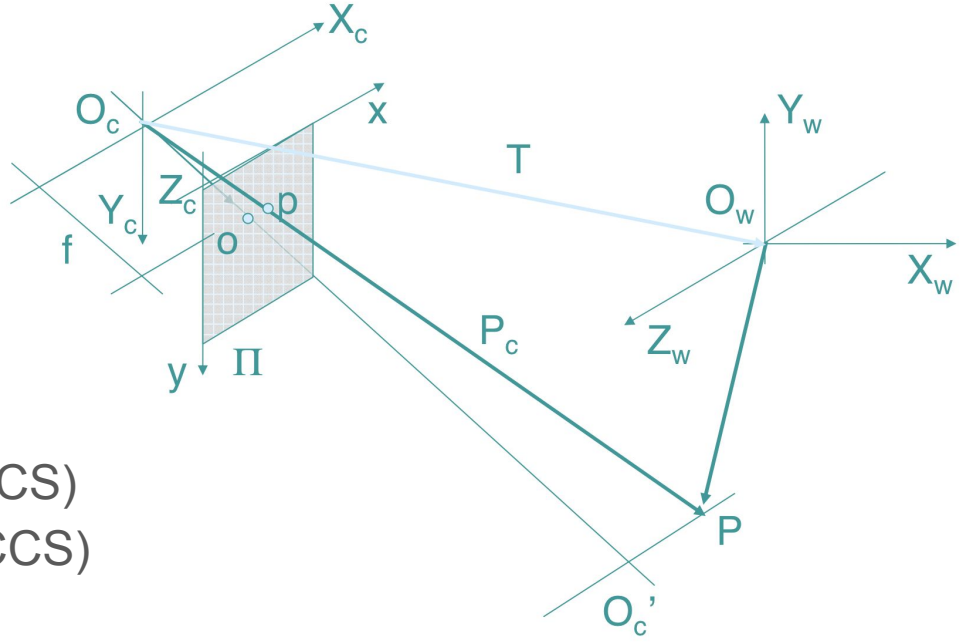
o: principal point

f: focal length

O_cO_c': optical axis

$X_wY_wZ_w$: world coordinate system (WCS)

$X_cY_cZ_c$: camera coordinate system (CCS)



All scene points on a line passing through the focus of the camera are mapped onto the same point of the imaging plane (for discrete, raster sensors: pixel).

Camera parameters

About CCS and WCS:

- In simple cases, the camera and world coordinate systems can be assumed to spatially coincide.
- However, in most cases (motion of the camera relative to the scene, or vice versa, presence of multiple cameras) distinguishing these systems from each other is essential.

Camera parameters can be divided into:

- external:
  Define the relationship of the camera coordinate system to the world coordinate system.
- internal:
  Determine how the scene is projected onto the camera sensor.

External camera parameters

Transition from the world coordinate system to the camera coordinate system: a composition of translation and rotation:

$$P_c = R(P_w - T) \qquad T = O_w - O_c = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix}$$

where:

- P_w: coordinates of scene point P in the world coordinate system
- P_c: coordinates of scene point P in the camera coordinate system
- T: translation vector
- R: rotation matrix (3x3)

Internal camera parameters

- The perspective camera has one main parameter: the focal length f.
- Translation between camera coordinates and image coordinates based on:
  - the location of the principal point $o=(o_x,o_y)$ in the sensor, and
  - pixel dimensions s_x and s_y:

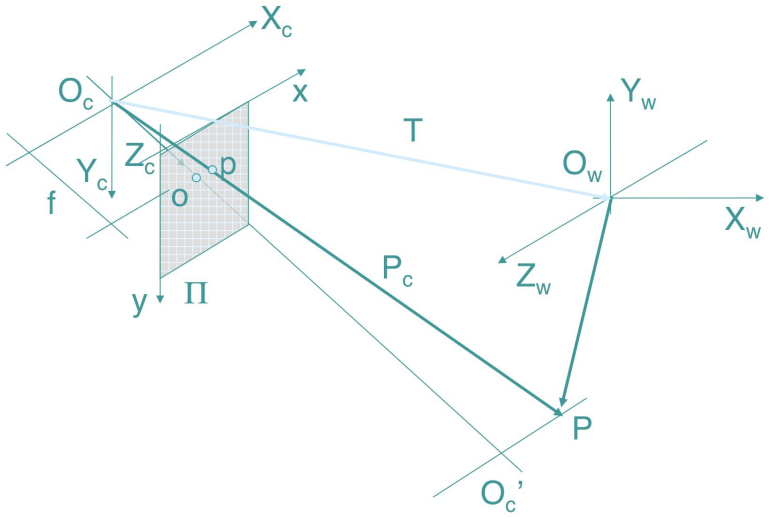
$$x = (x_u - o_x)s_x, y = (y_u - o_y)s_y$$

- For comparison: orthographic camera has no internal parameters

Projection of scene points

A point P=(X,Y,Z) in the camera coordinate system is given.
The coordinates of its projection p=(x,y,z) in the camera detector can be derived from the similarity of triangles $\Delta O_c po$ and $\Delta O_c PO'_c$.

$$x = f\frac{X}{Z}, y = f\frac{Y}{Z}, z = f$$



- The Z coordinate (depth)
  affects the x and y coordinates
  - with X and Y fixed,
    a change in Z implies changes in x and y.
  - with Z fixed,
    a change in X and Y implies changes in x and y.
- Key observation: given an observed point
  (x,y) in the raster, all we need to determine (X,Y) is Z.
- These observations justify the meaningfulness of stereovision approaches.

Digression: Simplified perspective camera

When X and Y are much smaller than the average Z (i.e., that the object size is significantly smaller than its distance from the camera), the earlier formula can be reduced to:

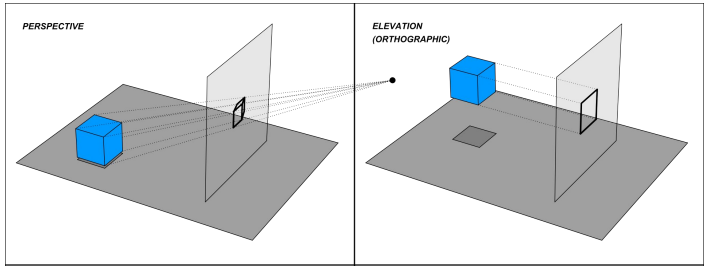
$$x = f\frac{X}{Z}, y = f\frac{Y}{Z}, z = f$$

which is *de facto* scaling of X and Y by a constant.
[A yet another related camera model: the affine camera model]

Another well-known camera (projection) model: orthographic camera (so-called orthophoto, parallel projection):

$$x = X, y = Y, z = f$$

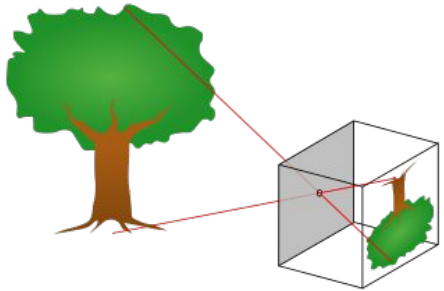
Used only at very large distances of the observer from the scene (compared to the size of the objects), e.g. aerial and satellite imaging.



Perspective camera: technical realization

The perfect realization of the perspective camera is the pinhole camera (camera obscura).

- However, perfect sharpness requires infinitely small opening (hole).
  - This in turn extends the exposure time to infinity.
- In practice, we try to approximate
  the properties of the pinhole camera using optics.



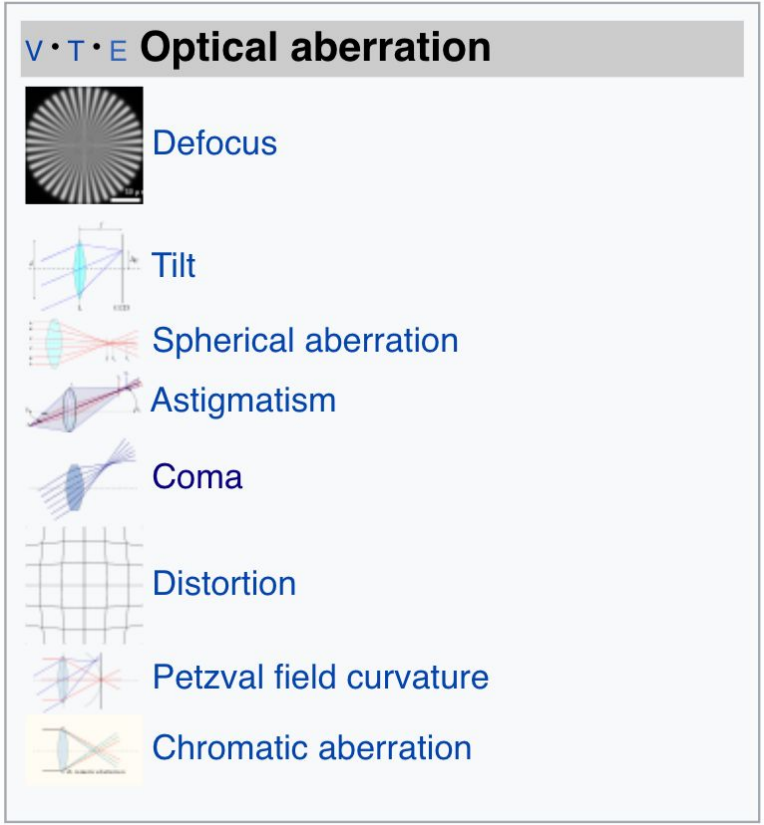
Unfortunately, no lens is perfect:

- Distortions introduced by the camera's optical system
  - result from the imperfections of the optical system (lenses, mirrors):
    - geometric aberrations: spherical aberration, coma (comatic aberration), astigmatism, field of view curvature, pillow and barrel distortion; occur at both lenses and mirrors.
    - chromatic aberrations: variation of light ray path in the function of wavelength; occur only at lenses.

Perspective camera: technical realization

Geometric aberrations occur because classical optics works only at [infinitesimally] small angles w.r.t. optical axis, and in general it is impossible to construct an optical system in which all rays originating at a given point in the scene end up at the same location in the sensor.

- (That is possible though under certain assumptions, e.g. imaging only selected areas in the scene, e.g. a single plane or two planes).



Recommended reading:
https://en.wikipedia.org/wiki/Optical_aberration

Stereovision

Stereovision

The use of two images of the same scene (*stereo pair*) to acquire information about depth in a scene, based on (mainly) analysis of the disparity between the images.

Advantages:

- relatively cheap
- analogous to human perception

Disadvantages:

- requires two cameras (usually),
- cameras must be accurately calibrated
  - E.g. the optical axes of cameras must be precisely parallel (in some settings).
- potentially high computational cost of disparity analysis algorithms.

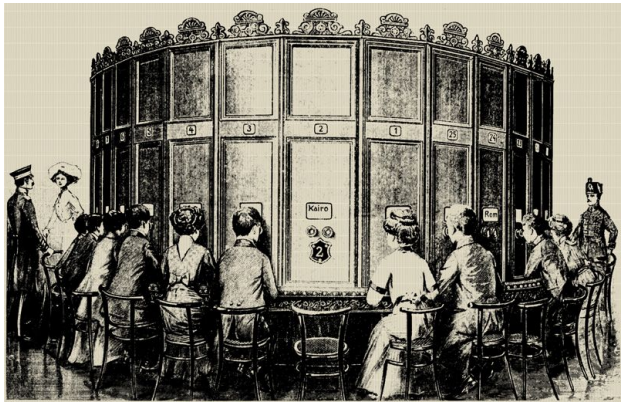
Canonical camera arrangement (simple stereo)

The canonical camera arrangement comprises two perspective cameras

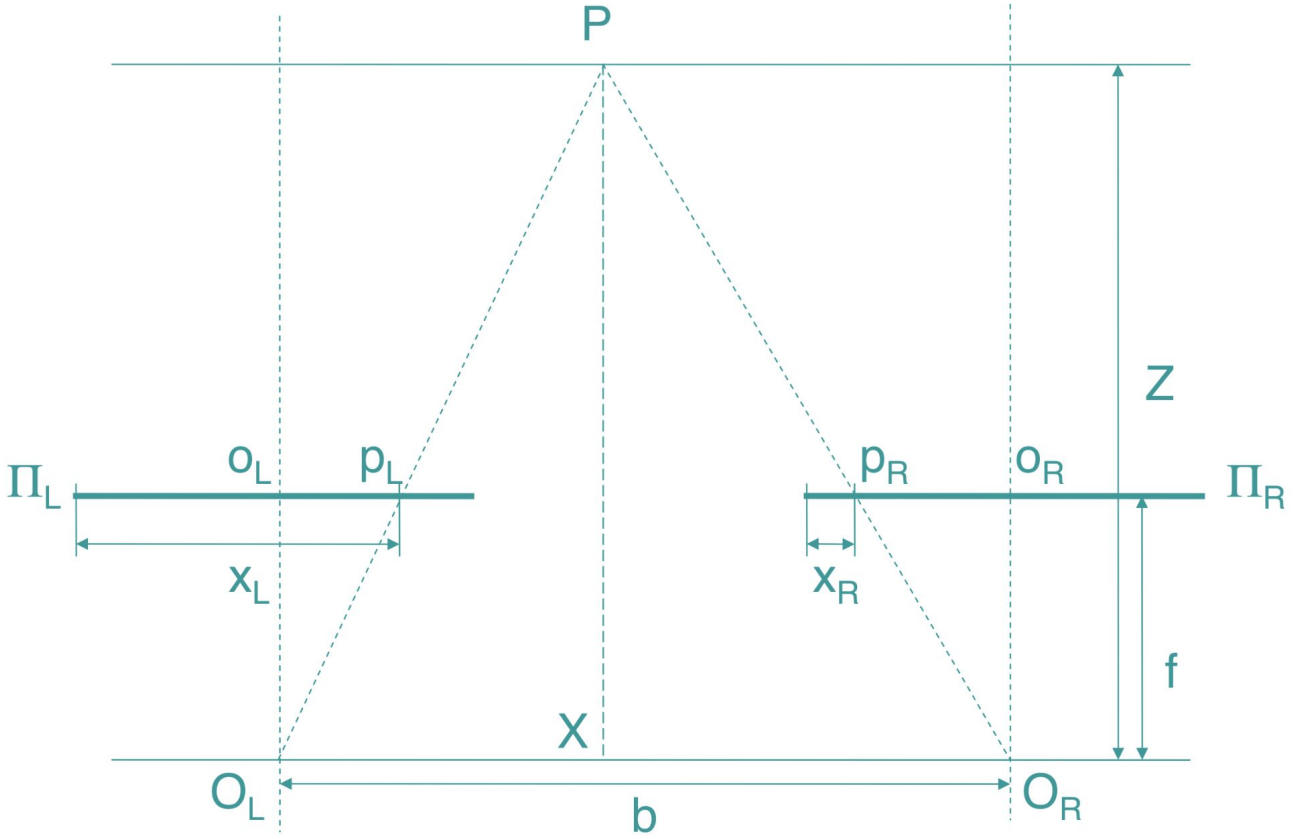
- with the same internal parameters,
- with parallel optical axes,
- placed at a distance b (base) from each other (distance of camera focal lengths).

It has been used for a long time in stereovision cameras

- Purpose: stereoscopy (to create the illusion of depth)
- Acquired images used in projection devices
- Historically: e.g. so-called Kaiser-Panorama



Canonical camera arrangement (simple stereo)



Canonical camera arrangement (simple stereo)

Legend to the previous slide:

- Π_L, Π_R: left and right projection planes (sensors)
- P: a point in the scene
- p_L, p_R: the projections of P on the sensors of the left and right camera
- x_L, x_R: the observed coordinates X of points p_L and p_R
- b: the length of the baseline, the segment connecting the focal points.
- Z: depth (the unknown variable, to be estimated)

Disparity: the difference in the observed position of the projections of point P on the camera sensors (horizontal disparity, in the X axis):

$$D_x(p_L, p_R) = x_L - x_R$$

Disparity (horizontal disparity)

Using the similarity of the triangles $\Delta p_L o_L O_L$ and ΔPXO_L, as well as $\Delta p_R o_R O_R$ and ΔPXO_R, the disparity can be determined as follows:

$$D_x(p_L, p_R) = x_L - x_R = \frac{bf}{Z}$$

Implication: knowing the disparity allows determining Z.

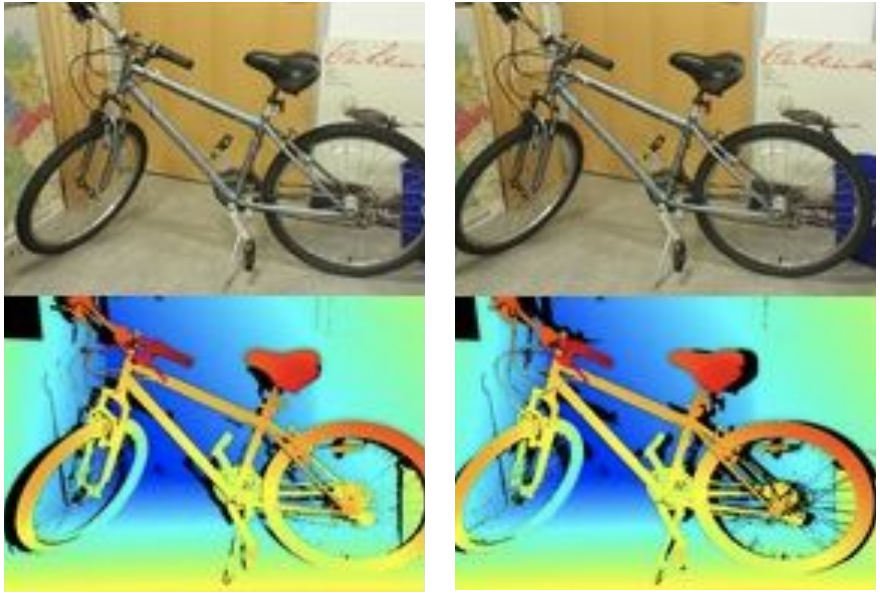
One can also define the vertical disparity D_y, and then aggregate D_x and D_y to the total disparity.

Example 1

Many test databases of stereo pairs with corresponding disparity have been developed. Below: an example from the Middlebury Stereo Datasets.

Legend:

- Left camera image
- Right camera image
- Ground truth images (disparity maps) obtained using a technique based on structured lighting.

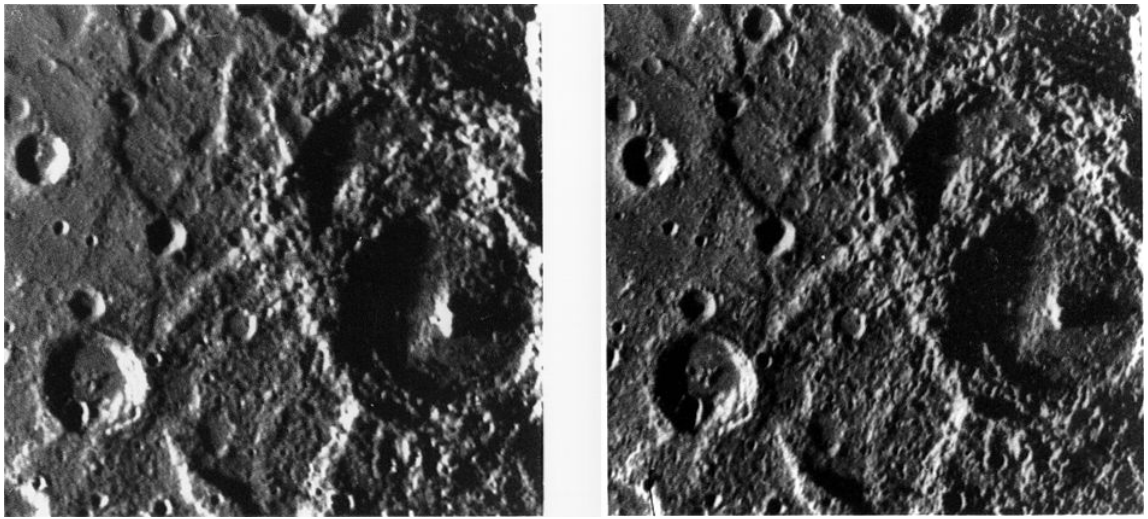


https://vision.middlebury.edu/stereo/data/scenes2014/

D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In German Conference on Pattern Recognition (GCPR 2014), Münster, Germany, September 2014.

Example 2

Stereovision does not always require two cameras: below is a stereo pair of the Boccacio crater (Mercury, 135km in diameter) taken by the Mariner 10 spacecraft. The images were taken as a result of two separate flybys over the planet's surface.



Source: NASA, https://nssdc.gsfc.nasa.gov/imgcat/html/object_page/m10_aom_15_11.html

Digression: Random stereograms

An excellent illustration of the usefulness of retinal disparity for human depth perception.

The simplest algorithm (more sophisticated ones also exist):

1. Create a random (salt-and pepper) image L
2. Copy L to R
3. Shift a selected area in R by a small vector (horizontally).
4. Represent (L, R) as a stereo pair.

Alternative interpretation: a scene composed of white 3D objects is being sprinkled with pepper.

Gazing at (L, R) from the ‘right’ distance gives the illusion of depth.

See https://en.wikipedia.org/wiki/Random_dot_stereogram https://en.wikipedia.org/wiki/B%C3%A9la_Julesz

Digression: Random stereograms

Top: two copies of the same random image; highlighted area in the right image.
Bottom: the highlighted area shifted 2 pixels to the right.

The problem of correspondence

The main challenge in stereovision

Knowing the disparity allows us to immediately determine Z (given known internal camera parameters). This does not pose a problem.

The challenge is the **correspondence problem**: determining the correspondence of the projections p_L and p_R of the point P.

- Note that we observe only p_L and p_R (more precisely: the entire projections).
- How can we be sure that a given p_L in the left projection corresponds to a candidate p_R in the right projection?

The task: find such pairs of points (p_L,p_R) that with high probability are projections of the same point P of the scene, whereby:

- A given p_L may not correspond to any p_R (e.g., as a result of occlusion).
- The algorithm should have an acceptable computational effort.

Main classes of algorithms

- Direct:
  - Disparity calculated directly from brightness (preliminarly filtered to eliminate noise).
  - Result: dense disparity map (disparity estimate available for each image point).
- Feature-based:
  - Disparity calculated on features/keypoints identified in the image (e.g., lines, corners).
  - Result: sparse disparity map (disparity estimated only at keypoint locations).

Another line of division between algorithms:

- iterating over points:
  - "priority for space": analyze consecutive image patches/sub-windows, and look for disparities in each patch (usually starting with the smallest disparity)
- iterating over disparities:
  - "priority for disparity": reverses the nested loops: for different expected disparities, search for matches across entire images.

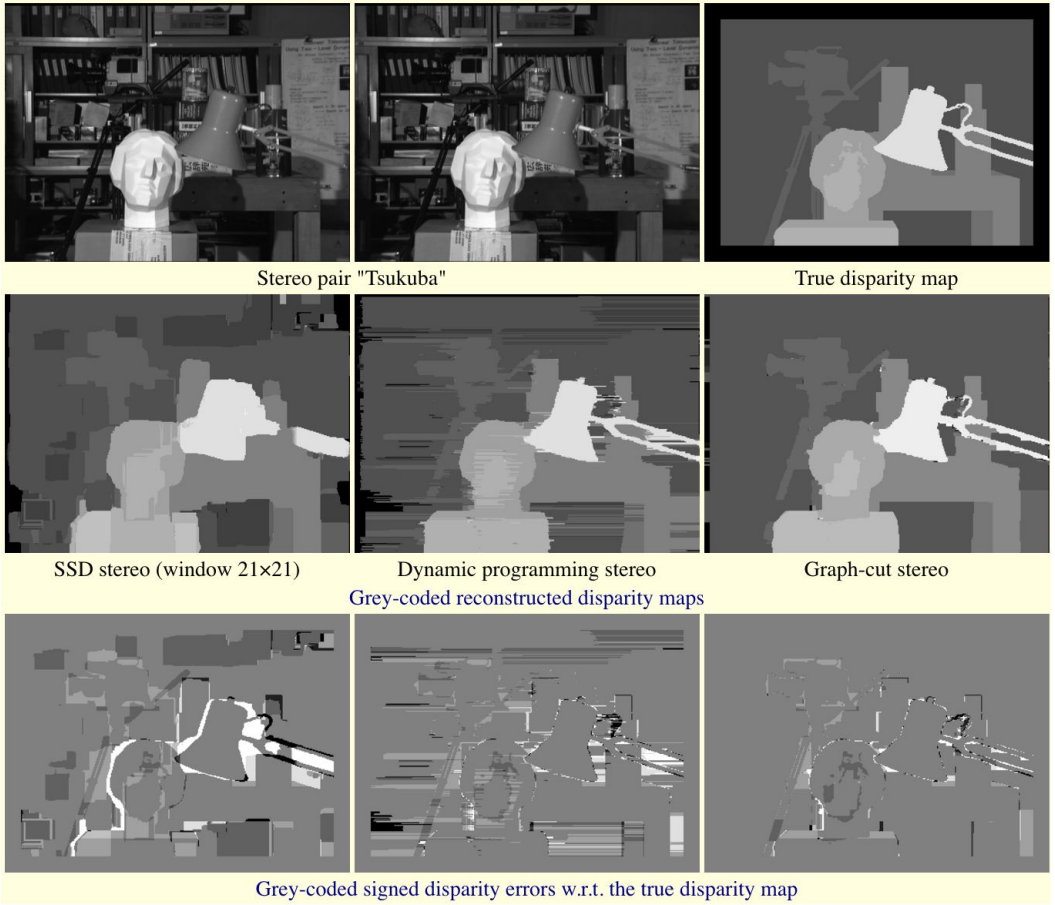
Concepts and formalisms used for correspondence

The number of

- correlation of image patches,
- energy functions, diffusion and relaxation (of ambiguities of matching),
- dynamic programming,
- gradient,
- probabilistic formulations.

Example

Depth estimation algorithms are evaluated by rigorously comparing them with other algorithms, based on reference datasets containing stereo pairs and target depth information ("true disparity map").



http://www.middlebury.edu/stereo

D.Scharstein and R.Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", Int. J. Computer Vision, vol.47 (1/2/3), pp.7-42, April-June 2002

Application areas

- robotics
- autonomous systems
- 3D object reconstruction
- 3D object recognition
  - usually model-based: similarity measurement of depth map and model in 3 dimensions
- cartography, climatology
  - interesting example: ocean wave analysis
- multimedia
  - image synthesis, e.g.
    - superimposing synthetic images on real ones
    - generation of unknown object views from known views

Selected factors and recommendations

- Image noise level:
  - High noise levels may disqualify direct methods; on the other hand, keypoints/features may remain largely unaffected by noise, and thus the indirect methods may fare better.
- Distance from imaged objects (and, consequently, range of disparity)
  - For instance, in satellite imagery objects are located very far from the cameras, so the disparity is relatively small, which may favor the direct methods.
- Computational effort required
  - Direct methods that create dense disparity maps are often computationally more demanding.
- Required precision/admissible error
  - Keypoint-based, indirect methods will usually only interpolate the estimates for locations between keypoints, and as such are typically less precise.

Depth estimation from a single image

Depth estimation from a single image

Stereovision is not the only depth estimation method.

Alternative approaches are based on, among others:

1. Knowledge-based methods
    a. Including methods based on machine learning.
2. Structured lighting
3. ‘Depth-from-X’ methods, including:
    a. Methods based on image focus/defocus
    b. Methods that extract depth from other ‘aspects’ of the scene (e.g. shadows or motion)
4. Specialized imaging methods
    a. Laser, radar, ultrasonic (sonar) rangefinders: measure the time taken for a reflected signal to return or change phase
    b. Newer techniques, such as light field imaging.

Knowledge-based methods

Hypothesis: A single image of a scene contains (usually) enough information to restore significant part of depth map.

Rationale:

- Images of real scenes are not entirely random: scenes consist of physical objects that have some universal properties, e.g:
  - they are often compact/continuous,
  - they are relatively well distinguishable from their surroundings,
  - may feature many straight (or nearly straight) edges - especially for man-made objects,
  - in the case of video scene analysis: they are subject to certain universal physical laws, e.g. characterized by inertia (e.g. objects often move along straight lines, cannot stop in an arbitrarily short time, they have permanence, etc.)

The methods in this category exploit these premises.

Knowledge-based method: An example (1)

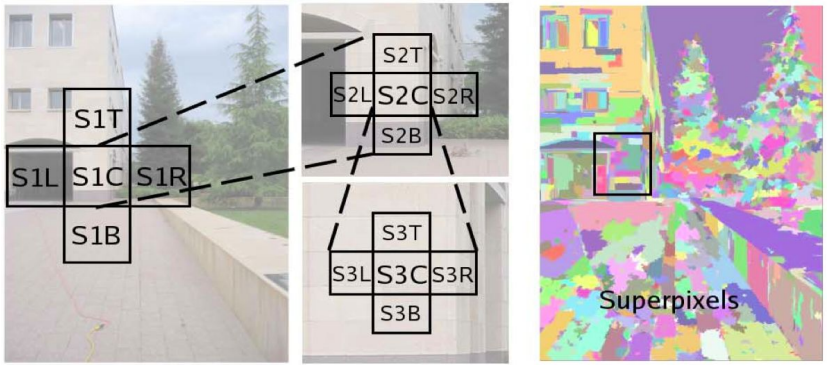
Assumption: the estimated scene model consists of multiple interconnected planes (facets). Steps of the method:

1. Segmentation into multiple small areas (*oversegmentation*), resulting in *'superpixels'*.
2. Extraction of superpixel features, subject to constraints:
   a. Adjacent superpixels often represent interconnected (contiguous) planes in a scene
   b. Adjacent superpixels often represent the same plane
   c. Straight lines in the 2D image correspond to straight lines in the scene
3. Construction and simulation of a spatial, stochastic Markov model that links superpixels together (Markov Random Field)
   a. The variables are the parameters of the 3D planes corresponding to the superpixels in 2D

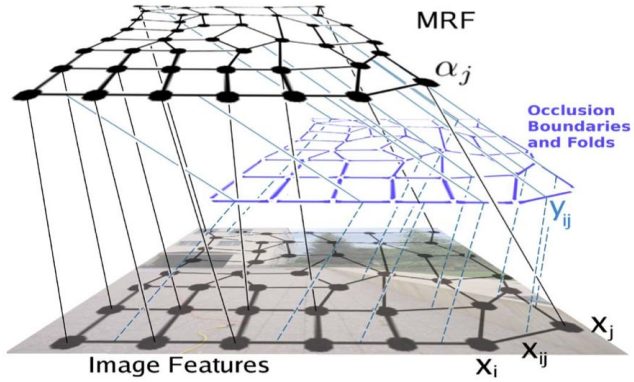
Saxena A, Sun M, Ng AY. Make3D: learning 3D scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009 May;31(5):824-840. DOI: 10.1109/tpami.2008.132.

Knowledge-based method: An example (2)

The acquired image, plane neighborhood relationship in the model, and result of image segmentation (pseudocolored superpixels).



Visualization of the Markov model of plane parameters in terms of 'explaining' the observed segmentation result.



Saxena A, Sun M, Ng AY. Make3D: learning 3D scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009 May;31(5):824-840. DOI: 10.1109/tpami.2008.132.

Knowledge-based method: An example (3)

Input image, the target (true) depth map (pseudo-colored), and depth estimation obtained with three different variants of the proposed method.

- The last image on the right illustrates the performance of the full variant of the method.

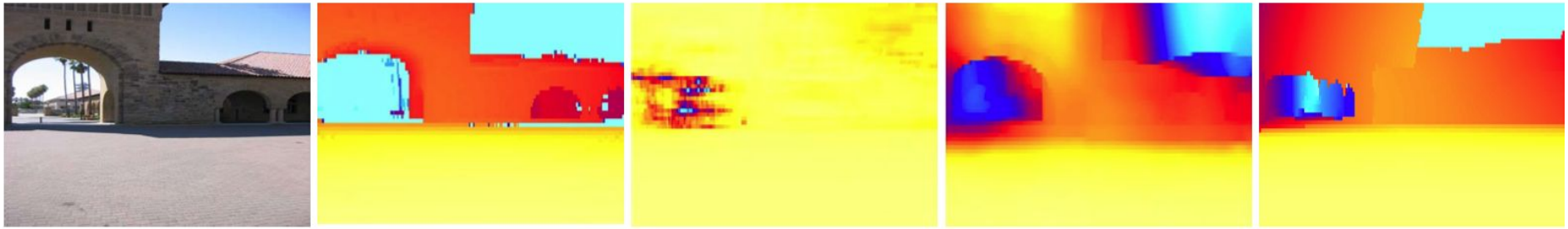
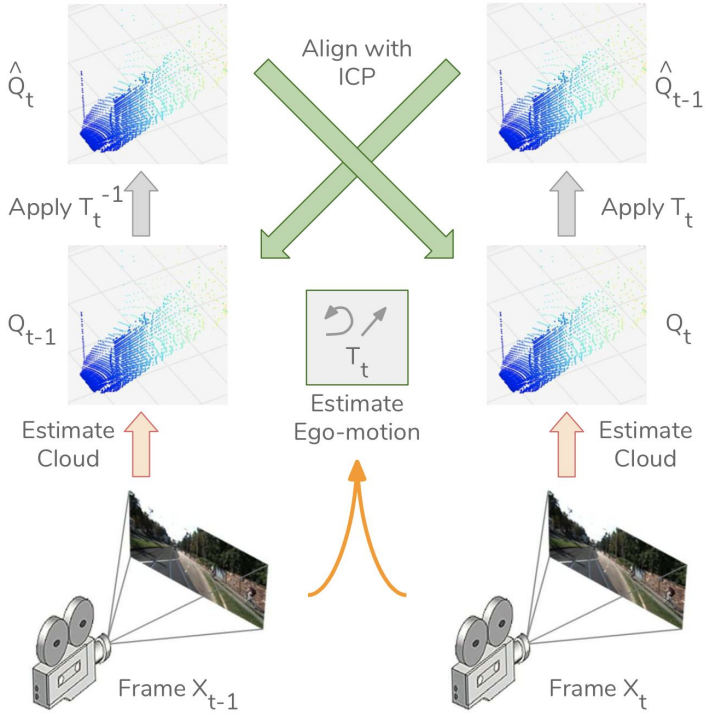


Figure 7. (a) Original Image, (b) Ground truth depthmap, (c) Depth from image features only, (d) Point-wise MRF, (e) Plane parameter MRF. (*Best viewed in Color*)

Saxena A, Sun M, Ng AY. Make3D: learning 3D scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009 May;31(5):824-840. DOI: 10.1109/tpami.2008.132.

Deep learning for depth estimation (an example)

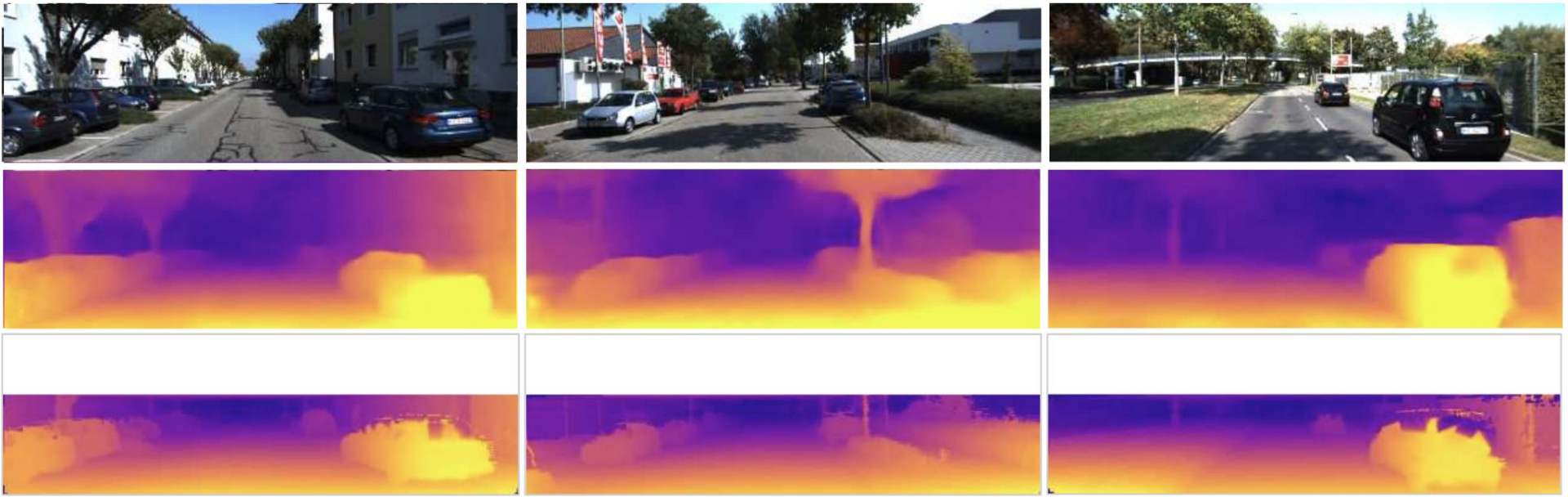
- Extracting depth in an unsupervised manner, from video sequences
- Simultaneous estimation of depth and *ego-motion*
- Uses a specialized loss function that simultaneously penalizes inconsistencies in motion and depth estimation.
- Loss function used simultaneously in 'forward' (frames (t,t+1)) and 'backward' (frames (t+1,t)) modes.



Mahjourian, R., Martin Wicke and A. Angelova. “Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 5667-5675. https://arxiv.org/abs/1802.05522

Exemplary results

- First row: input images
- Second row: the estimates obtained by the proposed approach
- Third row: actual depth (ground truth)



‘Depth-from-X’ methods

Methods based on image sharpness (focus)

- Depth from focus - determining the distance to objects/points in the scene by changing the focal length and testing the focus.
  - Sharpness in a part of the image can be measured by measuring the contribution of high spatial frequency components (even with edge detection filters).
  - Disadvantages: requires control of focal length, which in turn usually requires physical/mechanical control of the camera lens – time consuming (unacceptable in video sequence analysis).
- Depth from defocus:
  - Relies on the observation that an unfocused point in the scene is being projected as a disk. That disk can be described by a model, and by extracting the parameters of this model, the distance of the point can be determined (provided the optical parameters of the camera are known).
  - Mathematical analysis shows that just two images (with different defocusing) are enough to obtain the depth map.

Estimating depth from other image ‘aspects’

Methods extracting depth from other ‘aspects’ of scenes

May be collectively titled 'depth from X'

Examples:

- Depth from shading (structural analysis of shadow)
  - Use of light reflection/scattering characteristics to extract depth information
  - Disadvantages: requires strong assumptions, high computational complexity.
- Depth from motion
  - Exploiting changes in image content as the camera moves
  - In a sense, a generalization of stereoscopy

Controlled lighting

Controlled lighting

More precisely: structural lighting.

- Consists in projecting onto the scene a precise, regular pattern of known spatial characteristics, usually a line or a grid.
- As a result of interaction with scene elements, the pattern is ‘distorted’
- Knowing the characteristics of the pattern together with the acquired image enables depth reconstruction.

See for instance:

- An interesting tutorial on types of structured lighting: https://www.osapublishing.org/aop/fulltext.cfm?uri=aop-3-2-128&id=211561
- Examples of structured lighting emitters: https://www.stemmer-imaging.com/en/knowledge-base/structured-lighting/

Structural lighting

Advantages:

- often very precise,
- usually also fast.

Disadvantages: needs to use a source of light/radiation/energy:

- this is additional piece of equipment, often quite expensive,
- certain bands of the EM spectrum are harmful for humans (e.g. UV),
- additional energy/power consumption,
- excluded in some applications:
  - when the distance from observed objects is too high,
  - when the imaging system has to be passive (e.g. military, security).

New/alternative imaging techniques

Light-field cameras

Light-field cameras (light field photography, plenoptic photography)
Idea:

- Traditional cameras only record the amount of light arriving to a given sensor element (pixel).
- Light-field cameras register also the direction of the incident light rays (the vector).
    - I.e. the term 'field' is used here in the mathematical sense: *a vector field*.

This effect is usually achieved with so-called plenoptic cameras. A plenopti camera can be conceptually considered as a (usually large) number of small cameras (with overlapping sensors).

- Technically: a micro lens array projects an image onto a detector array.
- The reconstruction of the image is done by Fourier transform.

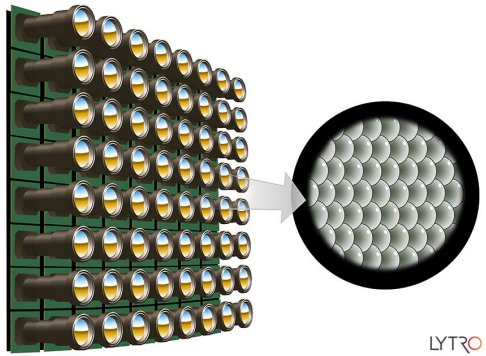
Advantages of plenoptic cameras

- High sensitivity (due to large aperture).
- No need to set the focus before taking an image, so less time preparing the camera for image acquisition.
- The *plane of focus* (PoF) can be changed after the picture has been taken (!)
- Allow reconstructing three-dimensional scenes.

The pioneer of plenoptic cameras: Lytro

The first manufacturer of of commercial plenoptic cameras (until 2018).

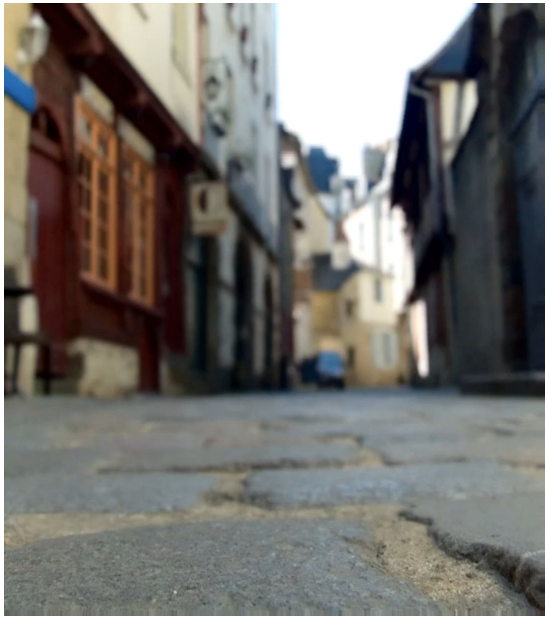
- Images saved in a special LFP (light field picture) format.
- Specialized viewer required (rendering the image for a given plane of focus)



Multi-camera arrays were the first Light Field capture systems, developed at Stanford University. This massive array has been replicated and miniaturized with the use of microscopic lenselets (microlens array) placed above the sensor in Lytro's ILLUM to capture the Light Field with a single camera. Illustration not to scale.

An image from plenoptic camera (Lytro): Example

The following three images are visualizations of the same image (acquisition) with a plenoptic camera, for different focal lengths.



A. Mousnier, E. Vural, C. Guillemot, Lytro first generation dataset, https://www.irisa.fr/temics/demos/lightField/index.html

Plenoptic cameras: summary

- Lytro went out of business in 2018.
- The technology has spread to other centers.
- See, among others: https://raytrix.de/