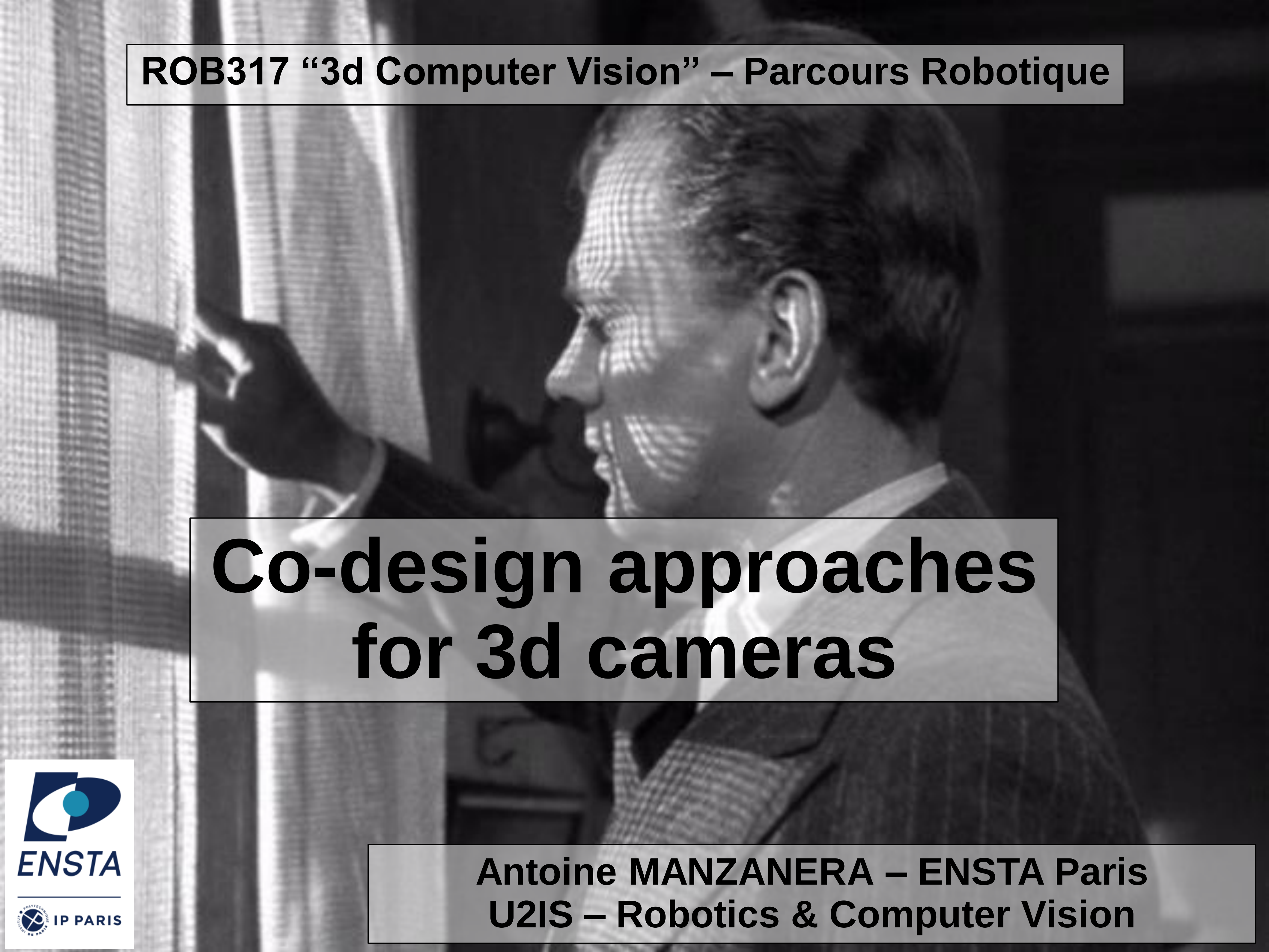


ROB317 “3d Computer Vision” – Parcours Robotique

# Co-design approaches for 3d cameras

Antoine MANZANERA – ENSTA Paris
U2IS – Robotics & Computer Vision

# Objectives of this lecture:

- Getting a global view of the bio-inspired and/or co-design opportunistic approaches, making the most of the different parts of a computer vision system (optics / mechanics / electronics / software) to increase its perception and analysis capabilities.

- Understanding the principle of the main categories of co-design approaches for the perception of 3d by a computer vision system.

**Co-design approaches for the perception of 3d using digital cameras**

# Outline of this lecture:

- **Part 1: Bio-inspiration?**
  - Stereovision and multi-view 3d
  - Other 3d cues
  - Learning based approaches
- **Part 2: Active 3d**
  - Time of flight
  - Structured light
- **Part 3: Passive 3d**
  - Plenoptic cameras
  - Depth from (de)focus
  - Coded aperture

# Part 1: BIO-INSPIRATION?

Almost all evolved biological vision systems use two eyes or more.

Stereovision is the main mechanism used by human for 3d sensing.

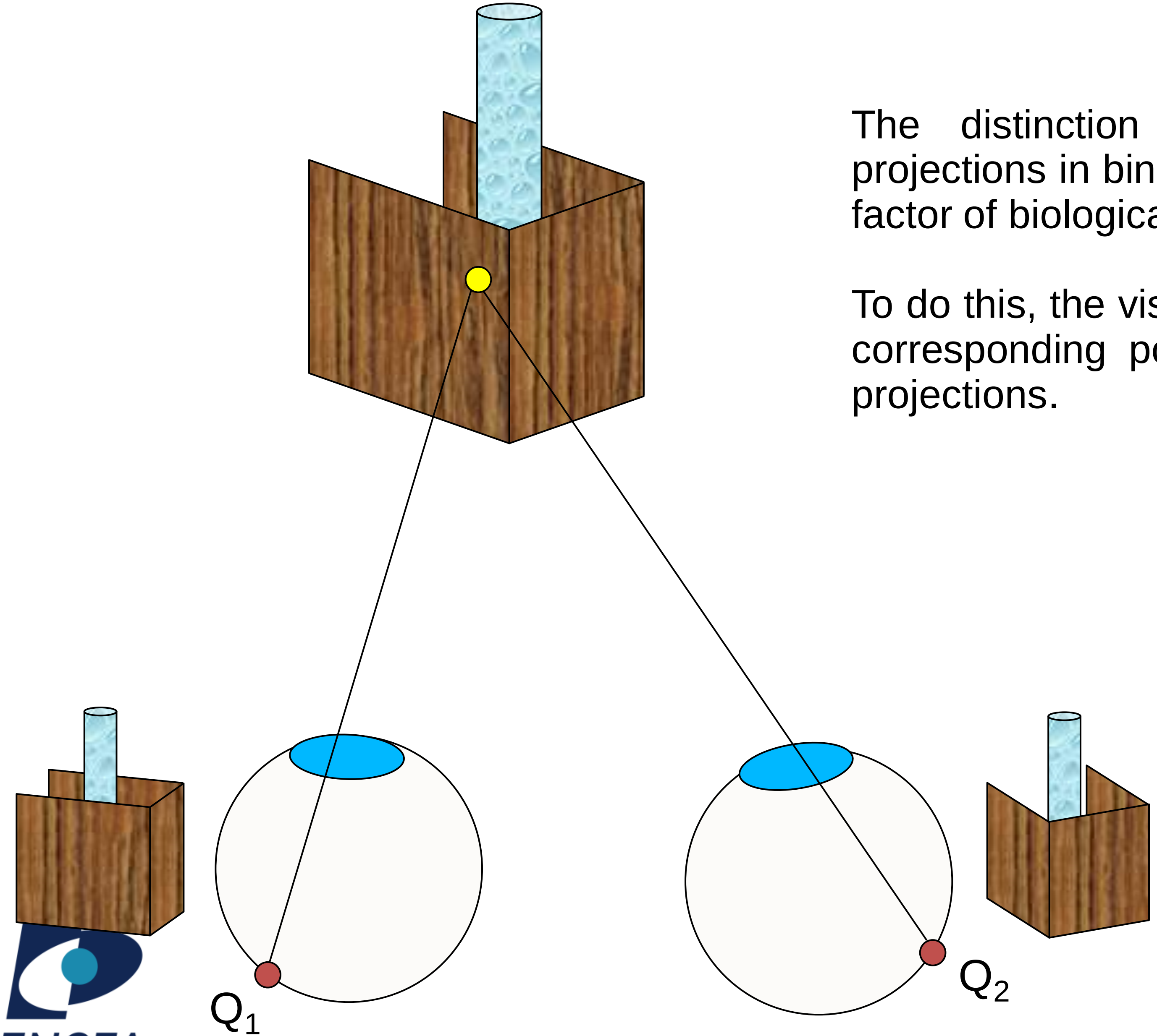
However, for some animals, the two fields of view do not overlap: differential motion parallax is used instead.

Stereovision and structure from motion are popular techniques for computer vision based 3d reconstruction algorithms.

But many other vision cues are also used by humans for 3d sensing, that may be exploited by future co-design systems.

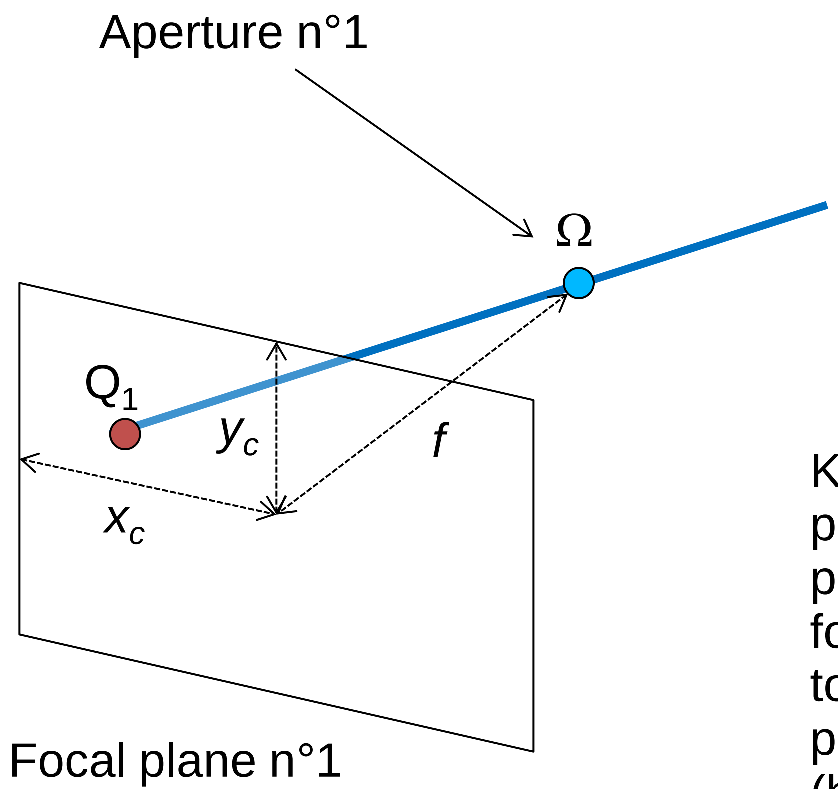
Those different cues are probably readily used by the learning based (CNN) approaches, though their interpretation is a tricky issue.

# VISUAL MATCHING AND 3D PERCEPTION



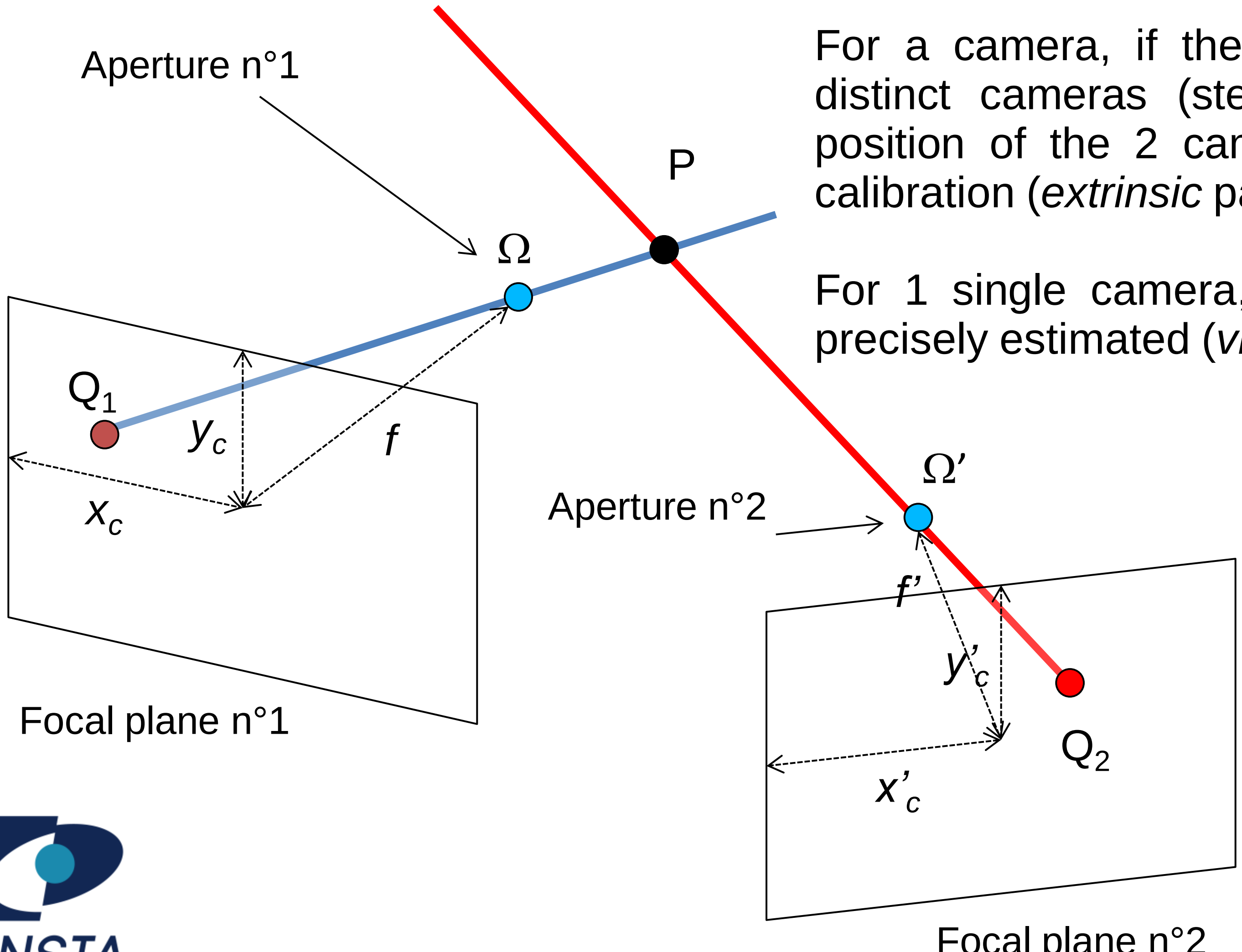

The distinction between left and right projections in binocular vision is an essential factor of biological 3d perception.

To do this, the visual system must match the corresponding points of the two perceived projections.




Knowing the geometry of the first focal plane, i.e. the position of the optical centre $(x_c, y_c)$, which is the projection of the aperture on the focal plane, and the focal distance $f$, which is the distance of the aperture to the focal plane, the optical path of every point $Q_1$ projected on the focal plane can be back-traced (back-projection of point $Q_1$ in blue).

For a camera, these so-called *intrinsic* parameters are estimated by calibration.




The same applies for a 2[d] focal plane. If $Q_1$ and $Q_2$ correspond to the same point (matching) then their back-projections intersect at this point P.

For a camera, if the 2 focal planes belong to 2 distinct cameras (stereovision), then the relative position of the 2 cameras must be estimated by calibration (*extrinsic* parameters).

For 1 single camera, its displacement has to be precisely estimated (*visual odometry*).

We then get:

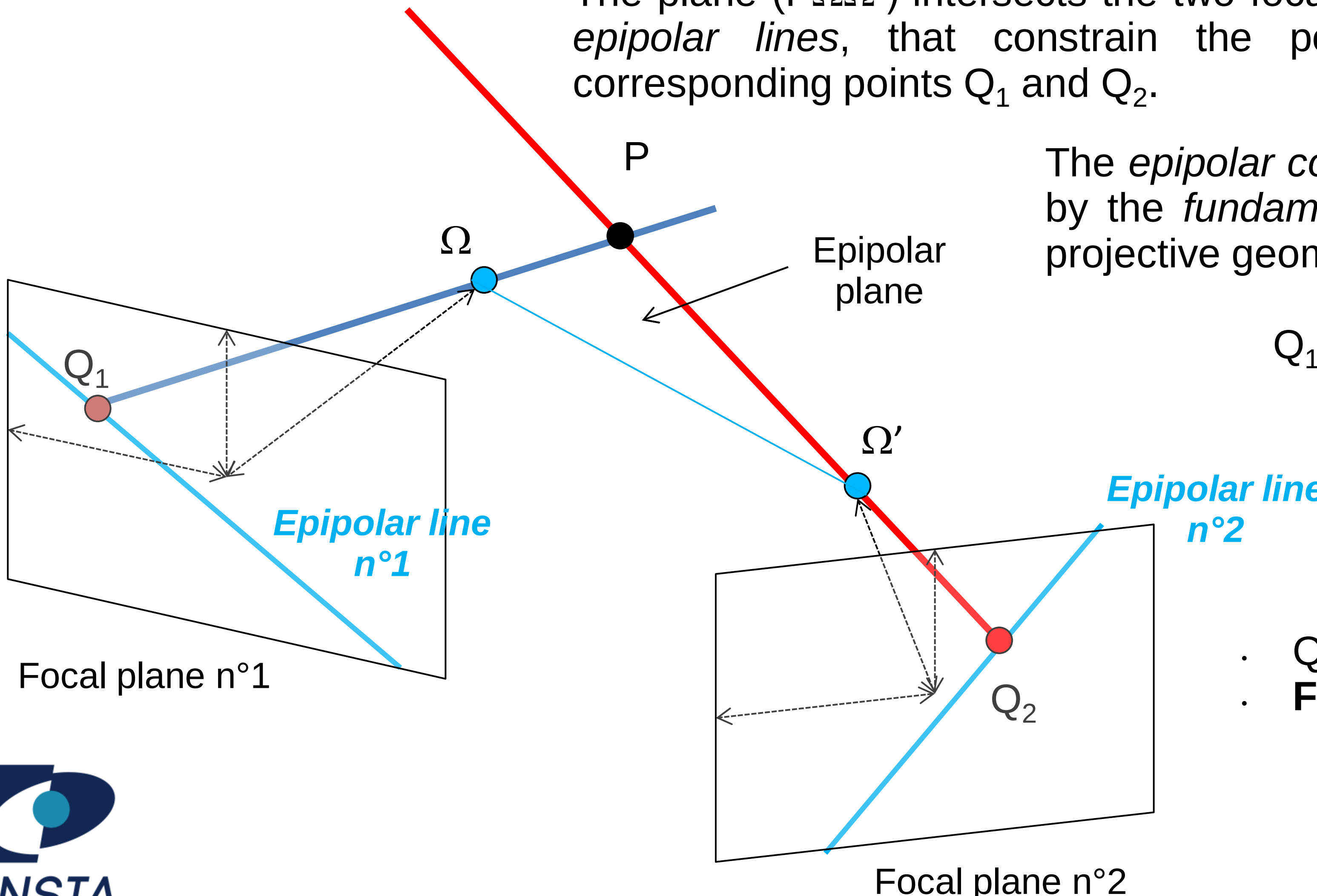
$$\Omega P = \Omega\Omega' \frac{\sin \widehat{\Omega}'}{\sin \widehat{P}}$$

$$\Omega' P = \Omega\Omega' \frac{\sin \widehat{\Omega}}{\sin \widehat{P}}$$

# STEREOVISION AND STRUCTURE FROM MOTION

The two lines corresponding to the optical paths are intersecting, and then coplanar.

The plane (PΩΩ’) intersects the two focal planes on so-called *epipolar lines*, that constrain the possible positions of corresponding points $Q_1$ and $Q_2$.

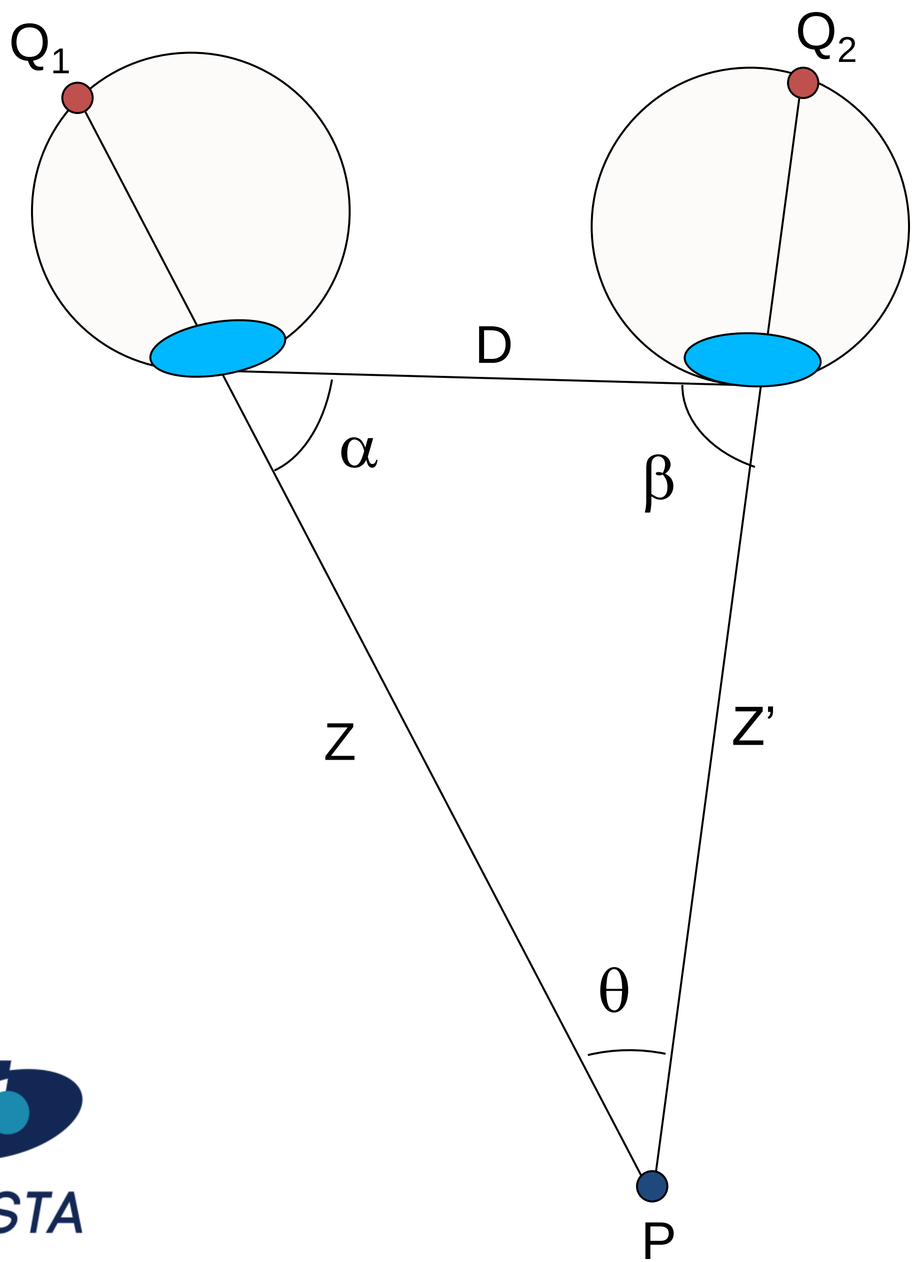



The *epipolar constraint* is expressed by the *fundamental matrix* **F** in the projective geometry framework:

$$Q_1 \mathbf{F} Q_2 = 0$$

- $Q_1$ **F**: Epipolar line #2
- **F** $Q_2$: Epipolar line #1

# DEPTH AND THE BINOCULAR VERGENCE



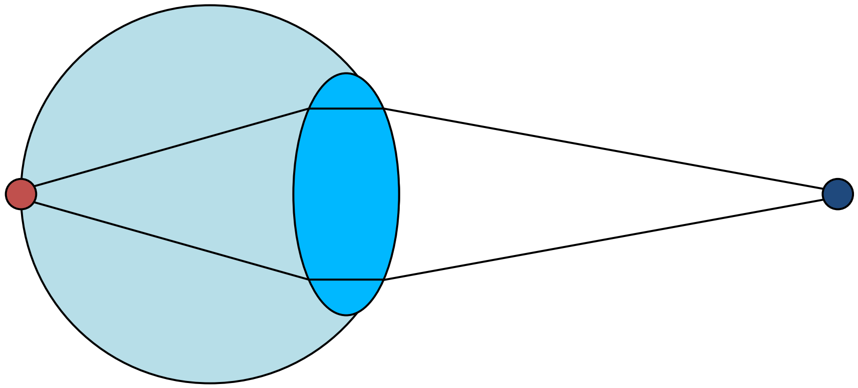

Triangulation principle:

$$Z = D\frac{\sin\beta}{\sin(\alpha+\beta)}$$
$$Z' = D\frac{\sin\alpha}{\sin(\alpha+\beta)}$$

Vergence angle: $\theta = \pi - (\alpha+\beta)$

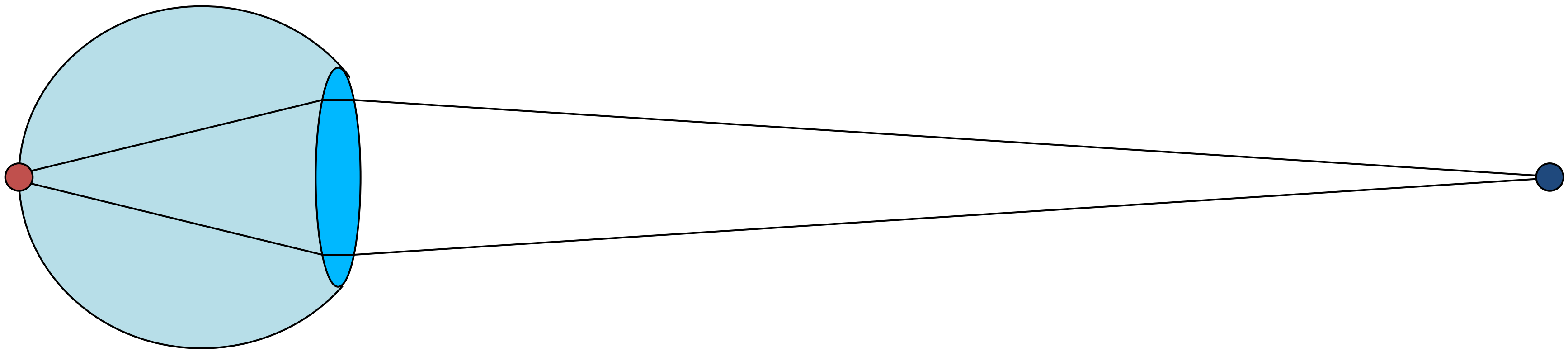
$$Z = D\frac{\sin\beta}{\sin\theta}$$
$$Z' = D\frac{\sin\alpha}{\sin\theta}$$

# DEPTH AND ACCOMMODATION (MONOCULAR)



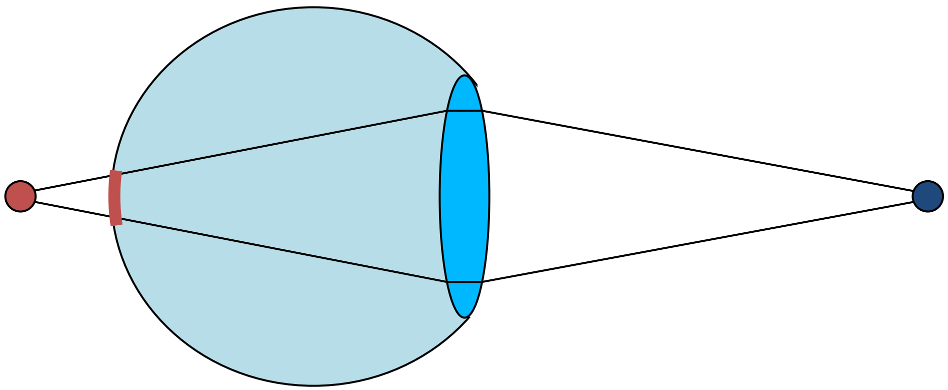
The accommodation mechanism consists in deforming the eye lens to adjust its focal in such a way that the image of the focalised object appears sharp on the retina.

*Short focal: near object appears sharp on the retina plane*



*Long focal: far object appears sharp on the retina plane*

# DEPTH AND ACCOMMODATION (MONOCULAR)

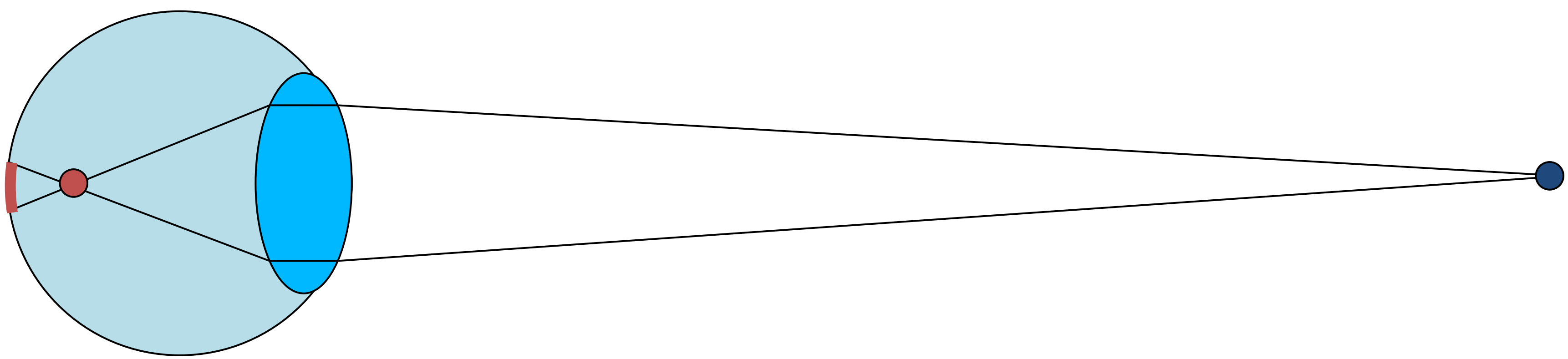


On the contrary, the points out of the focalisation plane $P_f$ form an image whose level of blur is proportional to their distance to $P_f$.

See: *depth from defocus*

(Note the ambiguity of the position due to the blur symmetry with respect to $P_f$).

*Focal too long: near object appears blurred on the retina plane*

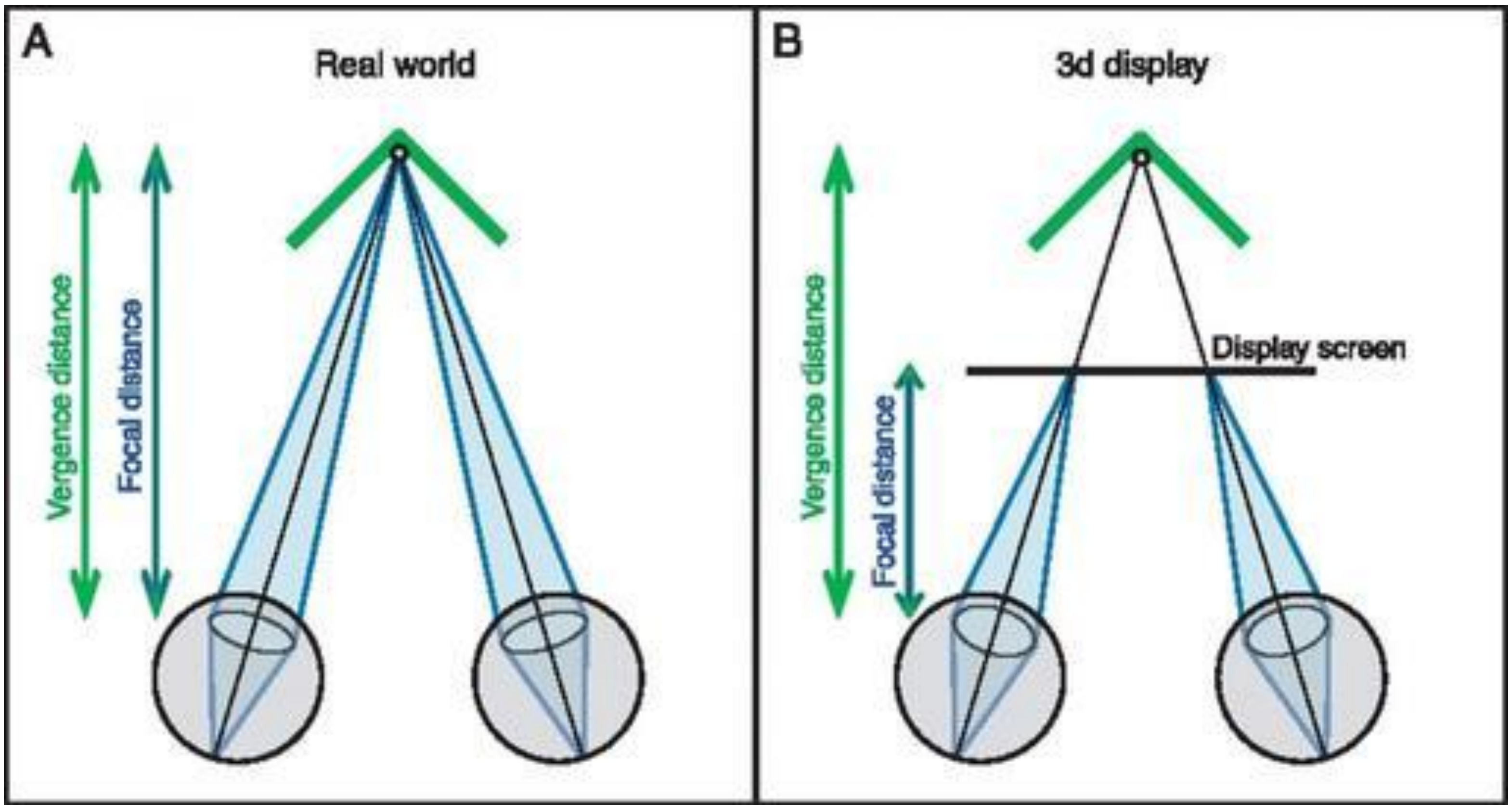


*Focal too short: far object appears blurred on the retina plane*

# STEREOPSIS *VS* 3D DISPLAY

In natural binocular vision (stereopsis), vergence and accommodation are congruent (left diagram).

It is however possible to put – more or less deliberately – in conflict these two functions (right diagram). Thanks to this mechanism, it is possible to get a sharp 3d perception using a 2d display.




[Hoffman 2008]

# 3D DISPLAY: ANAGLYPHS

*Šibenik City Hotel Hall (Anaglyph)*

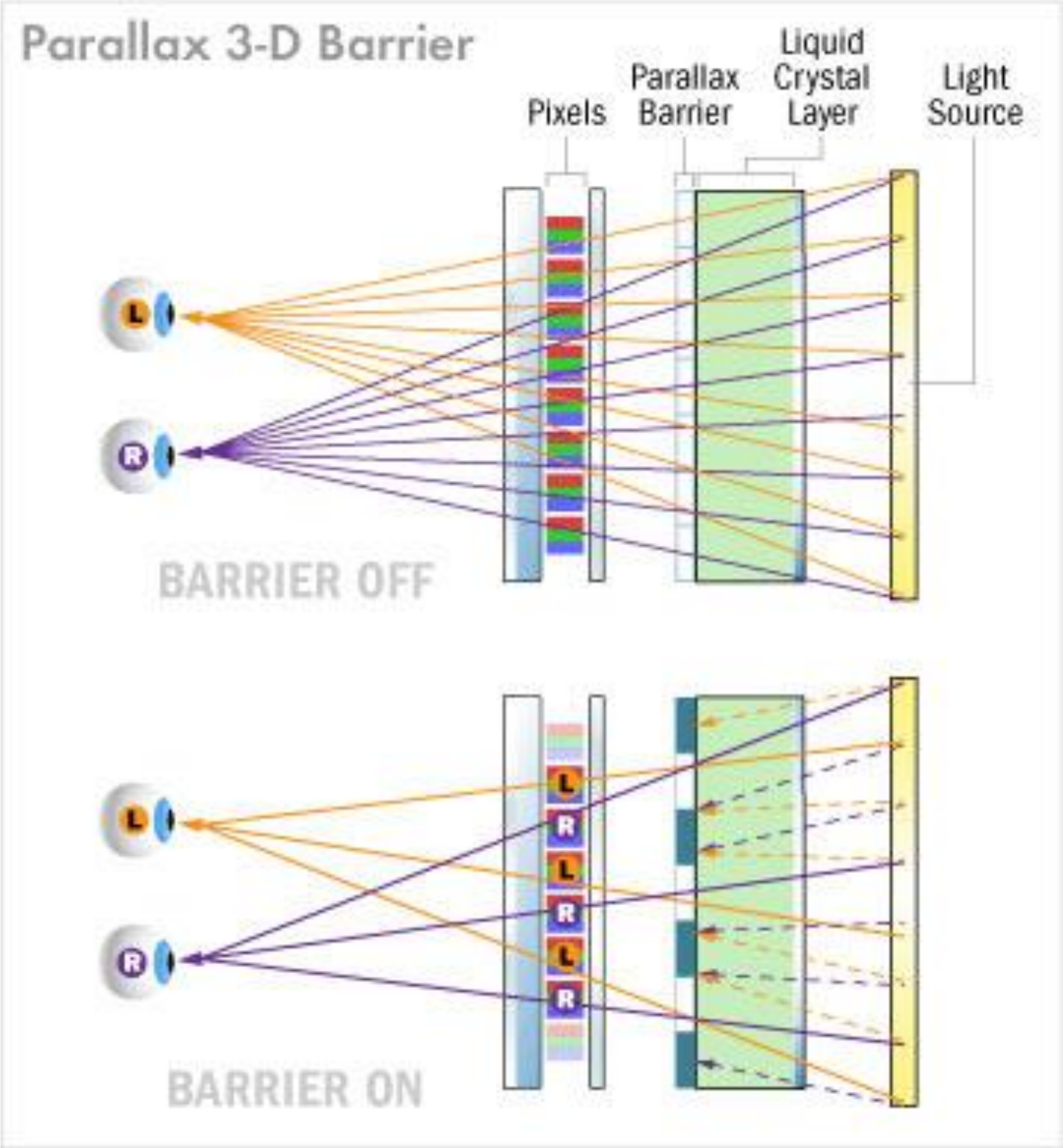
# 3D DISPLAY: AUTOSTEREOGRAMS

*L.M. Rutherford, Full Moon, stereo pair (1864)*

# 3D DISPLAY: PARALLAX BARRIER 3D SCREEN

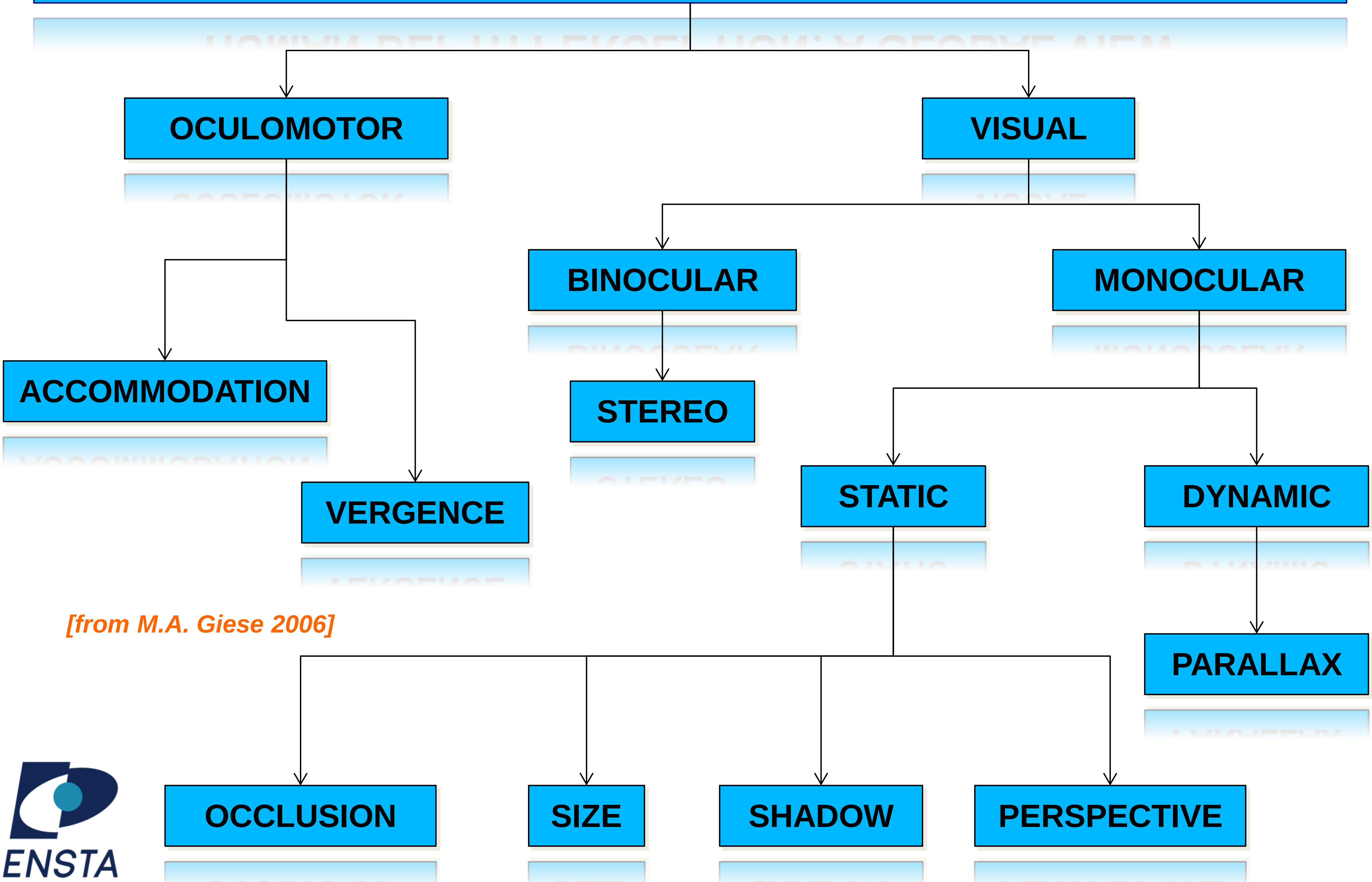
In parallax 3d-barrier screens, an opaque vertical grid is positioned between the liquid crystal layer and the colour filters (pixels), in such a way to separate by parallax pixels seen by the left eye from those seen by the right one.

Ex: Nintendo 3DS



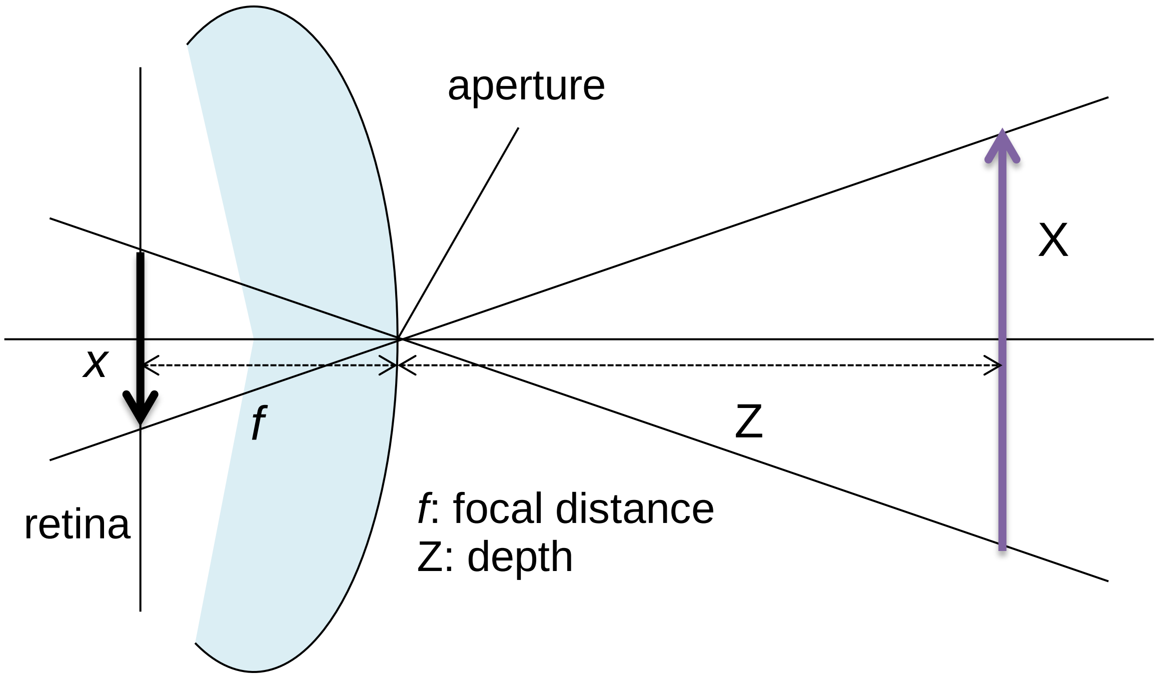

*[Ill. howstuffworks.com]*

# HUMAN DEPTH PERCEPTION: A GLOBAL VIEW




*[from M.A. Giese 2006]*

# PARALLAX AND THE OPTICAL FLOW




(O,X,Y,Z) 3d real coordinates

(O,x,y) 2d retinal coordinates

Perspective equation (pinhole model):

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

Time derivative (optical flow):

$$\dot{x} = f\left(\frac{\dot{X}}{Z} - \frac{X\,\dot{Z}}{Z^2}\right)$$

In the case of a pure translation along OX axis (horizontal travelling, $\dot{Z} = 0$; $\dot{X} = Cte$) :

$$\dot{x} = \frac{f\,\dot{X}}{Z}$$

and then:

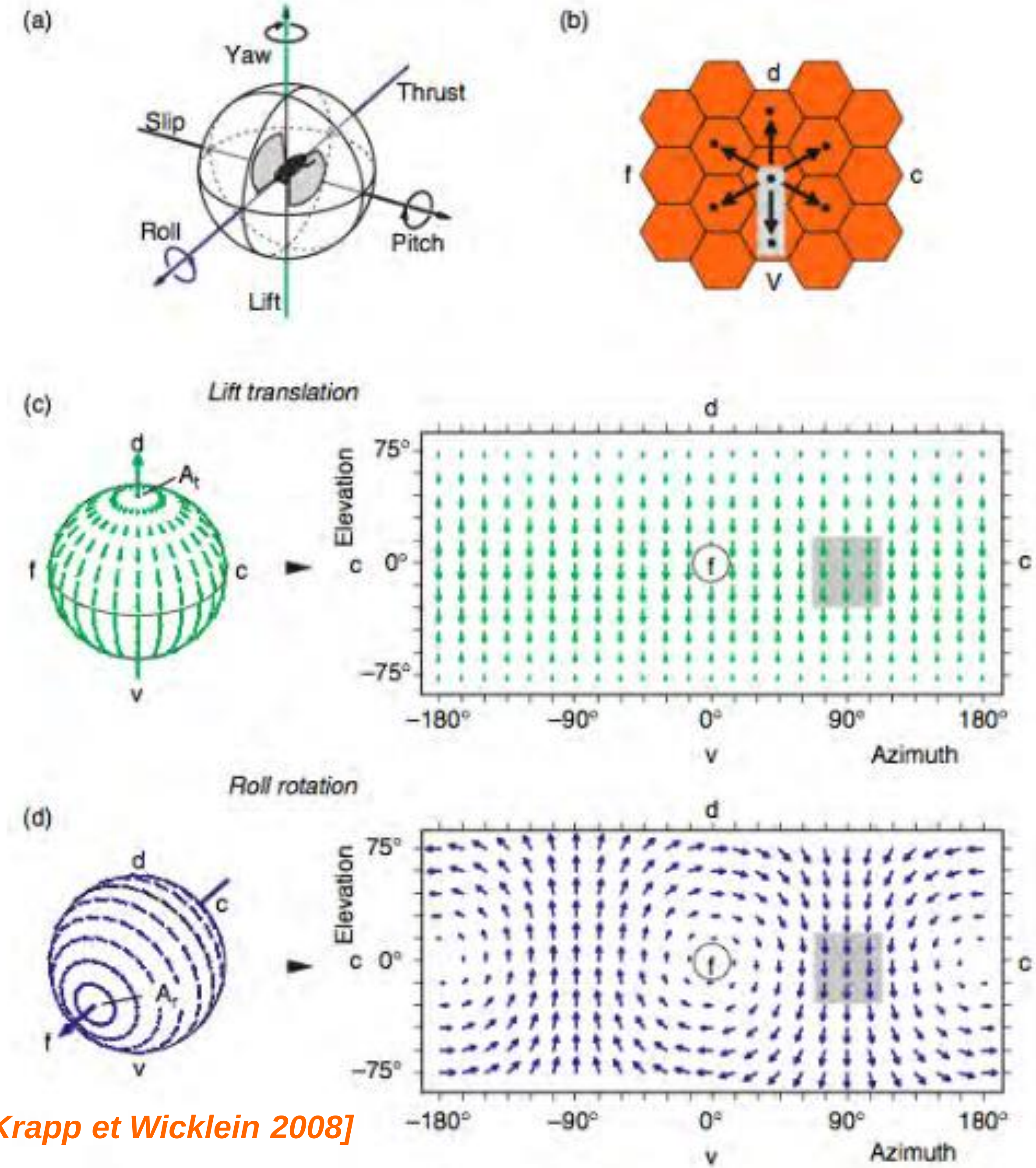
$$\boxed{Z = \frac{f\,\dot{X}}{\dot{x}}}$$

Depth is inversely proportional to the magnitude of apparent velocity.

# THE FLIGHT OF THE BUMBLEBEE

*[Jürgen Tautz 2008]*

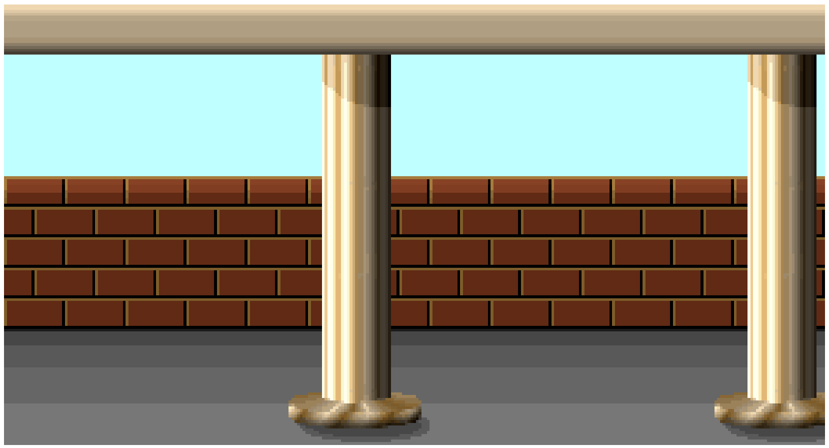
The bee is able to navigate in small corridors by controlling the direction of his flight from the balance of the spatial average apparent speeds estimated by his left and right eyes.



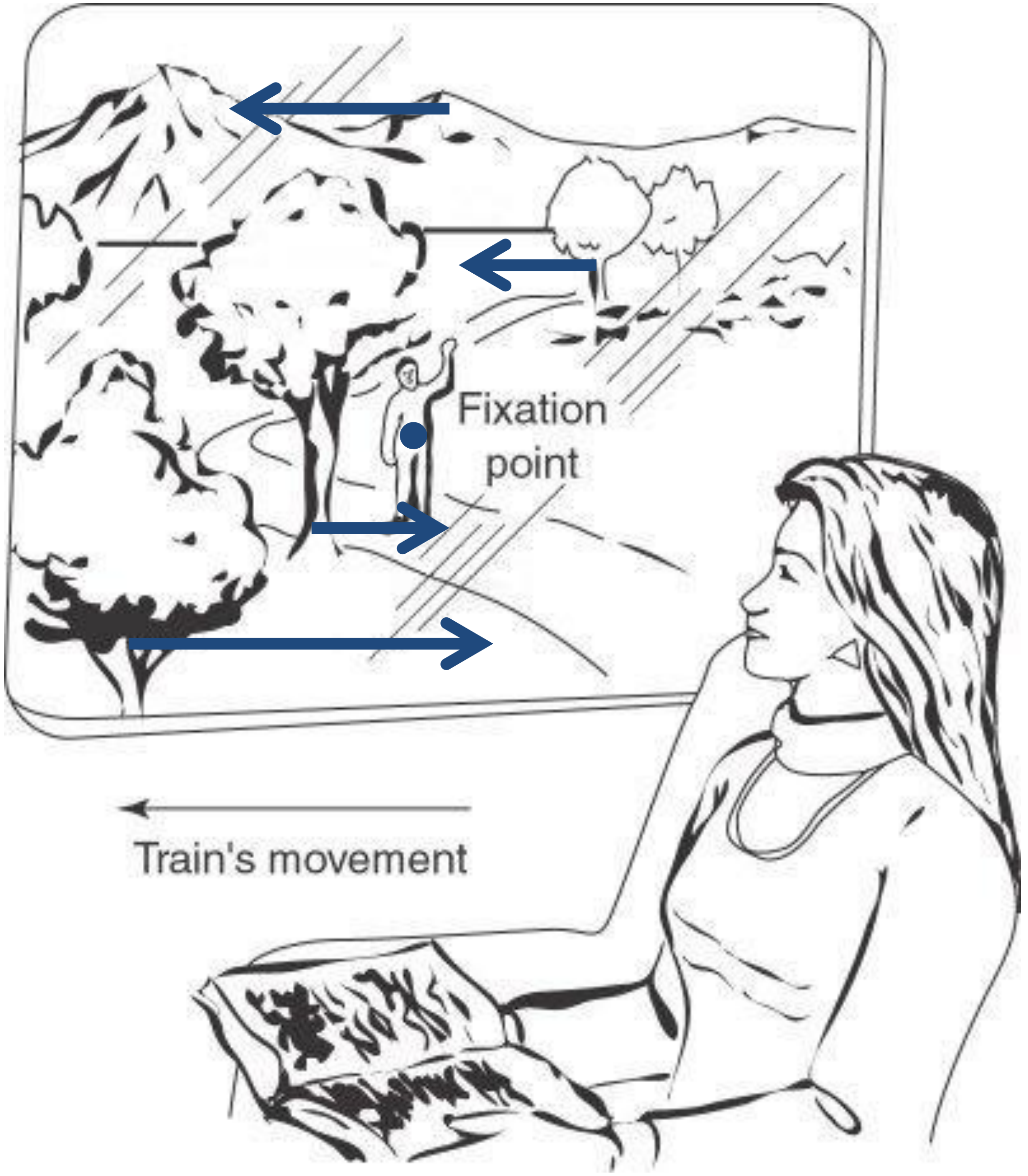

*[Krapp et Wicklein 2008]*

# DYNAMIC MONOCULAR 3D: PARALLAX

$$Z = \frac{f\,\dot{X}}{\dot{x}}$$



[© nvnews.net]

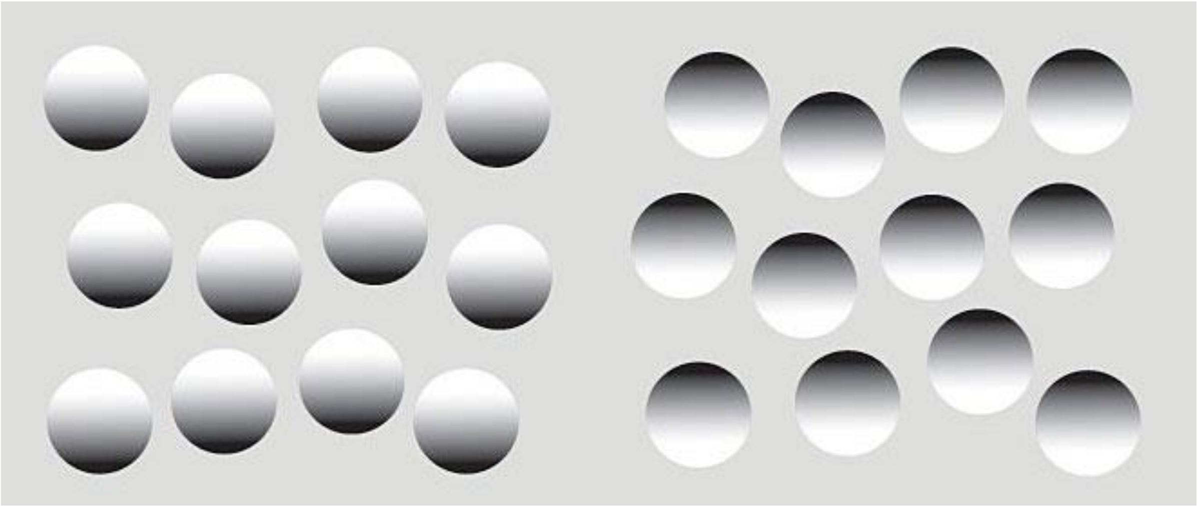



[Kenneth M. Steele 2014]

# STATIC MONOCULAR 3D: OCCLUSIONS

*Giotto – Pentecoste (c. 1305)*

# STATIC MONOCULAR 3D: SHADOWS



Self shadowing is a strong but ambiguous depth cue (light source position *vs* concavity).

Without shape prior, the concavity is determined by a prior of top lighting (left image).

When the shape prior is strong (face then convex), the concavity prior dominates the lighting prior (top-down effect, animation on the right).

See *shape from shading*


ENSTA

IP PARIS

# STATIC MONOCULAR 3D: SIZES

*Georges Seurat – Un dimanche après-midi à l'Île de la Grande Jatte (1884-86)*

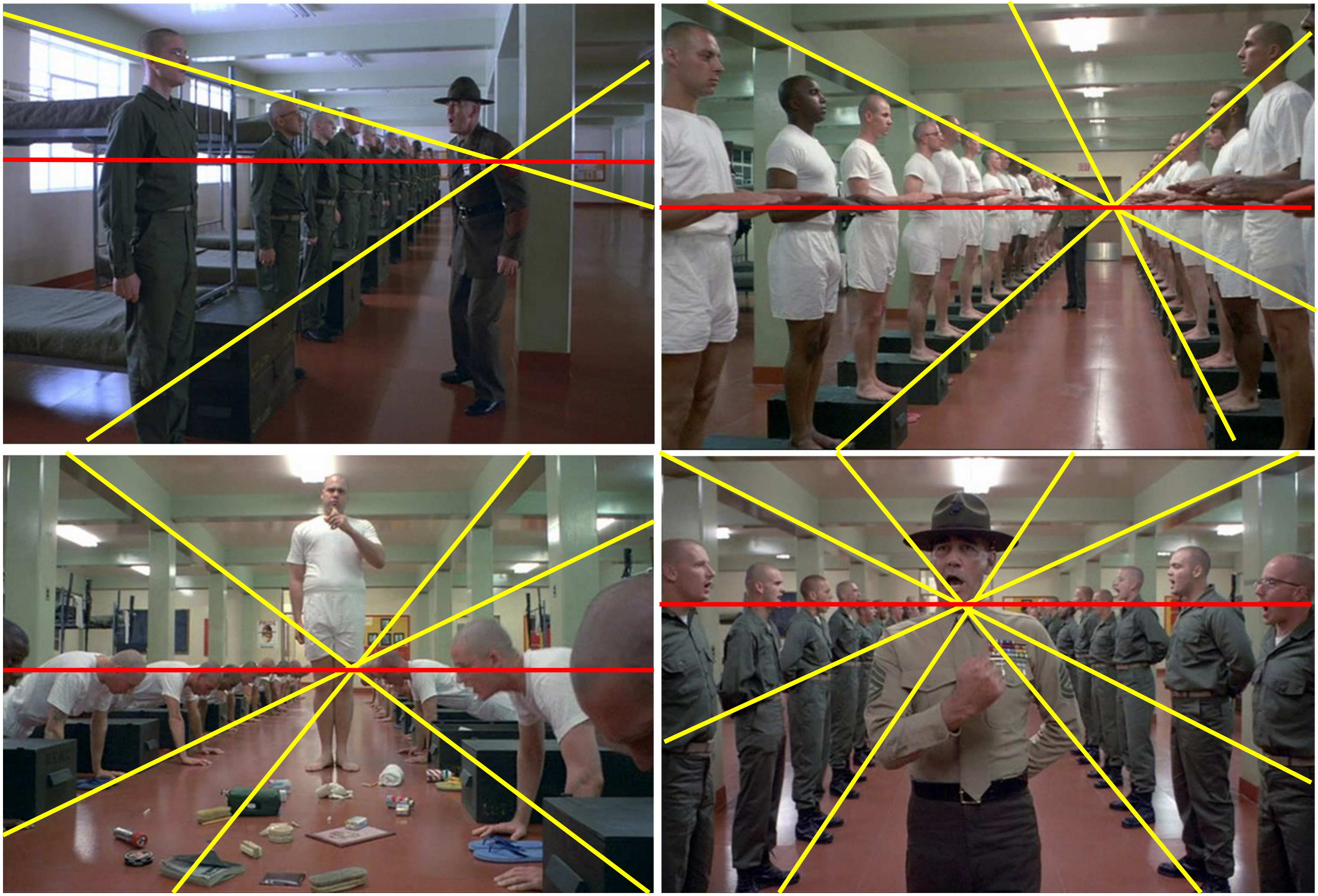
# STATIC MONOCULAR 3D: PERSPECTIVE

*Stanley Kubrick – The Shining (1980)*

# 3 STATIC  MONOCULAR 3D: SIZES *AND* PERSPECTIVE

*Stanley Kubrick – Full Metal Jacket (1987)*

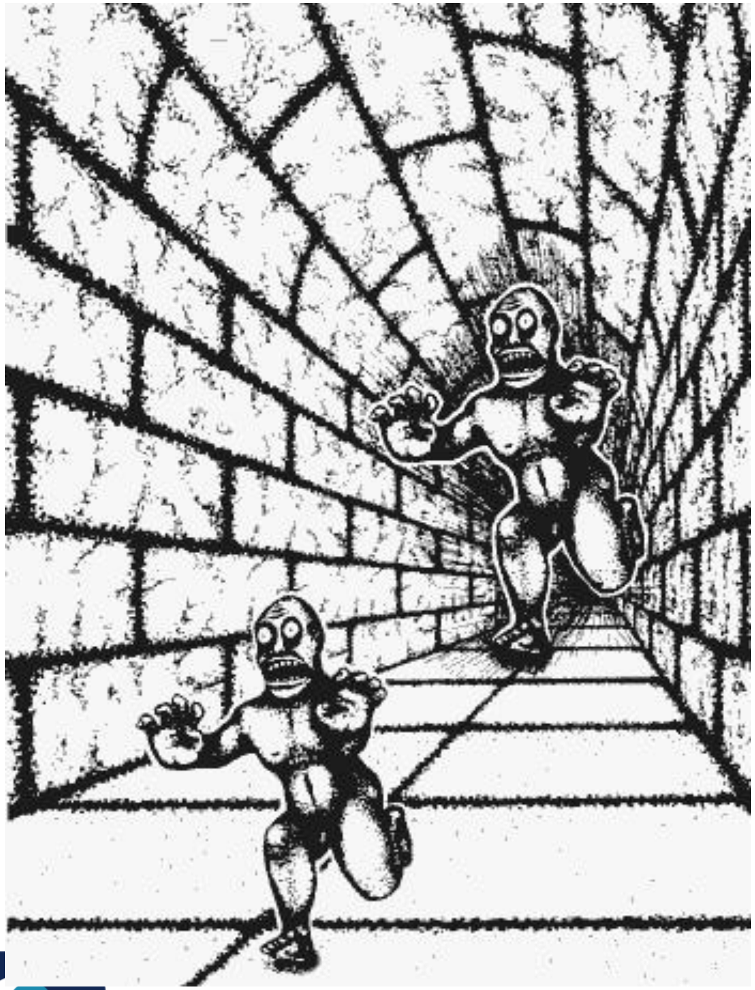
# PERSPECTIVE: HORIZON AND VANISHING POINT



*Stanley Kubrick – Full Metal Jacket (1987)*

# STATIC MONOCULAR 3D: SIZES *VS* PERSPECTIVE

Two popular examples of conflict between monocular depth cues:



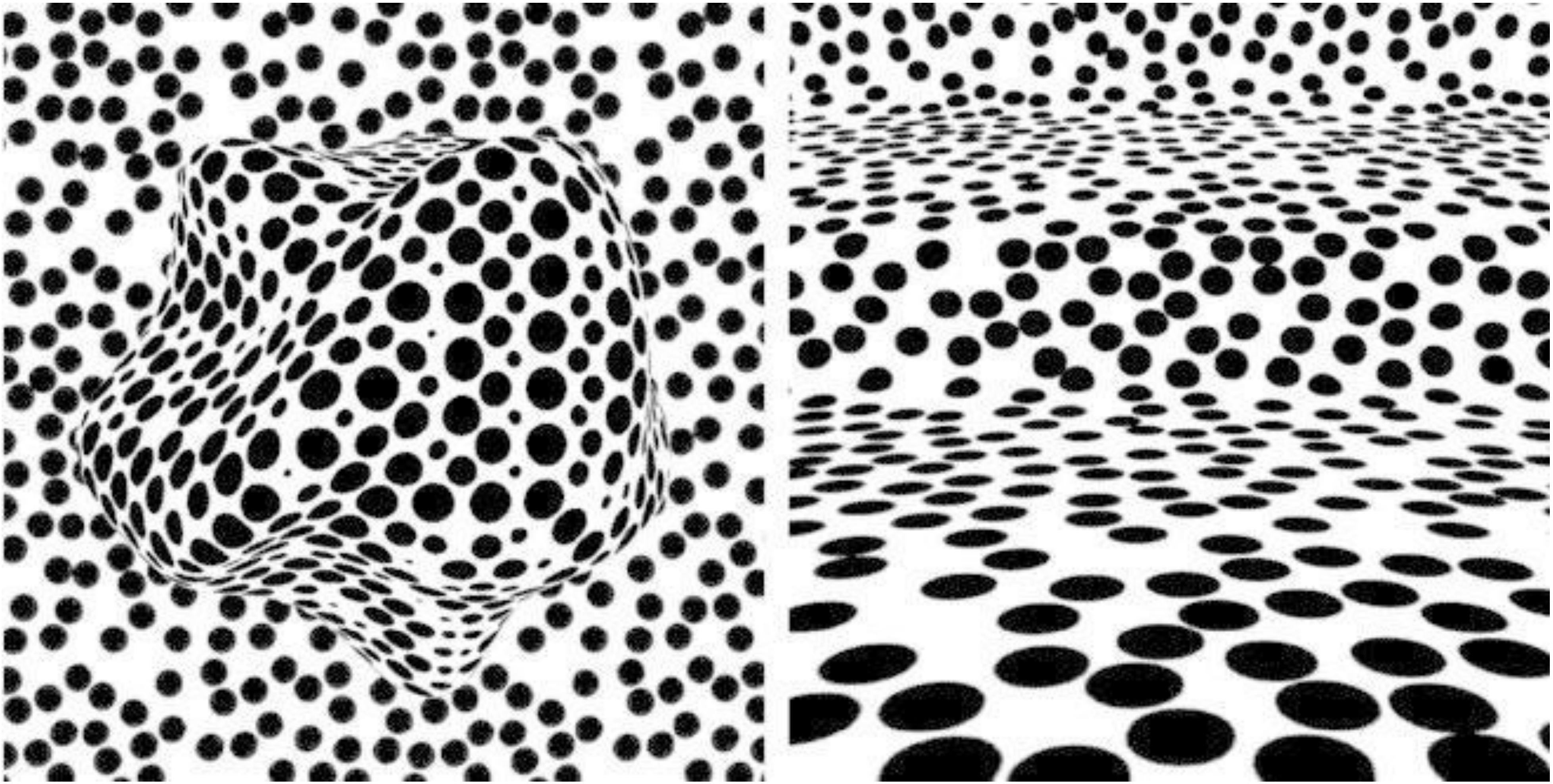
*The Hallway illusion*

*The Ame's room*

# STATIC MONOCULAR 3D: SIZES *VS* PERSPECTIVE

*Avignon TGV station: illusory space amplification created by accelerated perspective*

# STATIC MONOCULAR 3D: TEXTURE GRADIENTS



*[Ill. DrThomas @ Studyblue]*

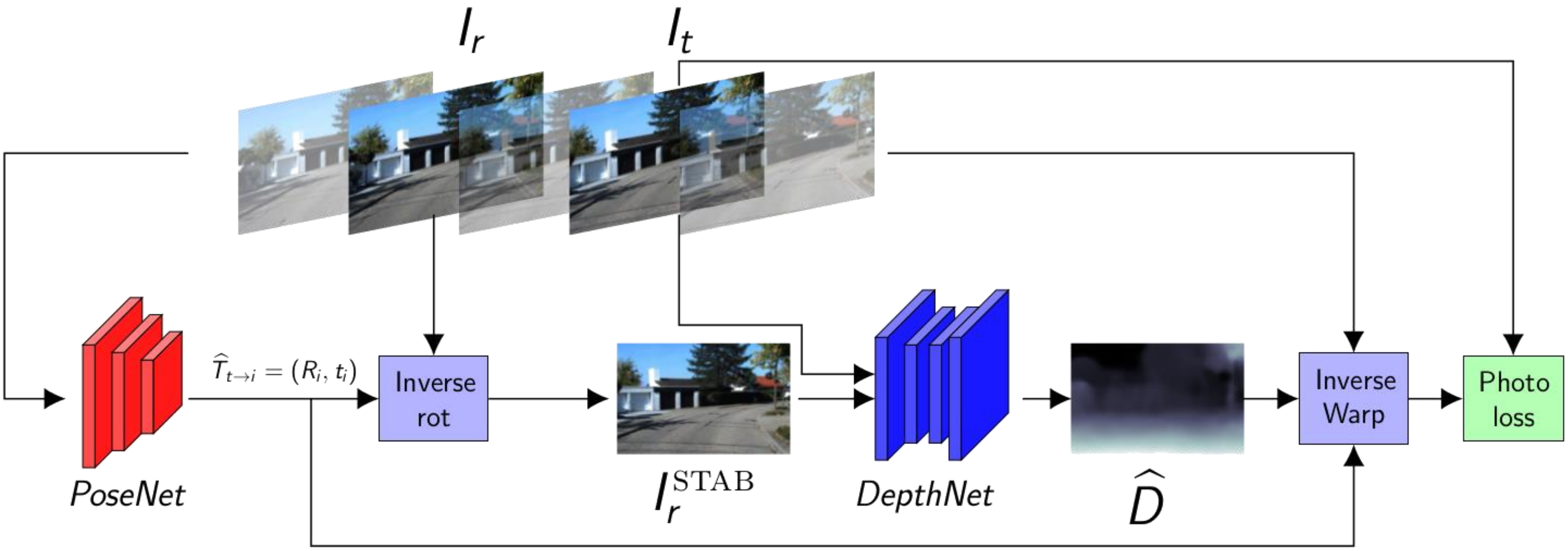
# TEXTURE GRADIENTS, SIZES AND PERSPECTIVES

*Gustave Caillebotte – Rue de Paris, temps de pluie (1877)*

# PERSPECTIVE: HORIZON AND VANISHING POINTS

*Gustave Caillebotte – Rue de Paris, temps de pluie (1877)*

# END-TO-END LEARNING OF DEPTH MAPS?




$$\forall i,\ t_i^{\mathrm{NORM}} = t_i \frac{T_0}{\epsilon + \|t_r\|}$$

Deep Neural Networks have the capability to exploit all possible 3d cues to predict dense depth maps from videos…

See upcoming lecture on deep learning based depth maps prediction

# **Part 2:** 3D CAMERAS / ACTIVE APPROACHES

Active 3d cameras aim at measuring the depth of every point from the scene that is projected on the image plane, using its response to a particular lighting.

The two fundamental components of the system are then:

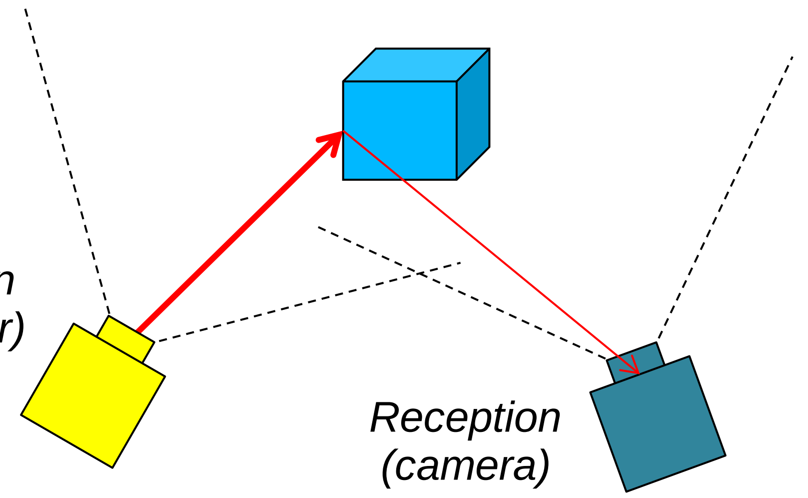
1. A lighting system controlled in time and space
2. A sensing device to analyse the illuminated scene

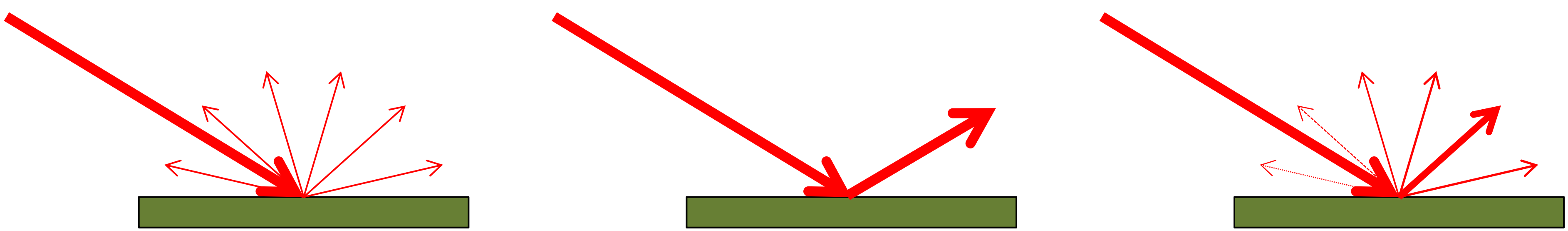
Such systems are active in that they *emit* a light signal (not to be confused with the other sense of « active vision  », i.e. that « moves to see »).

# ACTIVE APPROACHES AND DIFFUSION MODELS

For 3d active cameras, it is assumed that every point illuminated by the projector in the camera field of view reflects a part of its light toward the optical centre of the camera.

The nature of light diffusion at the measured point then has a major influence in the depth estimation…



*Lambertian diffusion* | *Perfect specular reflection (mirror)* | *Semi specular diffusion*

Note that this may also be an issue for passive approaches (e.g point matching between two poses).

# ACTIVE 3D: "TIME OF FLIGHT" CAMERAS

3d « time of flight » (ToF) cameras measure the distance $d_x$ between a point X projected in $x$, from the propagation time $t_x$ of light (with speed $c$) from its emission by the projector until its reception by the photosensor associated to $x$, after being reflected by point X:

$$d_x = \frac{c.t_x}{2}$$

Unlike scanner like (e.g. LIDAR) systems, the light emitted by ToF cameras (usually laser infrared LED) illuminates the whole scene simultaneously.

Different technologies can be used to measure time of flight:

- Direct time measuring (impulsion light)
- Phase estimating (time-modulated continuous light).

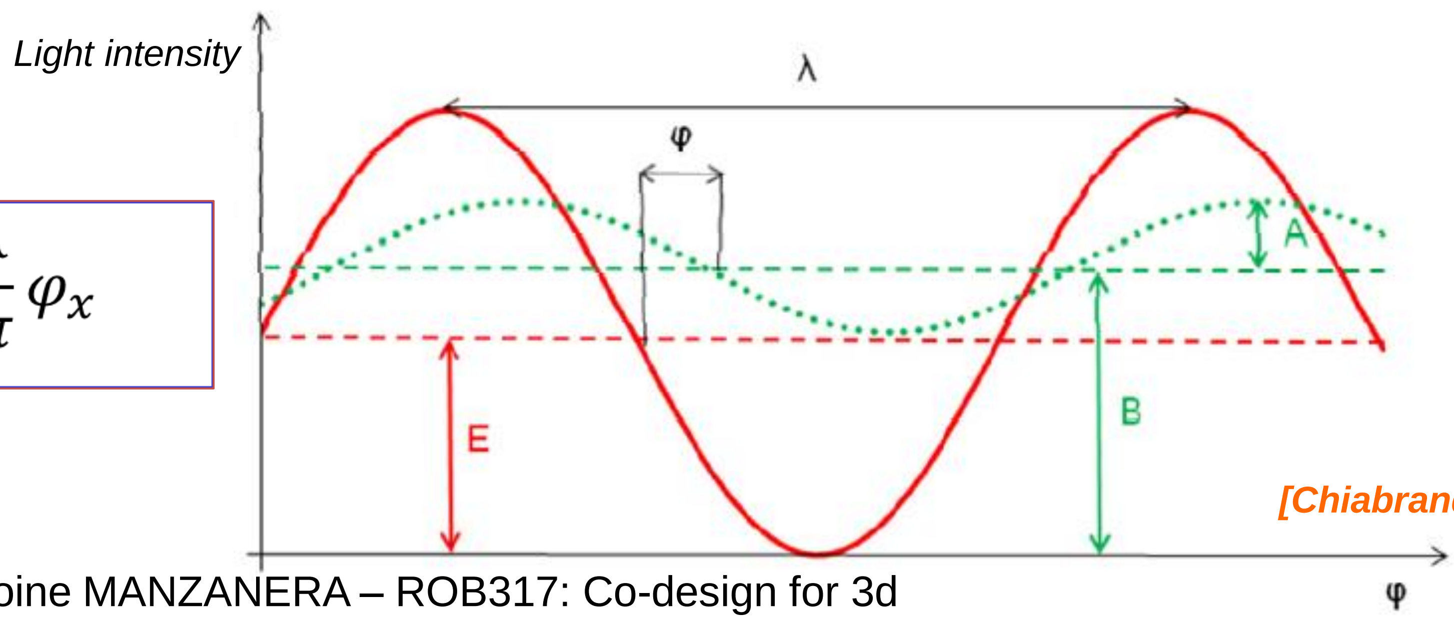

*[CamCube - ©PMDTech]*


*[Kinect v2 for Xbox One - ©Microsoft]*

# ACTIVE 3D: ToF CAMERA BASED ON PHASE ESTIMATION

- The scene is uniformly illuminated with a light whose intensity varies in time according to a sine signal (in red) with amplitude E.
- The signal received in pixel *x* (in green) has the same frequence, a weaker amplitude A depending on the reflectivity of the point and a phase shift φ depending on its distance.
- The signal received is also shifted in intensity (offset) of a value B due to the background light present in the scene.
- This signal is sampled and the phase shift φ is deduced from the measured intensities.
- The modulation period λ (typ. 50 ns) is large with respect to the time of flight to avoid phase ambiguities, but small with respect to typical acquisition times to allow repeating the measure (time filtering).

$$d_x = \frac{c\lambda}{4\pi}\varphi_x$$




[Chiabrando 2009]

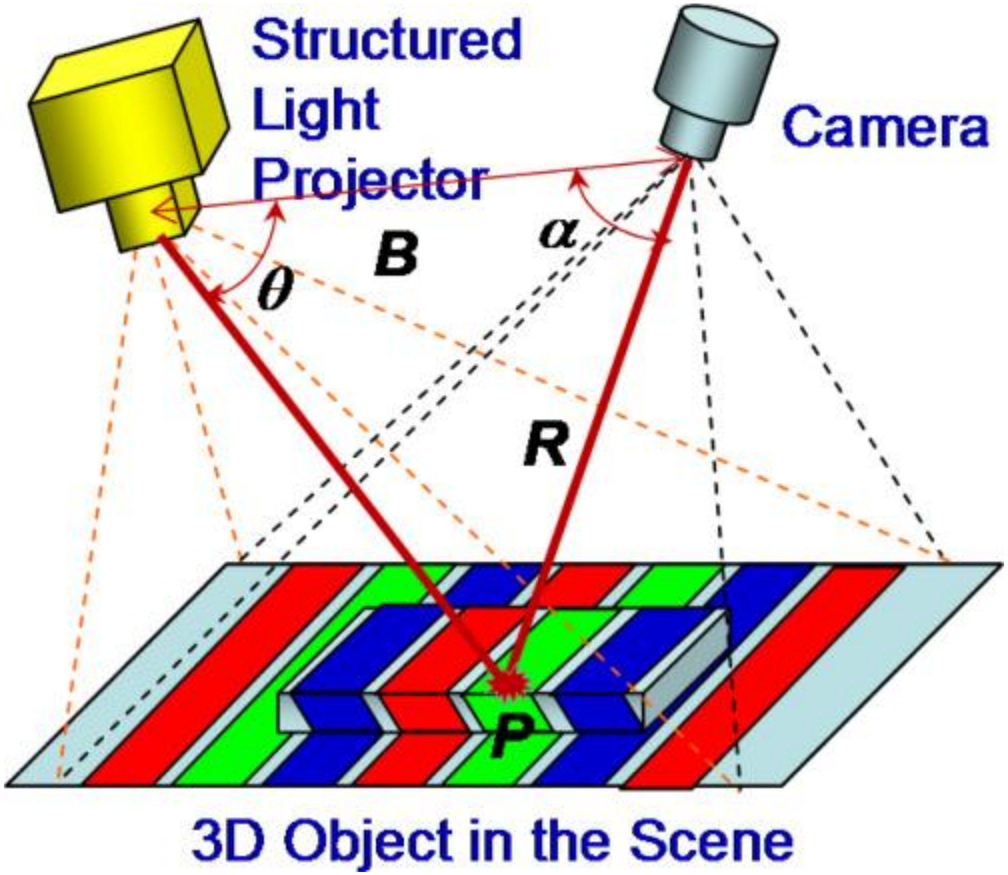
# ACTIVE 3D: “STRUCTURED LIGHT” CAMERAS

Structured light 3d cameras interpret the deformation of a 2d image projected into the scene to recover depth information.

They are based on the same triangulation principle as stereovision:

$$R = B\,\frac{\sin\theta}{\sin(\alpha+\theta)}$$

The structure of projected 2d images determines a spatial coding that plays a major role in triangulation.

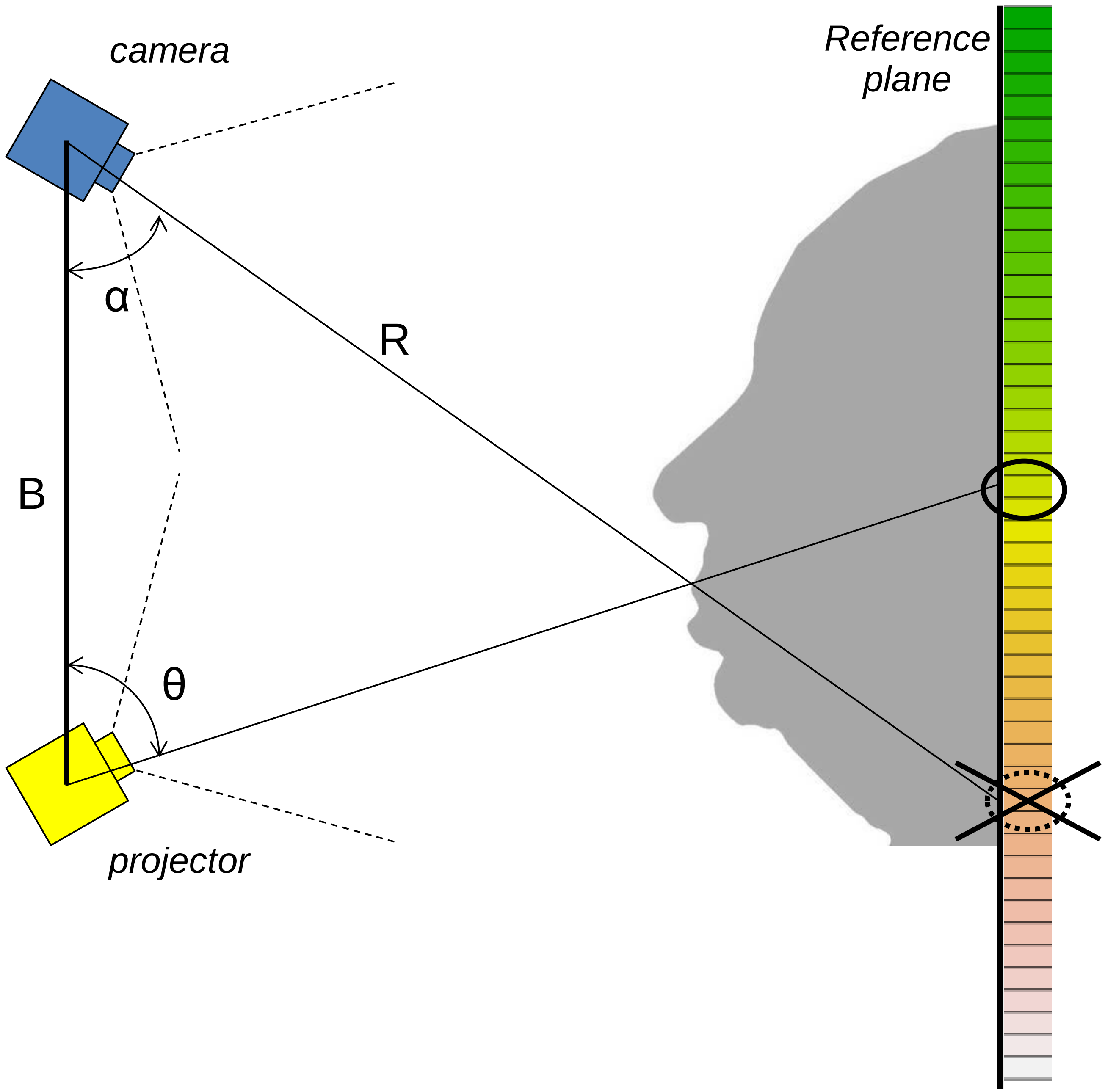



[Geng 2011]

# ACTIVE 3D: "STRUCTURED LIGHT" CAMERAS

$$R = B\,\frac{\sin\theta}{\sin(\alpha+\theta)}$$

The angle α is provided by the position of the point in the image, and the angle θ by the corresponding colour (or pattern) in the reference plane:



# ACTIVE 3D: "STRUCTURED LIGHT" CAMERAS

Also:

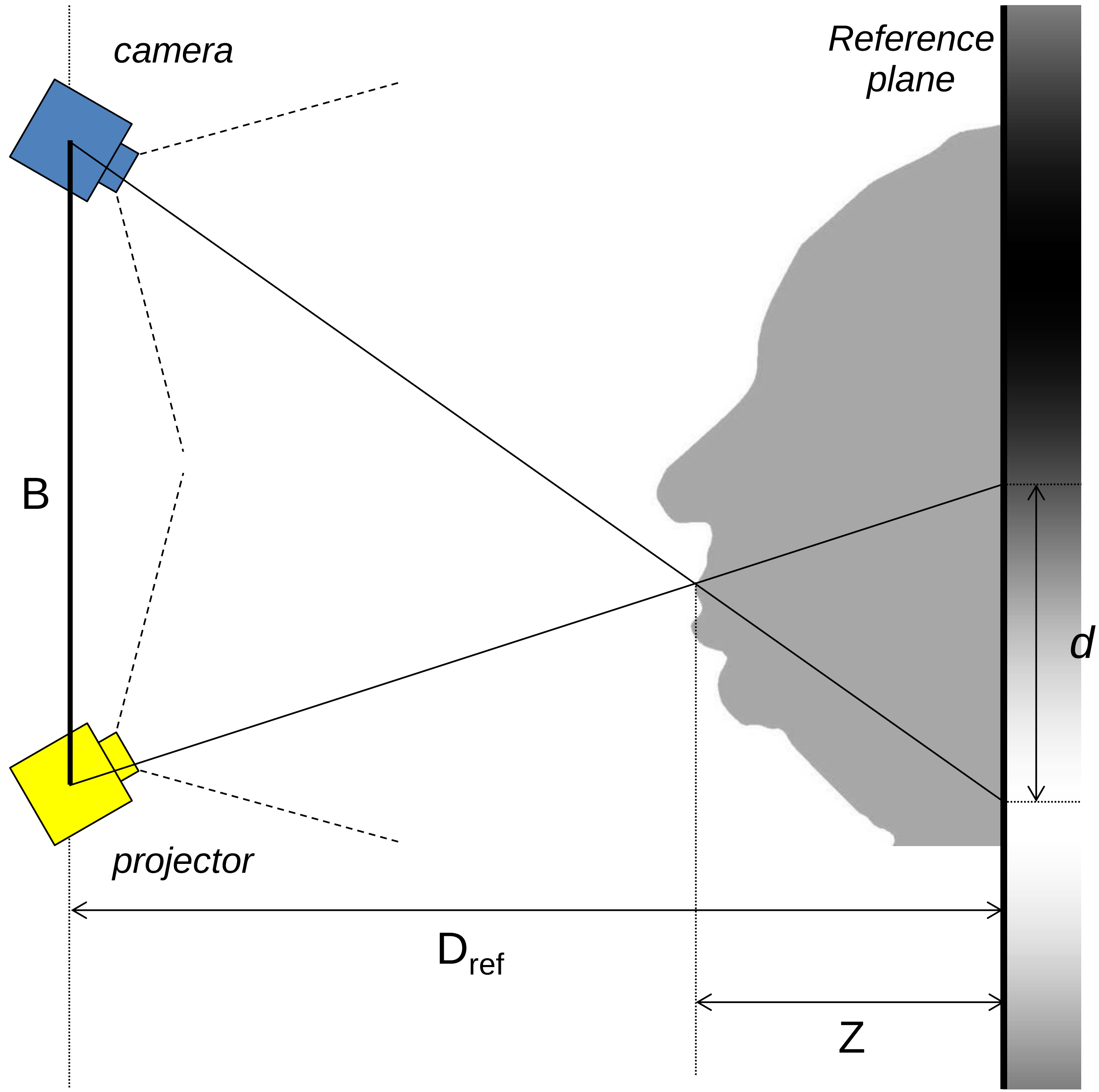
$$\frac{d}{B} = \frac{Z}{D_{ref} - Z}$$

And then:

$$Z \approx \frac{D_{ref}}{B} d$$

So, if the projected image is a sinusoidal ramp, depth can be deduced from the phase shift:

$$Z \propto \Delta\varphi$$

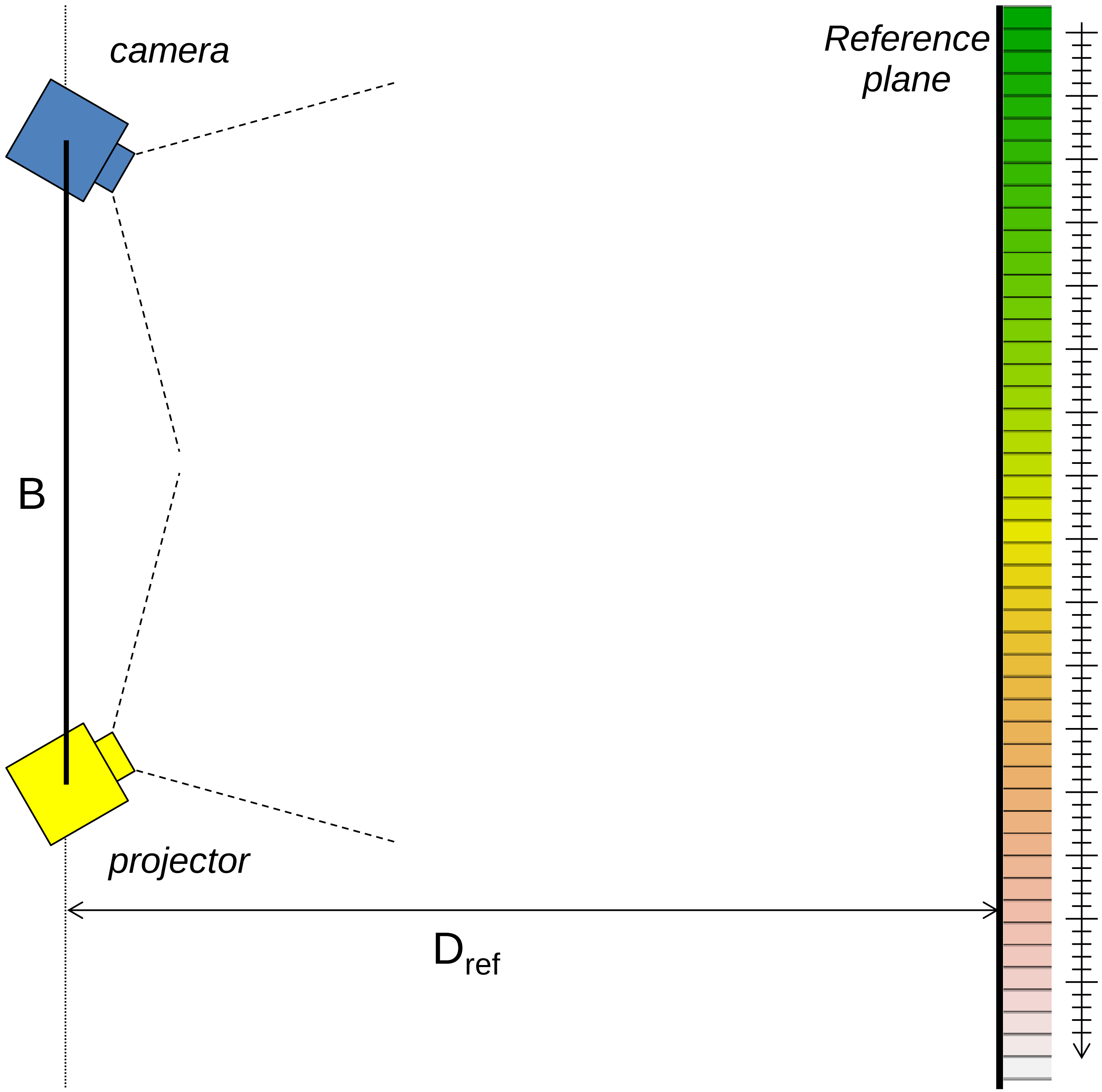


# "STRUCTURED LIGHT" CAMERAS: CALIBRATION

Like stereo systems, the camera and the projector must be calibrated to:

(1) Determine the back-projection line for every pixel of the captured image.

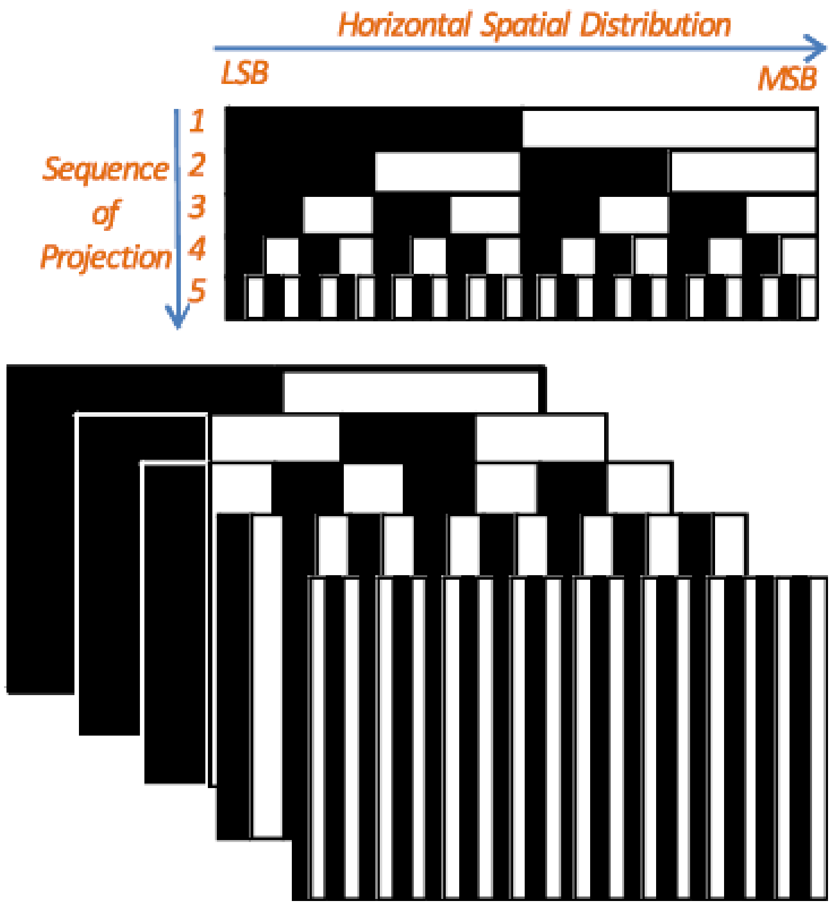
(2) Associate to each pattern of the projected image the direction corresponding to its projection on the reference plane.



# STRUCTURED LIGHT: WHICH PATTERNS?

- Ideally, every point should be uniquely indentified from its value/colour…
  - …but all the values must be easily distinguishable!
- A point can also be identified using its neighbourhood…
  - …but then each neighbourhood must be unique!
- Depth being associated to an angle, a 1d target (band) is sufficient…
  - ..but using a 2d target may solve ambiguities!
- Several targets may also be sequentially combined...
  - …but then the acquisition time increases!

# STRUCTURED LIGHT: SEQUENTIAL TARGETS




***Binary sequence $2^5$***

***[Posdamer 1982, from Geng 2011]***

Binary targets allow to optimally discriminate the different values.

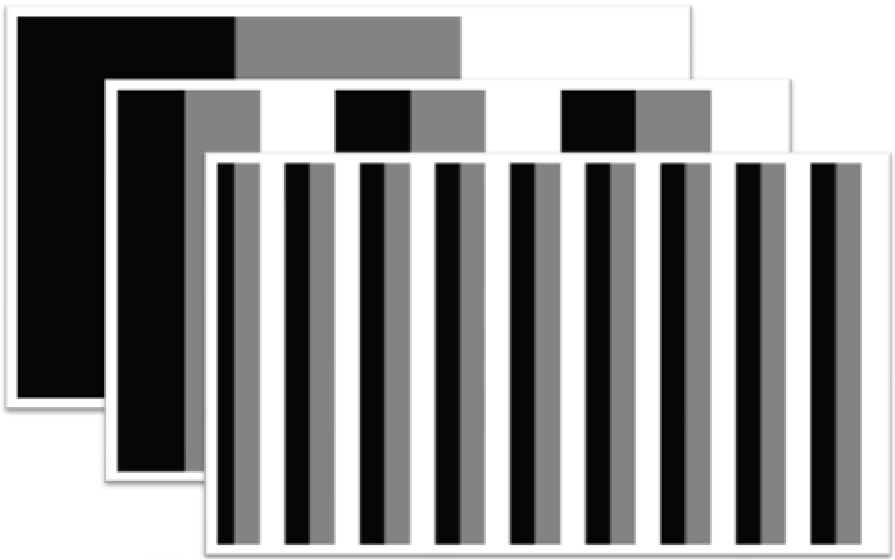
Depth resolution depends on the number of distinct values and then, for sequential techniques, on the acquisition time.

***[from Naramsimhan 2006]***

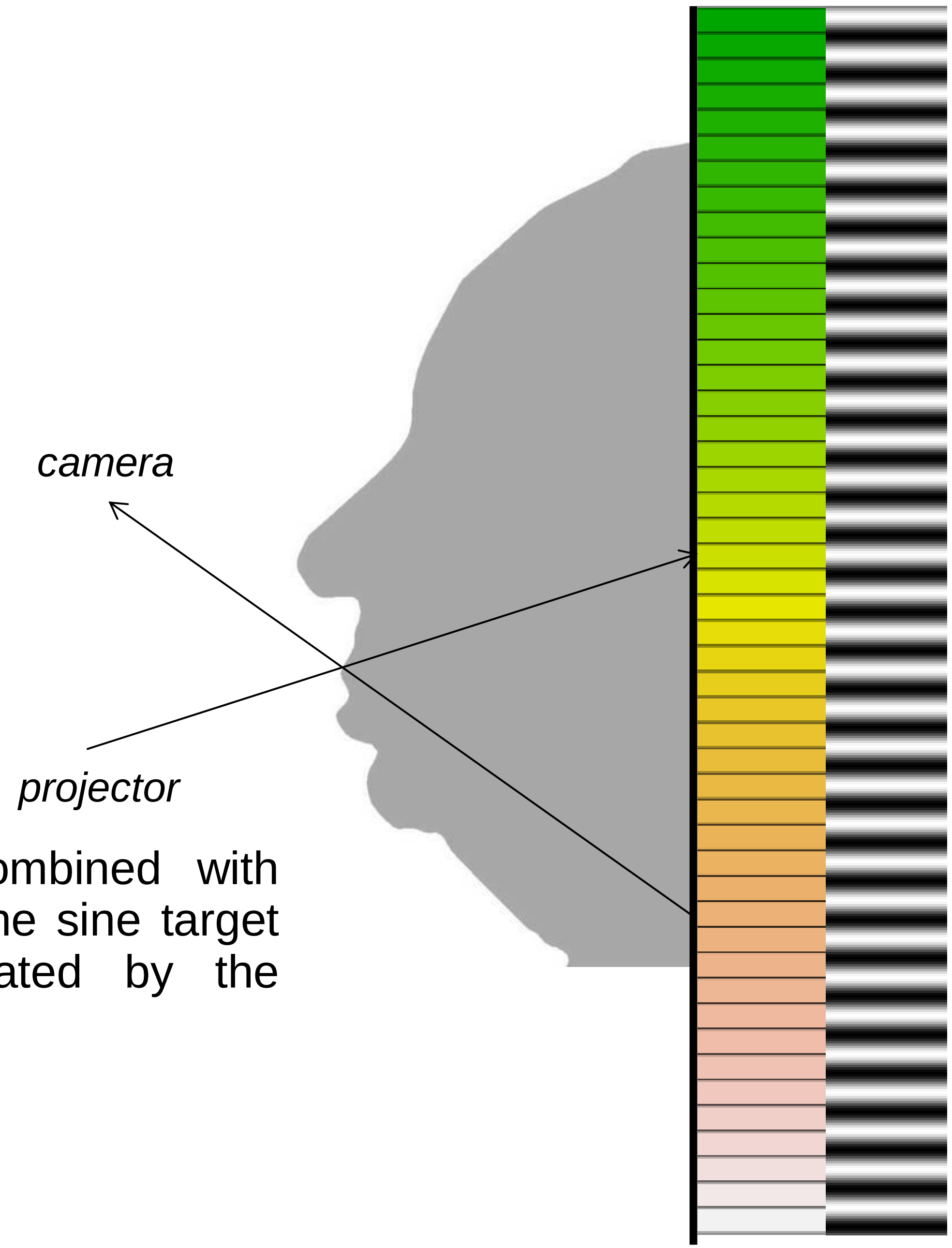
ENSTA IP PARIS

# STRUCTURED LIGHT: SEQUENTIAL TARGETS

Increasing the number of bits for a trade-off contrast / acquisition time:



***Ternary sequence $3^3$***

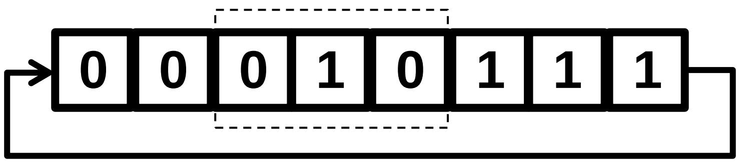



Rectangular signal targets can also be combined with sinusoid targets (on the right): the phase of the sine target allows to refine the coarse depth estimated by the rectangular target.

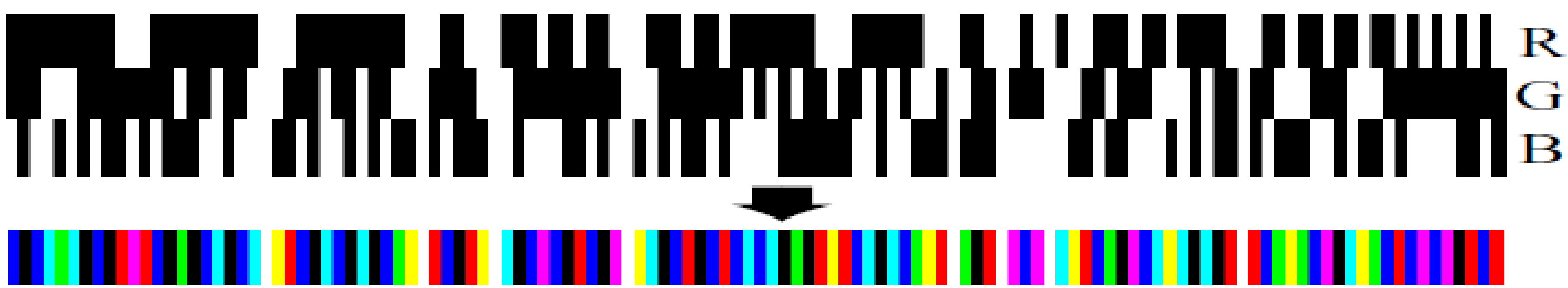
# STRUCTURED LIGHT: UNIQUE “SNAPSHOT” TARGET

- To better distinguish the values, rectangular (runs) targets are preferred to continuous ones (ramps).
- To be able to locally discriminate points using quantised values, local patterns (neighbourhoods) can be used instead of the value alone.
- But then, each pattern must define a *unique* position.

De Bruijn’s sequence *B(n,k)* are words from a *n*-symbols alphabet such that all the sub-words of length *k* are different.




*De Bruijn’s sequence B*(2,3)

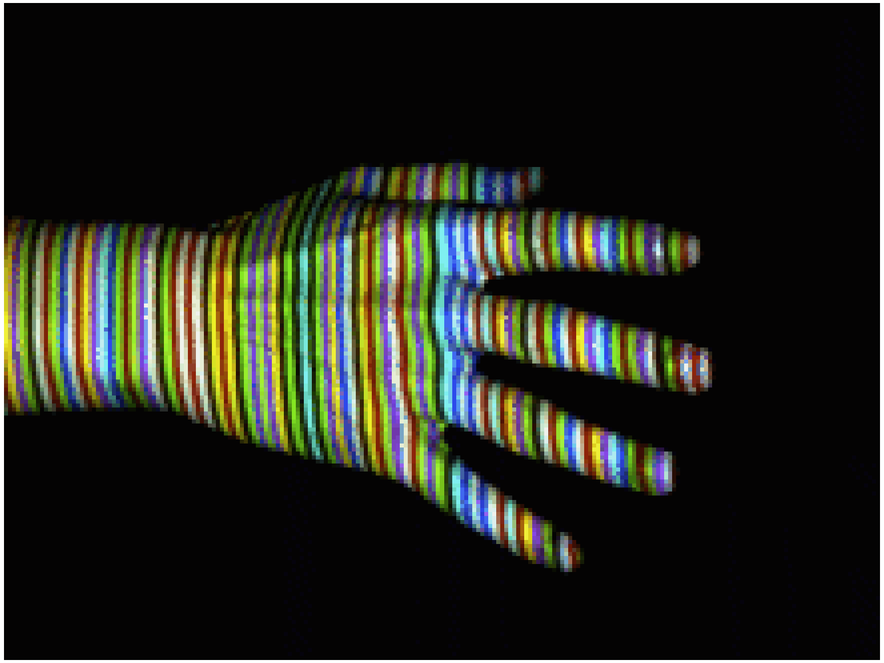


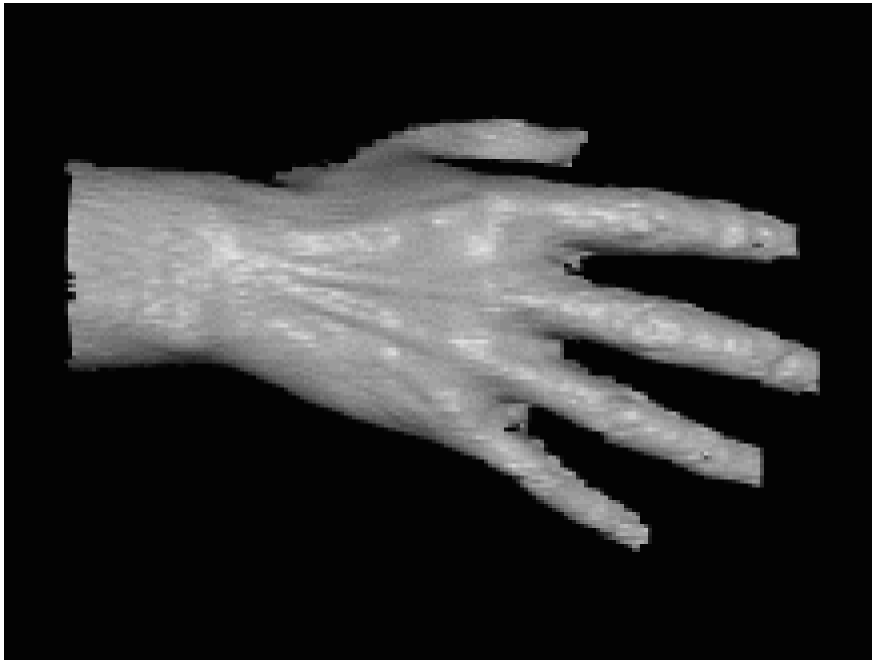

*Colour De Bruijn’s sequence B*(5,3)

**[Zhang 2002]**

# STRUCTURED LIGHT: UNIQUE “SNAPSHOT” TARGET

Using a unique (« snapshot ») target reduces significantly the acquisition time and then allows to acquire mobile scenes:

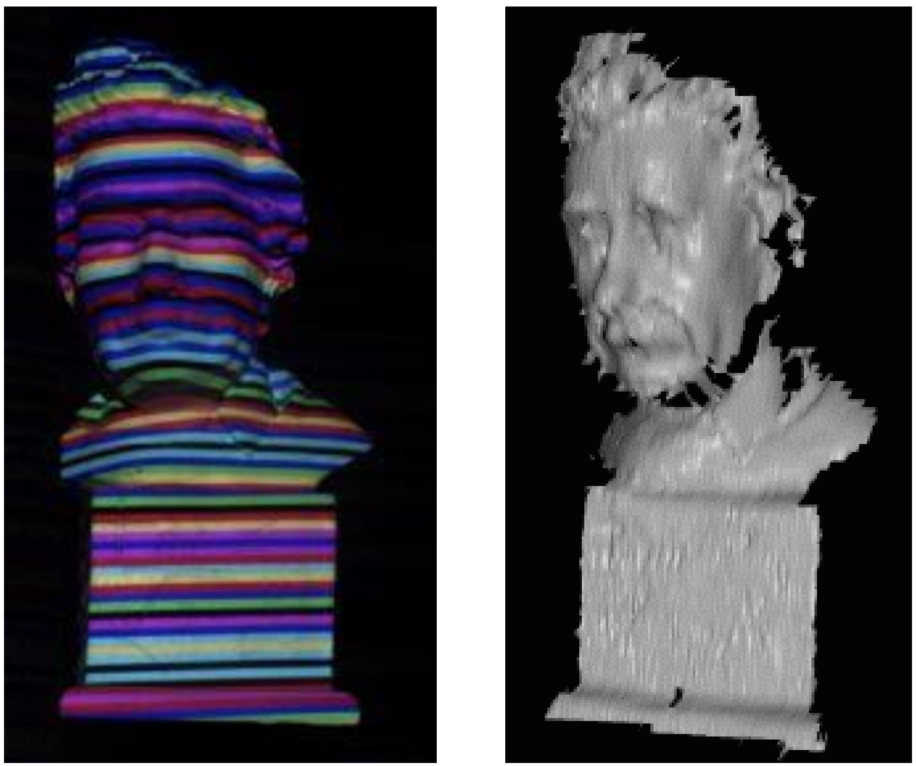




[Zhang 2002]

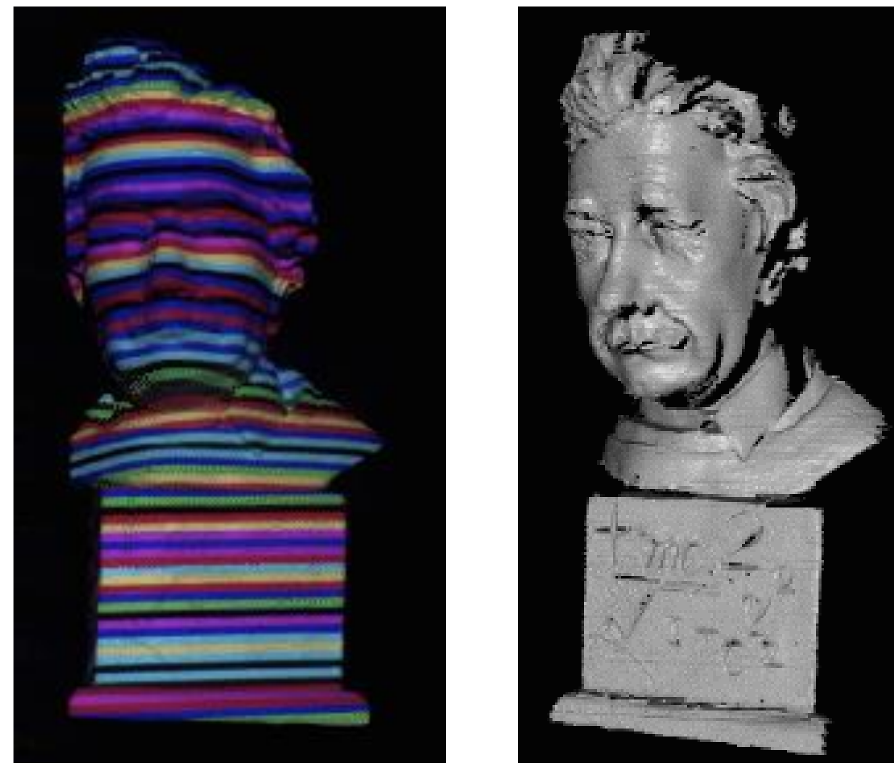
# DE BRUIJN’S TARGET: SNAPSHOT *VS* SEQUENTIAL

Targets designed for snapshot acquisition can be used with *phase shifts* for sequential acquisitions, to improve both robustness and resolution (static scenes):



*« Snapshot » acquisition*

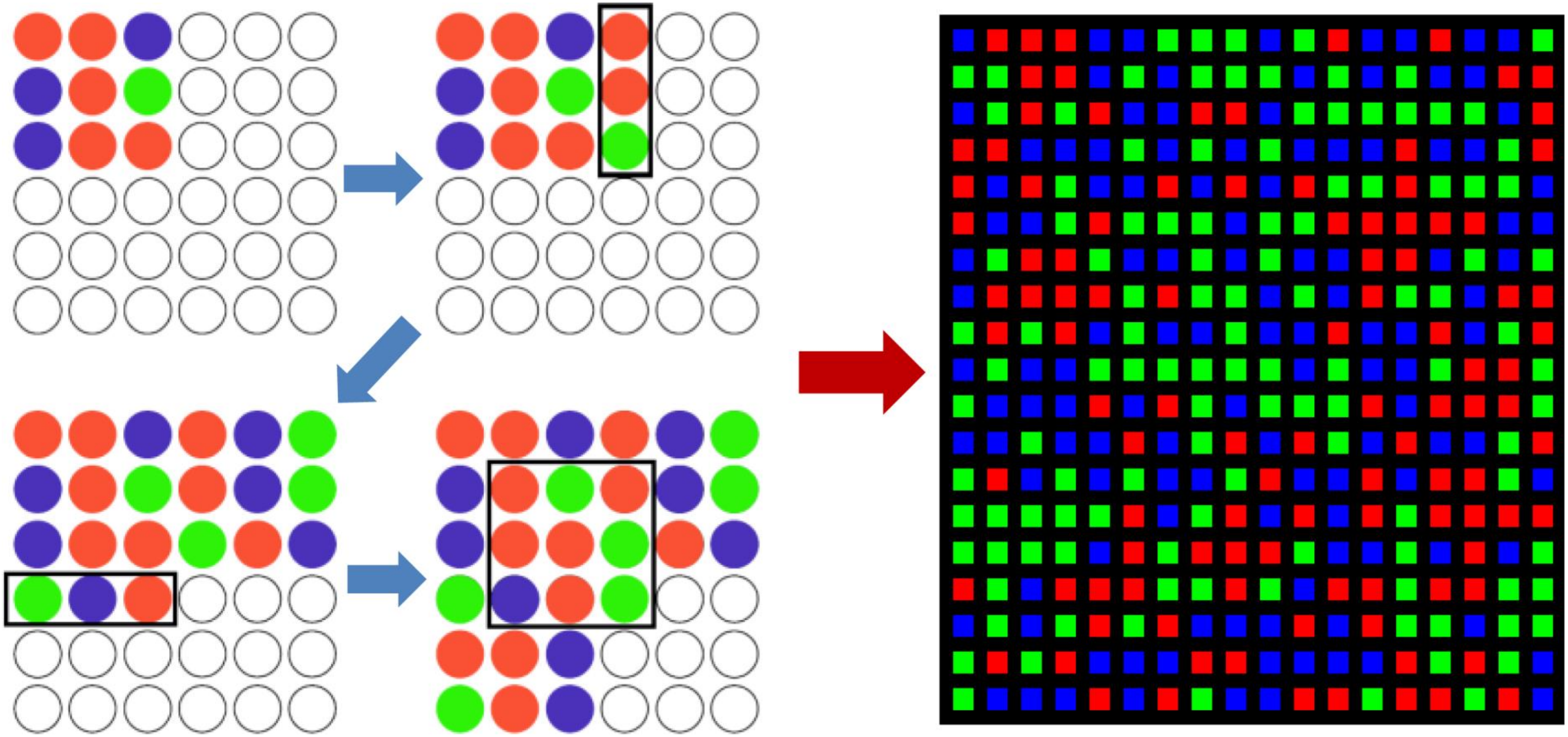
[Zhang 2002]



*Sequential acquisition:*
*7 interlaced targets*

# STRUCTURED LIGHT: UNIQUE “SNAPSHOT” TARGET

2d « snapshot » target by pseudo-random patterns generated using a brute-force algorithm:

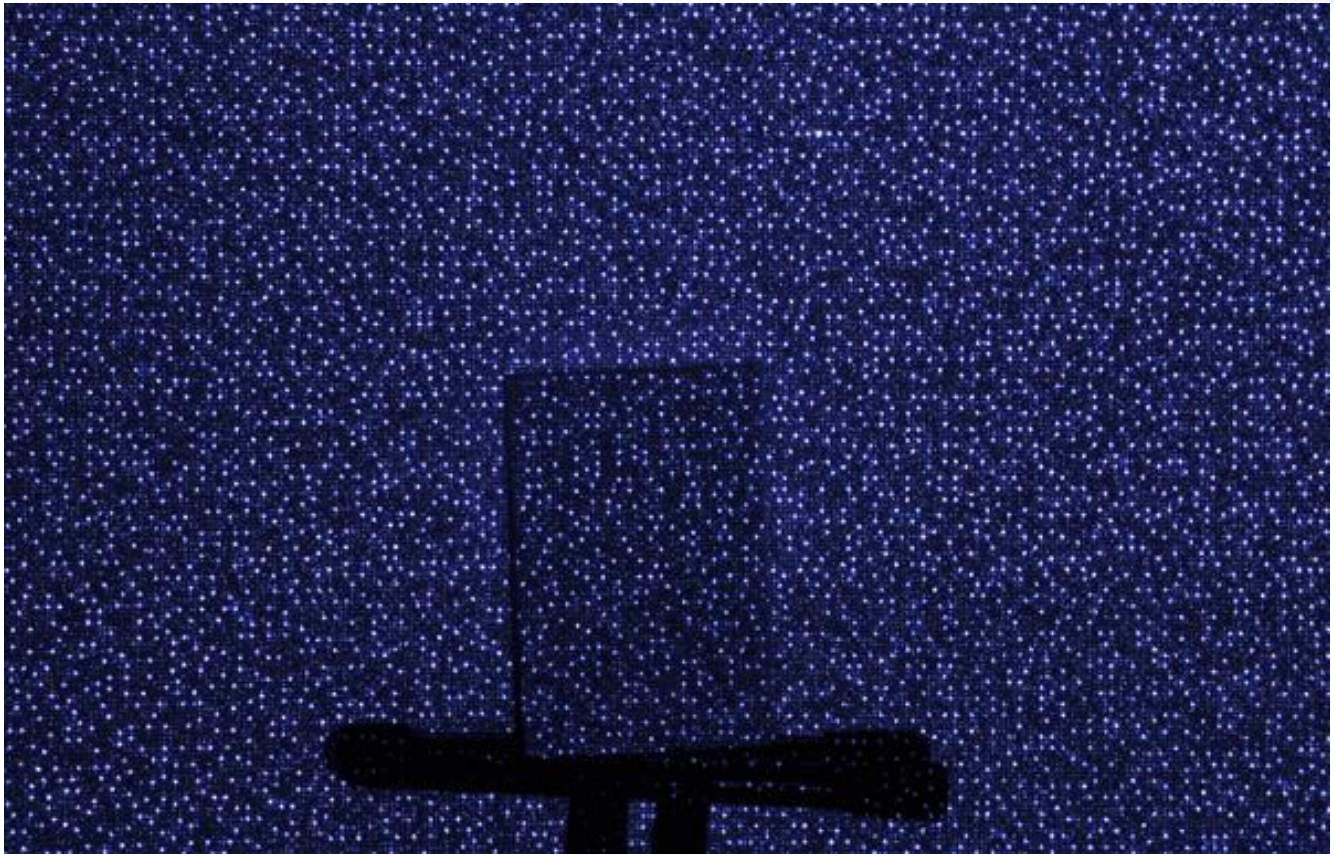


*[Geng 2011]*

# STRUCTURED LIGHT: UNIQUE "SNAPSHOT" TARGET

The first version of the Kinect™ includes an RGB camera associated to a structured light 3d camera using a pseudo-random patterned infra-red light.


[Kinect v1 - © Microsoft]



[© futurepicture.org]

For energy and / or discretion purposes, it may be better for an observation system, not to emit light.

The passive techniques get the information using only the light intensity captured by the photosensors.

The approaches presented in this chapter are all based on a non-pinhole aperture associated to a lens, by making the most of the focus and blur information:

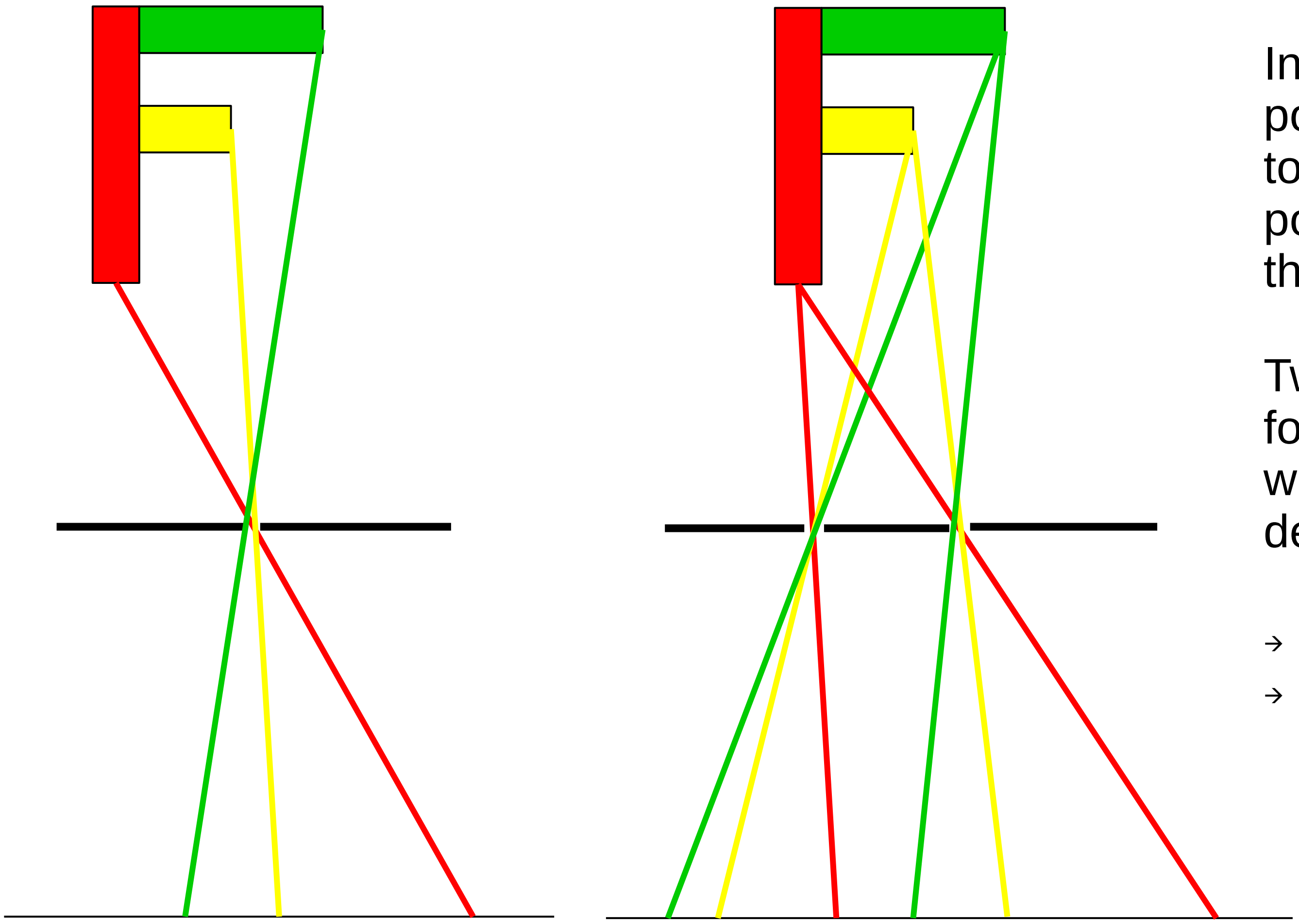
- Plenoptic camera
- Depth from (de)focus
- Coded aperture


***Plenotic camera 3d Raytrix™***


***Light field plenotic camera Lytro™***

ENSTA
IP PARIS

# PASSIVE APPROACHES: PINHOLE…



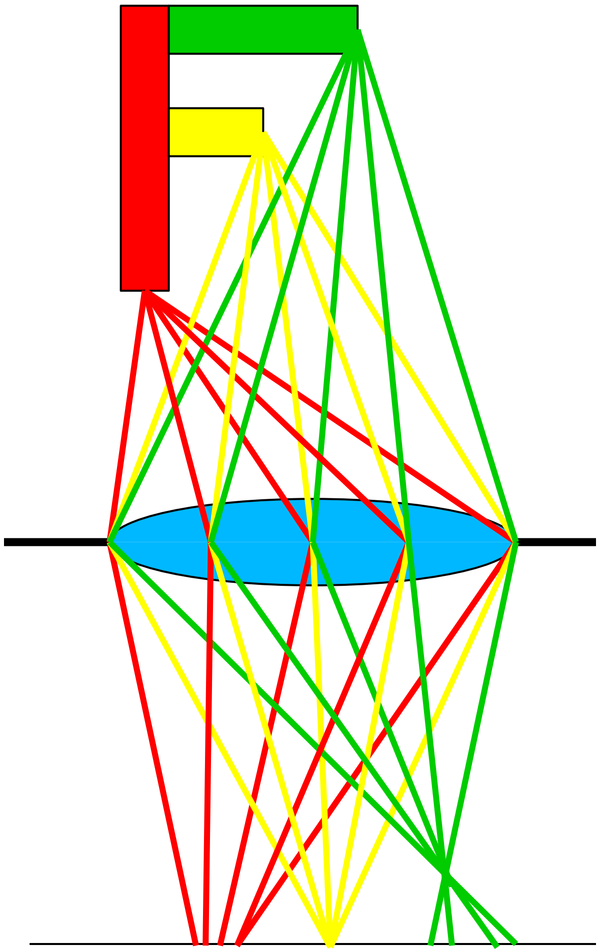
(a) (b)

In a pinhole camera (a), every point of the image corresponds to a unique optical path. All the points appear sharp, whatever their depth.

Two distinct points of view (b), form different images, from which depth information may be deduced.

- → *Stereovision*
- → *Structure from Motion*

# PASSIVE APPROACHES: …*VS* LENS



With a lens on the aperture, each point of the scene illuminates the focal plane along many different optical paths, corresponding to the line beam formed by the cone whose basis is the aperture.

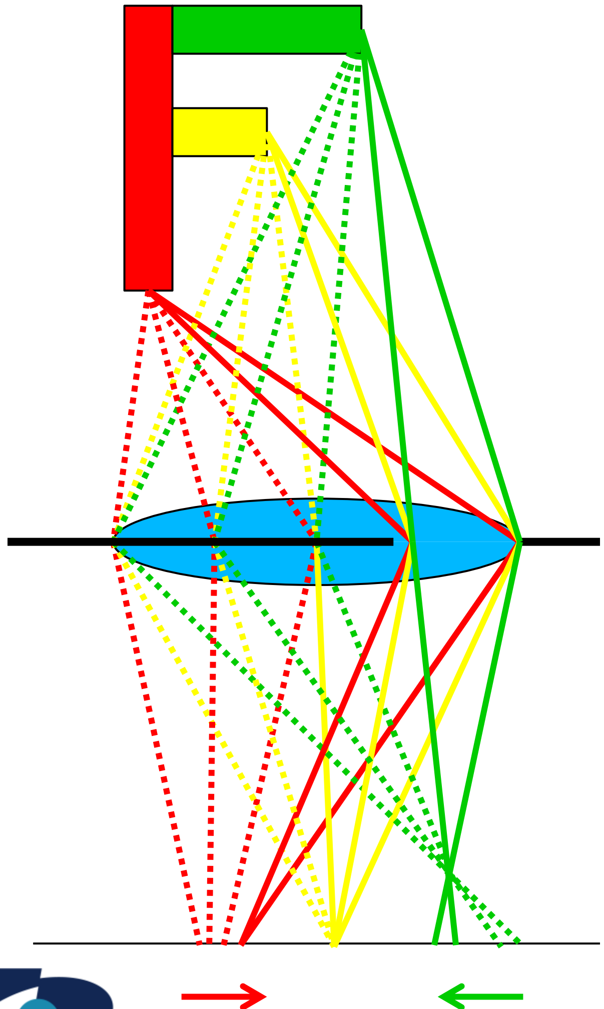
Each path line corresponds to an infinitesimal portion of the aperture, through which the scene is perceived under a particular angle.

Each infinitesimal portion then forms a pinhole-like image, and the image formed by the lens corresponds to the sum of those many « pinhole images ».

If the point is in the conjugate plane of the focal plane (sharpness plane), all the different paths converge on the image, and the point appears sharp, otherwise it appears more or less blurred depending on its distance to the sharpness plane.

→ *Depth from (de)focus*

# PASSIVE APPROACHES: LENS AND APERTURES



By using an excentric aperture (figure), a sub-set of the optical paths is selected, reducing both the blur and the light intensity.

Points in the sharpness plane (yellow lines) remain at the same location in the image plane.

Closer points (**red lines**) are deviated in the direction of the aperture.

Further points (**green lines**) are deviated in the inverse direction.

→ *Coded aperture:*
*Modify the geometry of the aperture for an easier interpretation of the blur (≈ point spread function of the aperture).*

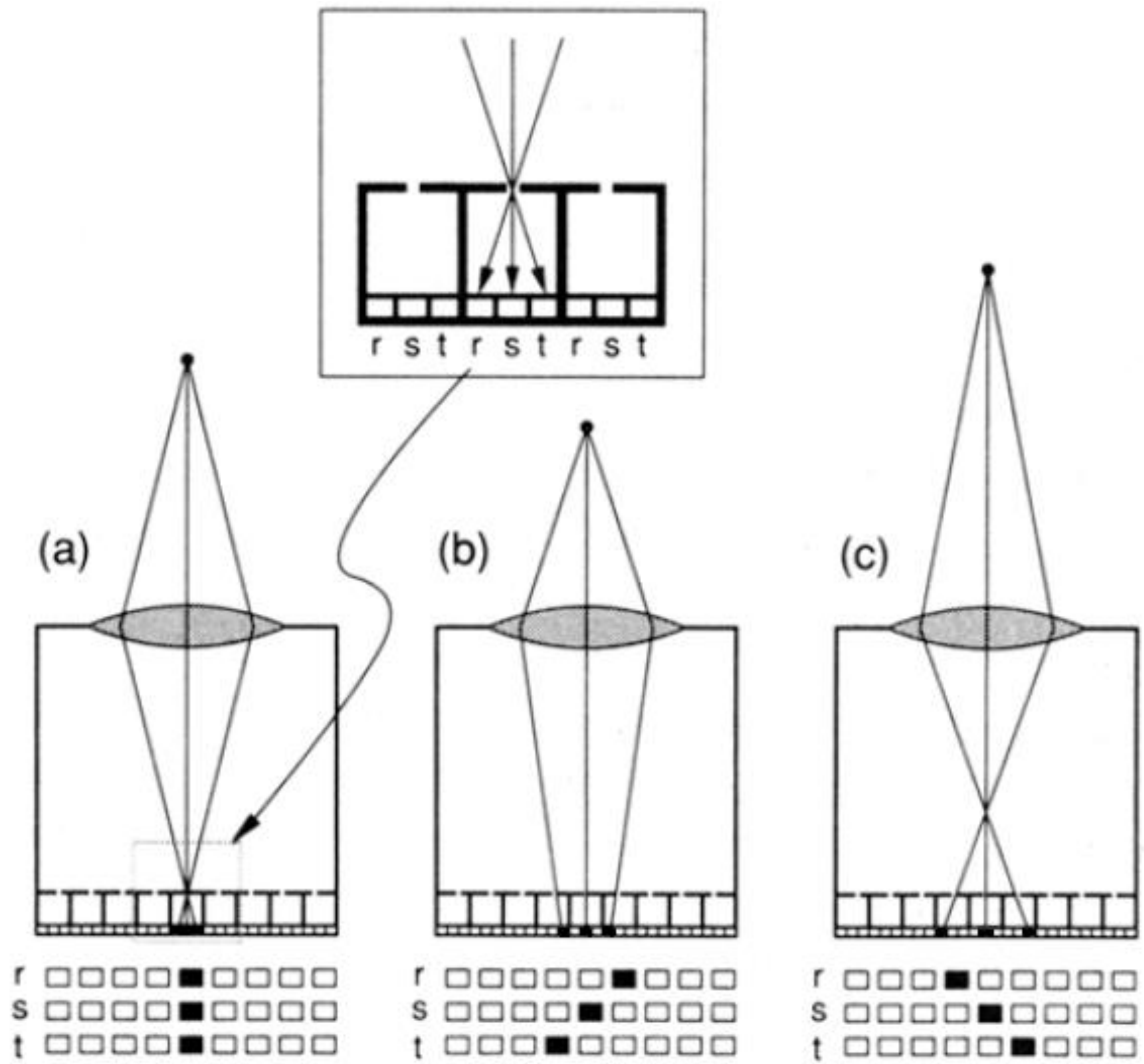
→ *Plenoptic camera:*
*Separate physically the different optical paths within sub-beams focalised on distinct parts of the sensor.*

# PASSIVE 3D: PLENOPTIC CAMERA

In a plenoptic camera, the optical paths are separated within sub-beams that are focalised on different parts of the sensor.
(Figure: mini-pinholes, but also 1d lenticular grid, or 2d micro-lens grid).

The captured information is then composed of one macro-image made of many hyper-pixels (or micro-images).
(See figure:
- Macro-image of 1x9 hyper-pixels.
- Hyper-pixel of size 1x3.)

The plenoptic image then captures a 4d information: $I(x,y,\xi,\varsigma)$, where $(x,y)$ is the direction of a point illuminating the aperture (light cone), and $(\xi,\varsigma)$ a particular view of this point through the aperture.

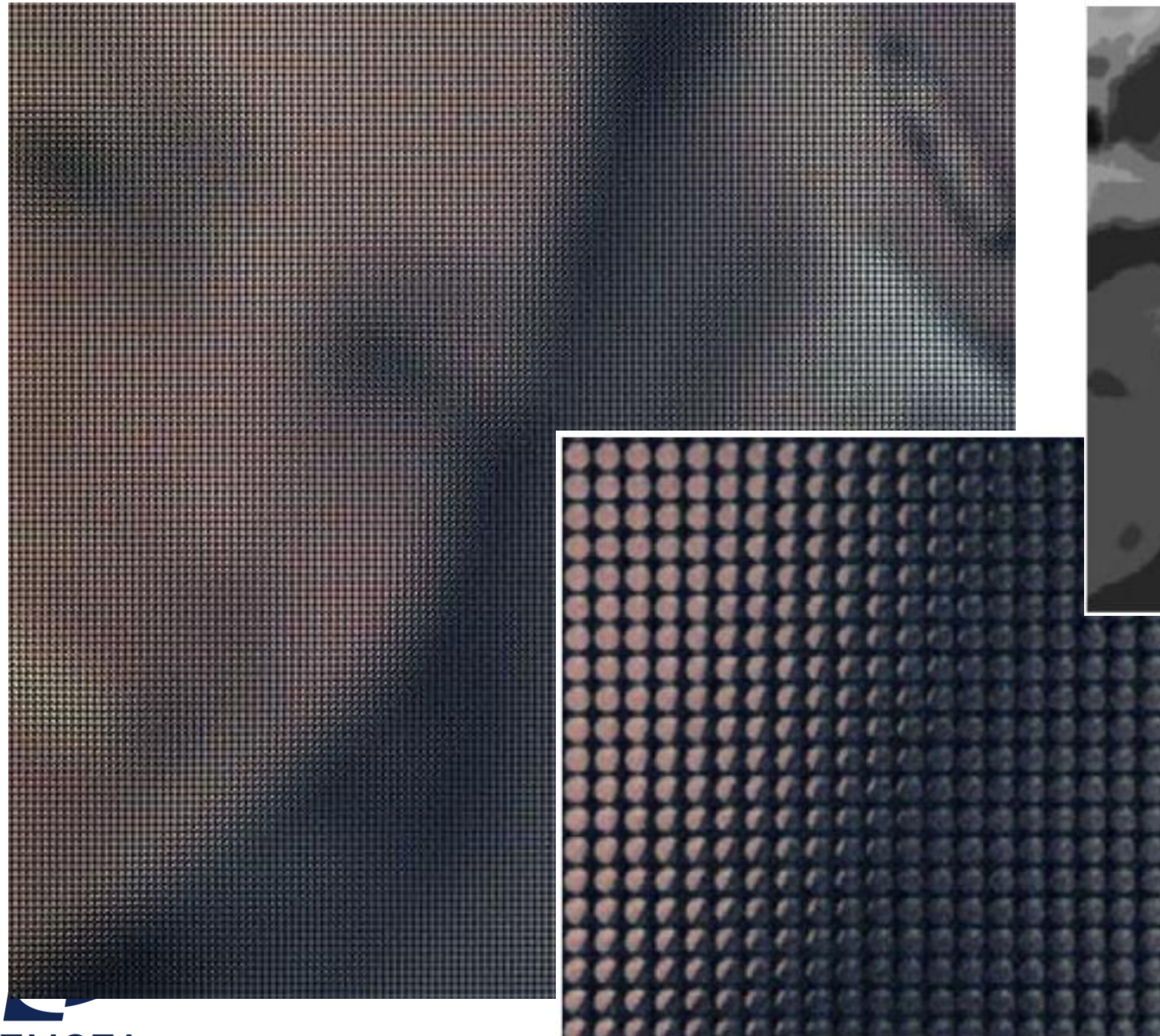



[Adelson 1992]

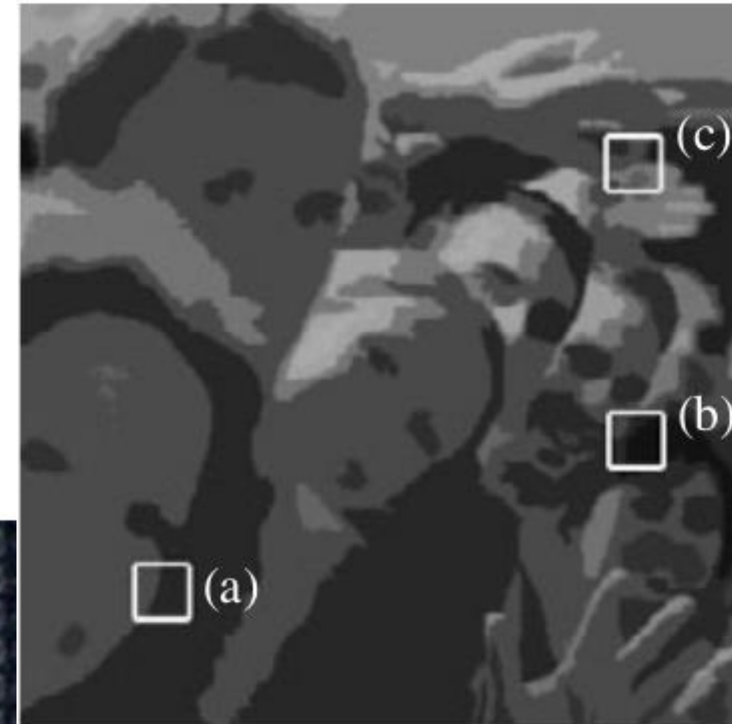
# PLENOPTIC CAMERA: MACRO-IMAGE

*[Ng 2005]*

# PLENOPTIC CAMERA: MICRO-IMAGES



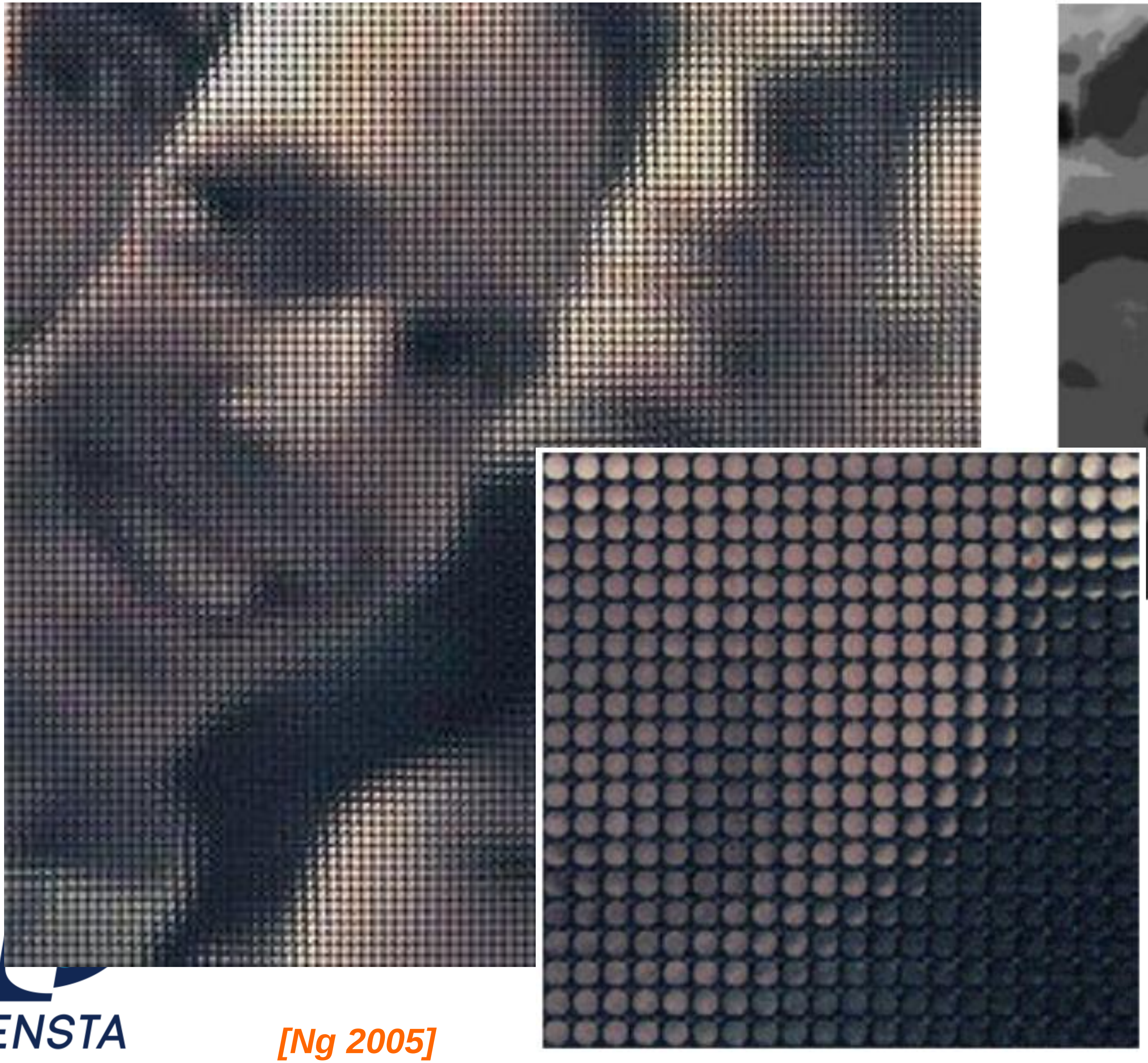
(a)




Micro-images (hyper-pixels) of points that are closer than the sharpness plane *have a contrast in the same sense* as the macro-image.

[Ng 2005]

# PLENOPTIC CAMERA: MICRO-IMAGES



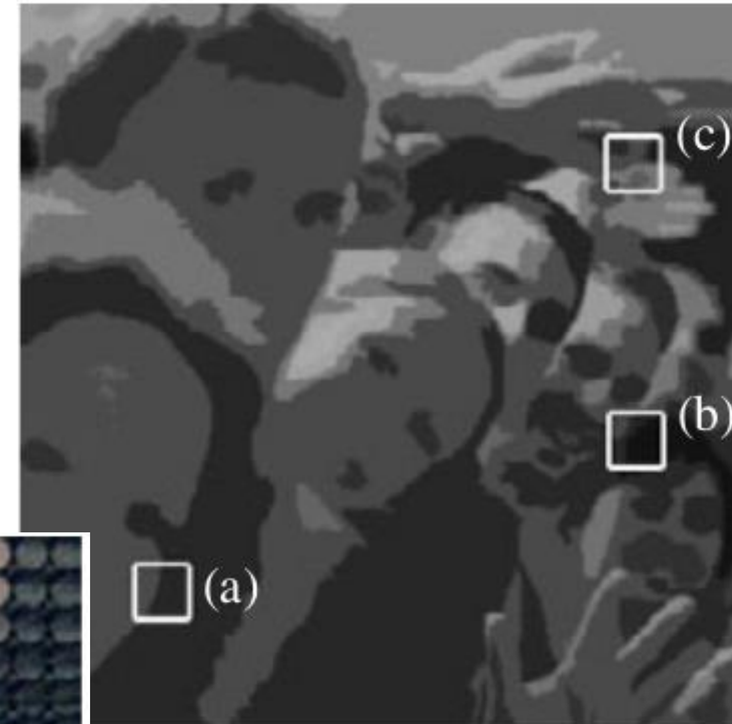

(b)

Micro-images (hyper-pixels) of points that are further than the sharpness plane *have an inverted contrast* with respect to the macro-image.

*[Ng 2005]*

# PLENOPTIC CAMERA: MICRO-IMAGES

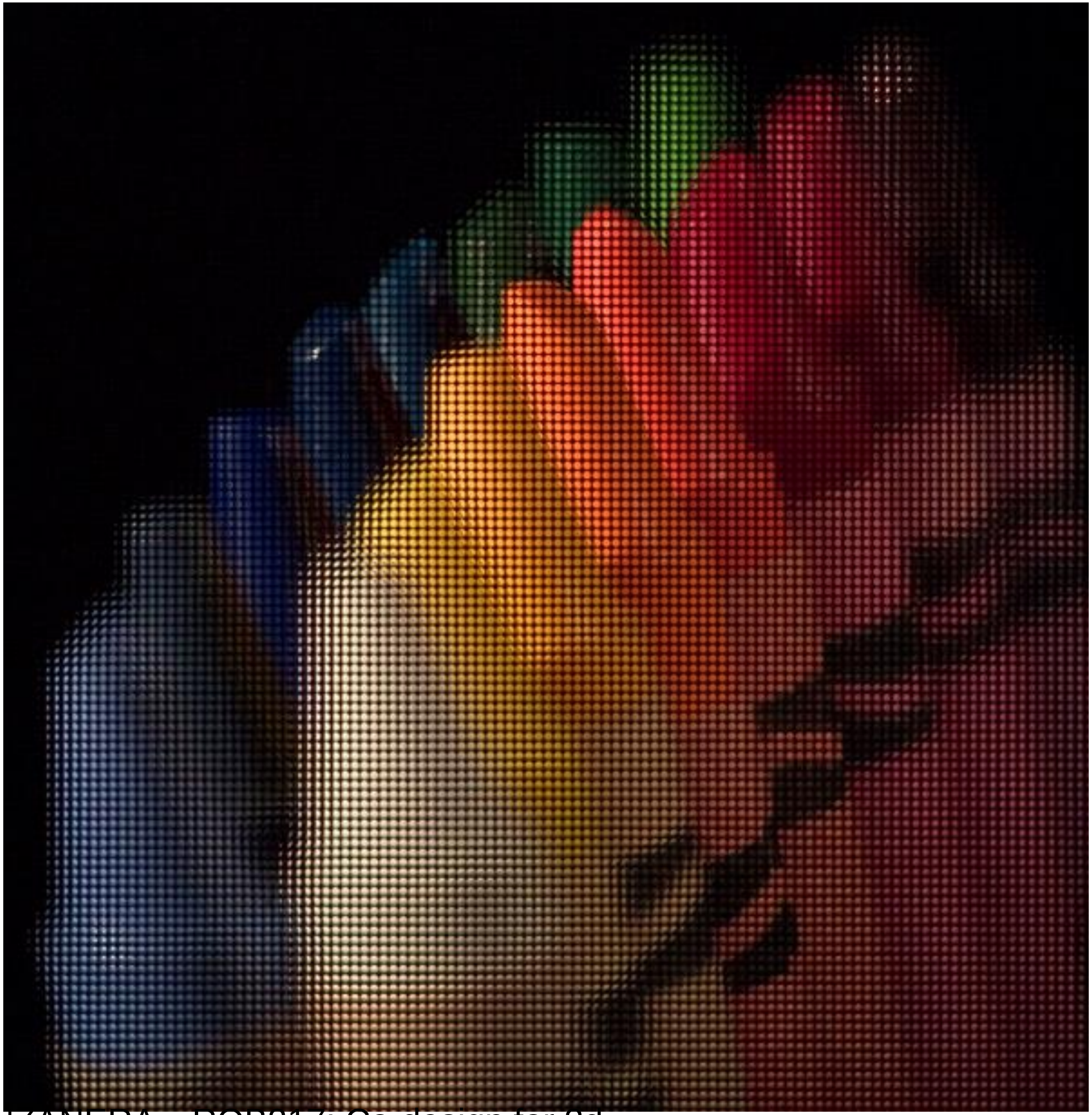
(c)




Micro-images (hyper-pixels) of points in the sharpness plane *form homogeneous regions.*

[Ng 2005]

# PLENOPTIC CAMERA: MACRO-IMAGE

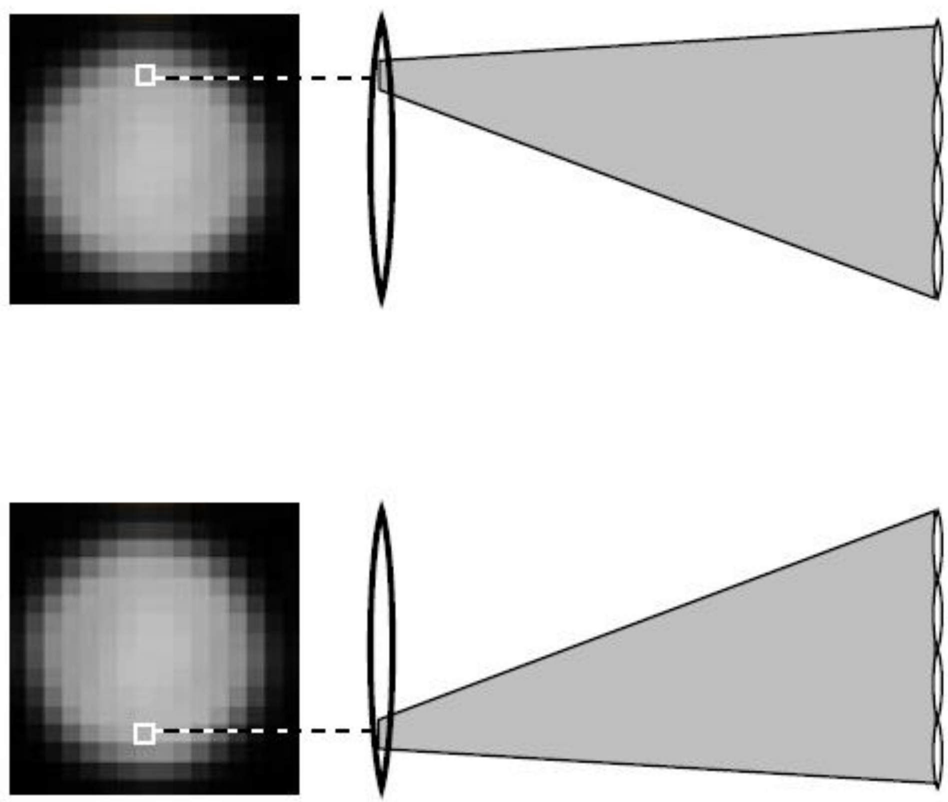


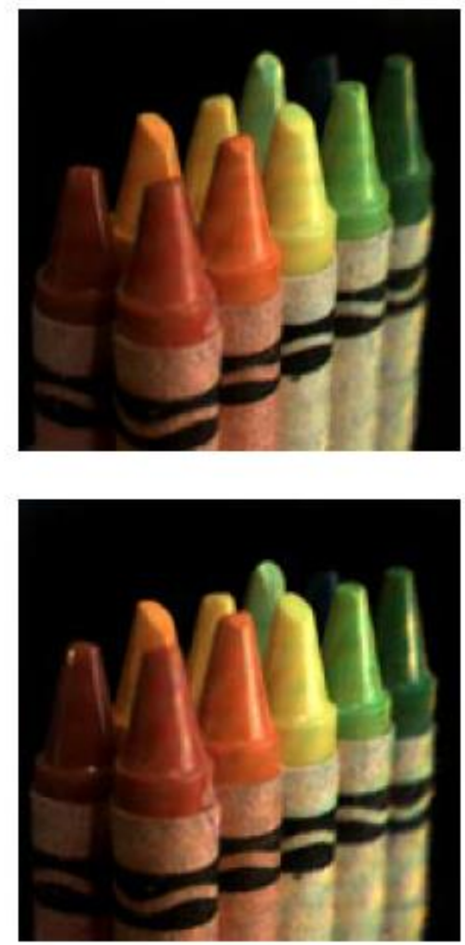
*[Ng 2005]*

# PLENOPTIC: MACRO-IMAGE AND DUAL MACRO-IMAGES

Dual macro-images are made by recomposing *m* x *m* sub-sampled images of size *n* x *n* from the homologous pixels of all the micro-images, where *n* x *n* is the number of micro-images (resolution of the macro-image), and *m* x *m* is the resolution of the micro-image.

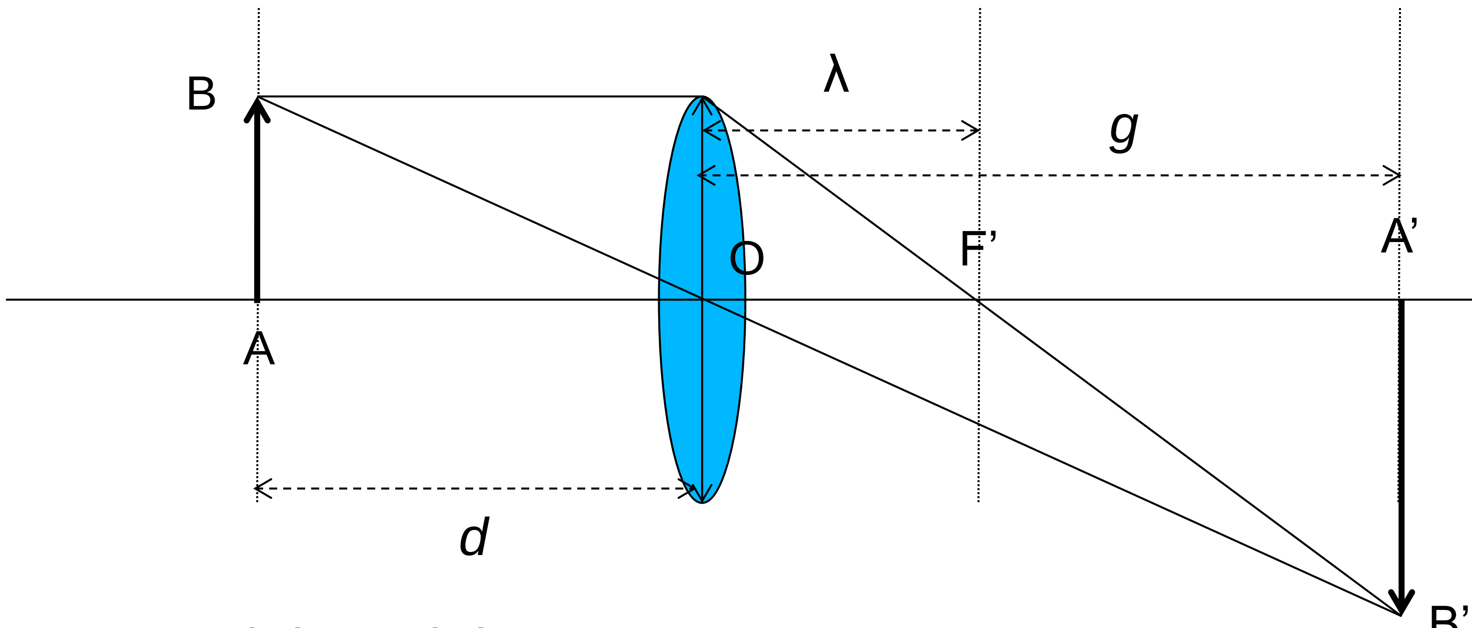
Dual macro-images then correspond to a partition of the aperture into distincts viewpoints and then present parallax differences, from which depth information can be deduced by matching (*single-lens stereo*).





***[Ng 2005]***

# GEOMETRY OF THE THIN CONVERGENT LENS




λ: focal distance of the lens

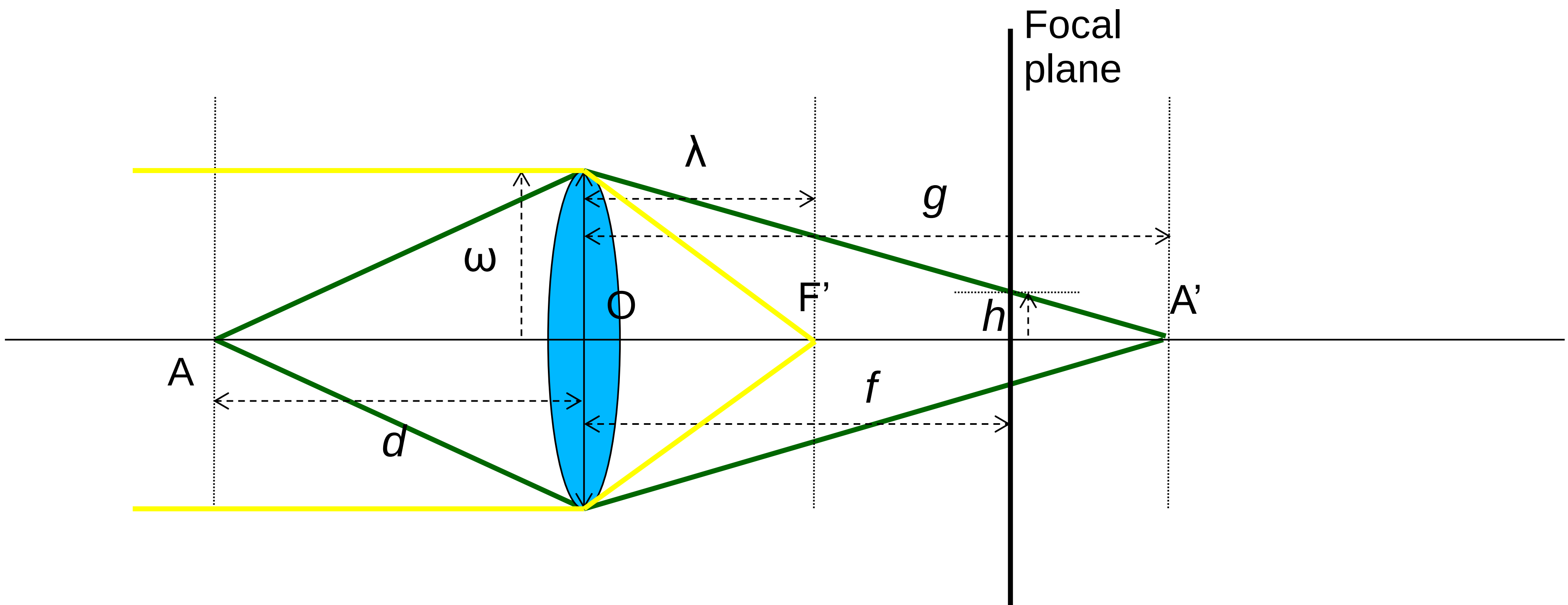
$$\frac{OA}{OF'} = \frac{OA}{OA'} + 1$$

and then: $$\boxed{\frac{1}{d} + \frac{1}{g} = \frac{1}{\lambda}}$$

Thin lens equation

- *d*: distance of point B to the aperture plane (depth)
- *g* : distance between the aperture and the focalisation plane of point B

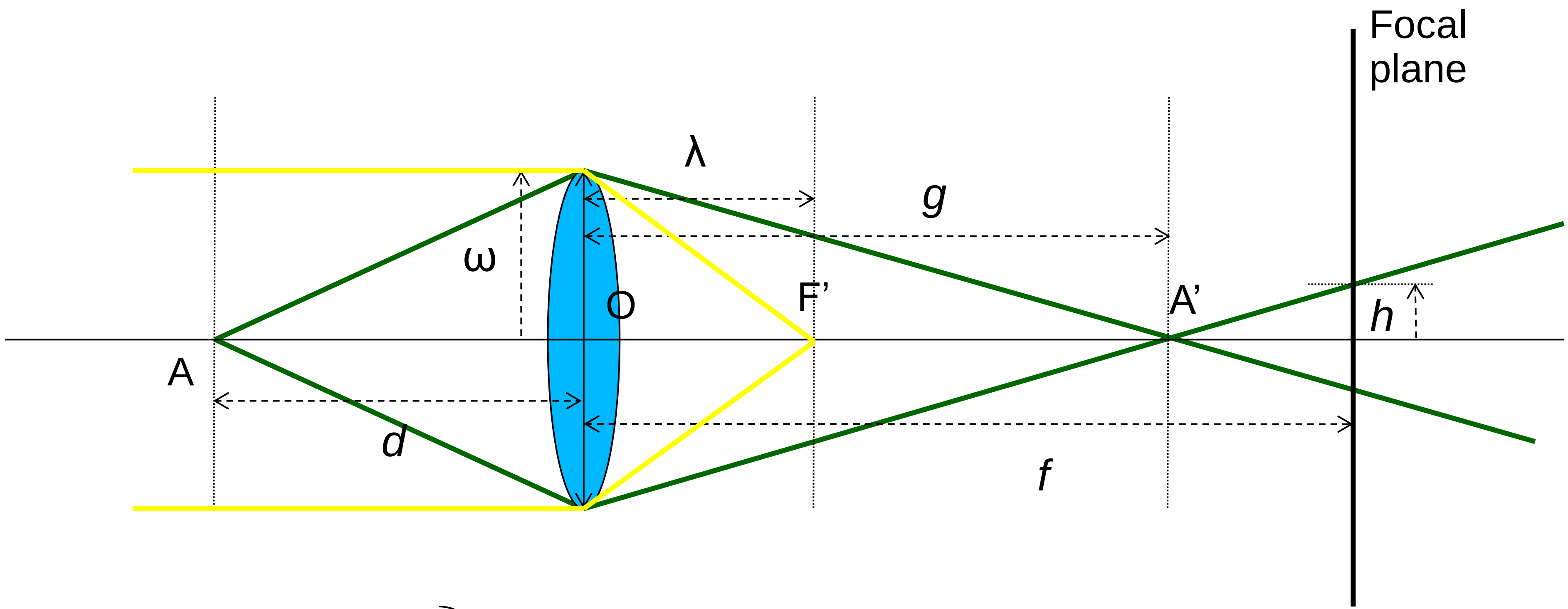
# RELATION FOCUS / DISTANCE: SHORT FOCAL




$$\left.\begin{aligned} \frac{f}{g} &= 1 - \frac{h}{\omega} \\ \frac{1}{d} + \frac{1}{g} &= \frac{1}{\lambda} \end{aligned}\right\} \quad \frac{1}{d} = \left(\frac{1}{\lambda} - \frac{1}{f}\right) + \frac{h}{f\omega}$$

λ: lens focal
ω: aperture
*f*: image focal
*h*: defocus width

# RELATION FOCUS / DISTANCE: LONG FOCAL



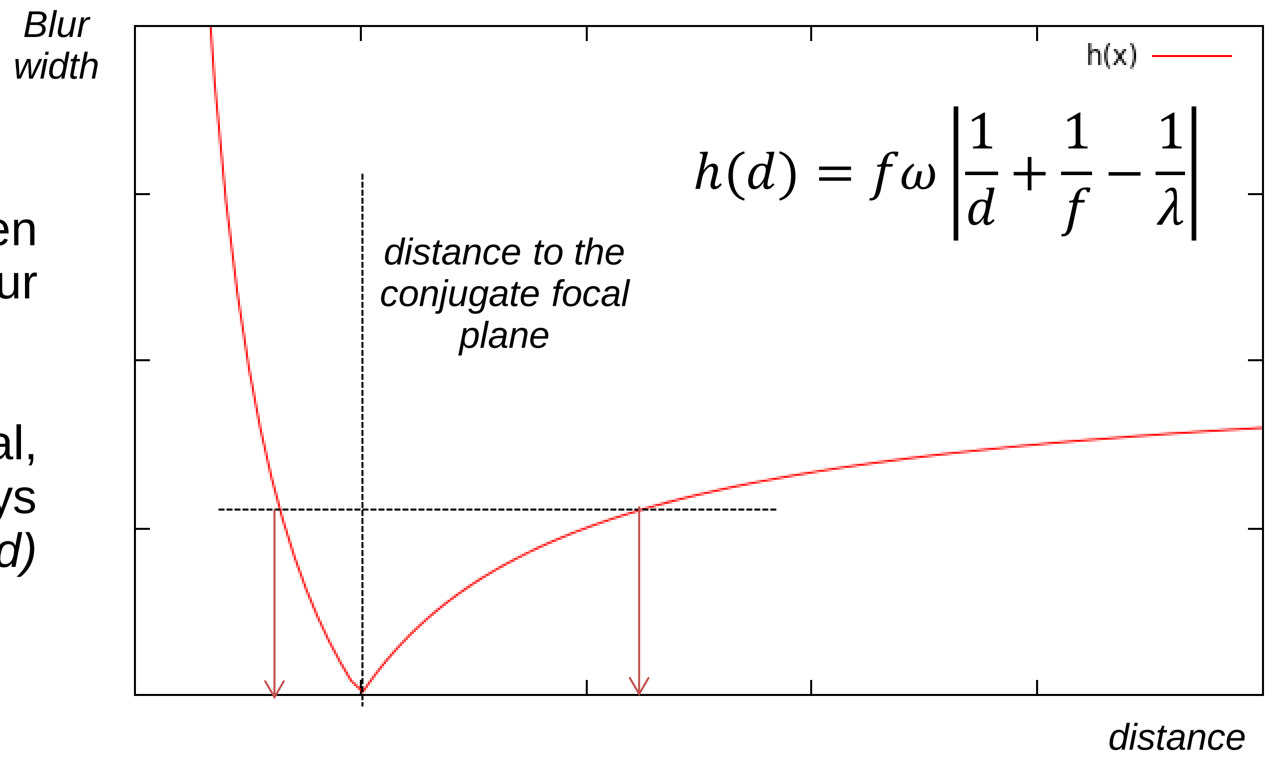

$$\left.\begin{aligned} \frac{f}{g} &= 1 + \frac{h}{\omega} \\ \frac{1}{\lambda} &= \frac{1}{d} + \frac{1}{g} \end{aligned}\right\} \quad \frac{1}{d} = \left(\frac{1}{\lambda} - \frac{1}{f}\right) - \frac{h}{f\omega}$$

λ: lens focal
ω: aperture
*f*: image focal
*h*: defocus width

# PASSIVE 3D: DEPTH FROM (DE)FOCUS

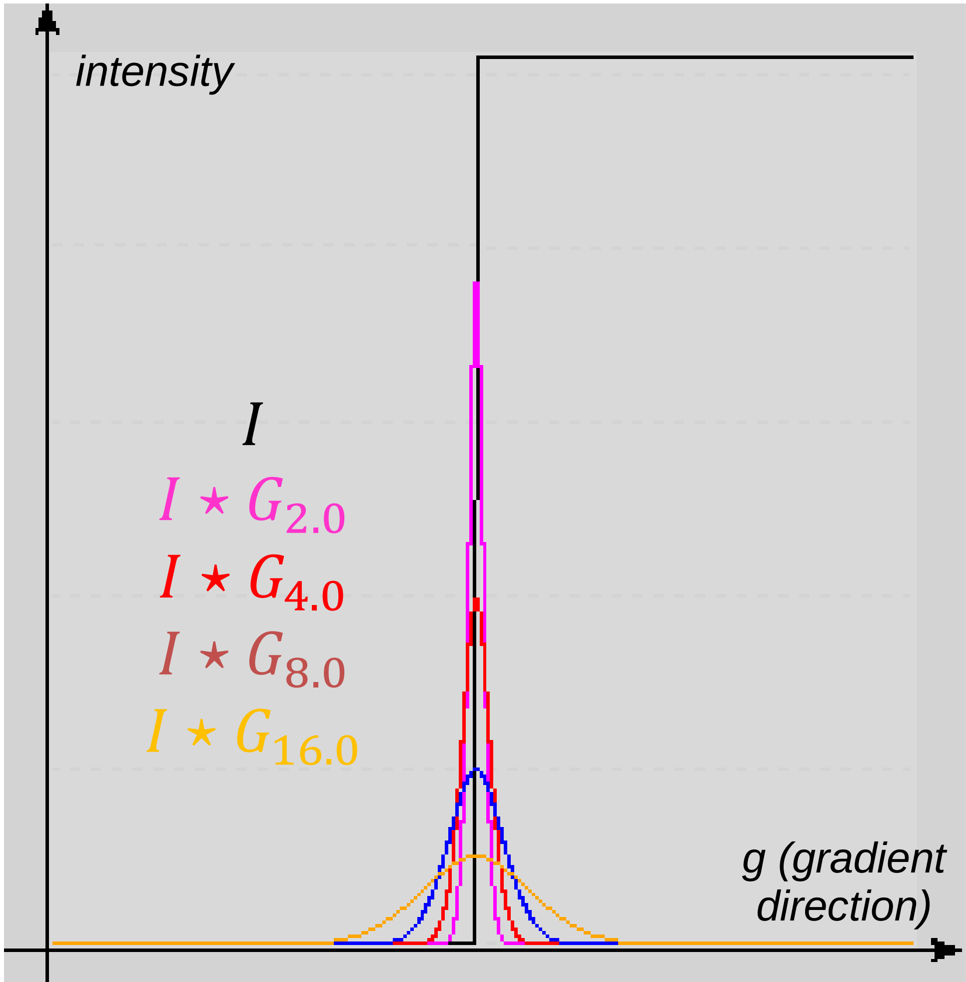
Estimating the distance can then be made by estimating the blur width in the image.

Without prior on the image focal, a single measure is always ambiguous, the function *h(d)* being not injective (Figure).




To perform direct measurement of the blur width by image processing, an hypothesis on the structure of the sharp image is necessary: impulsion, step-like contour, in order to predict the effect of blur on this structure.

# PASSIVE 3D: DEPTH FROM (DE)FOCUS




If the blur is modelled by a 2d Gaussian convolution whose standard deviation depends on *h*, *h* can be deduced from the effect of blur on a step-like contour structure, by measuring the local maximum of the gradient value in the direction orthogonal to the step.

Those structures correspond to the classic definition of contours, i.e. the zero-crossings of the second derivative in the gradient direction *g*:

$$C_I = \left\{ x; \frac{\partial^2 I}{\partial g^2}(x) = 0 \right\}$$

*Question*: how to justify the use of a Gaussian blur model when the geometric optics predicts a gate (square) function?

# PASSIVE 3D: DEPTH FROM (DE)FOCUS

$$I(x) = (I^H(x), I^S(x), I^V(x))$$

$\frac{\partial I^V}{\partial g}(x)$ (gradient magnitude)

# PASSIVE 3D: DEPTH FROM (DE)FOCUS

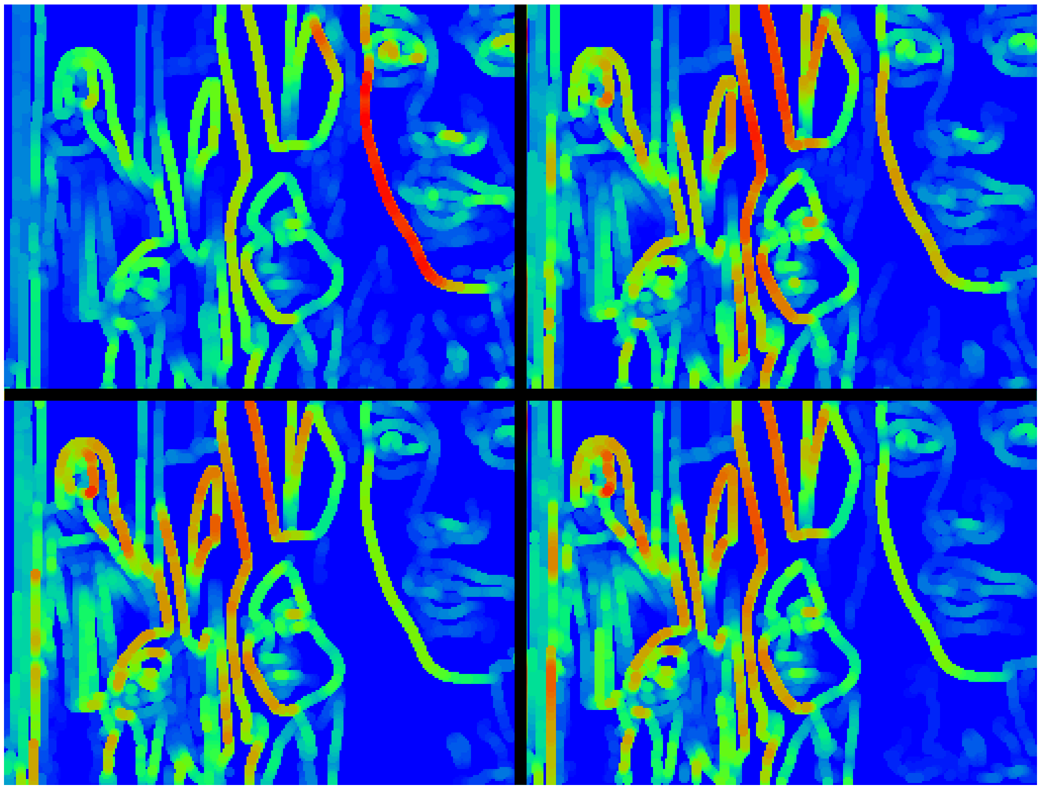
$I(x)$

$C_I = \left\{x; \frac{\partial^2 I^V}{\partial g^2}(x) = 0\right\}$ (contours)

# PASSIVE 3D: DEPTH FROM (DE)FOCUS

Measuring the gradient magnitude along the contours allows estimating the blur width *h*, but remains ambiguous regarding the position with respect to the sharpness plane.

$I(x)$



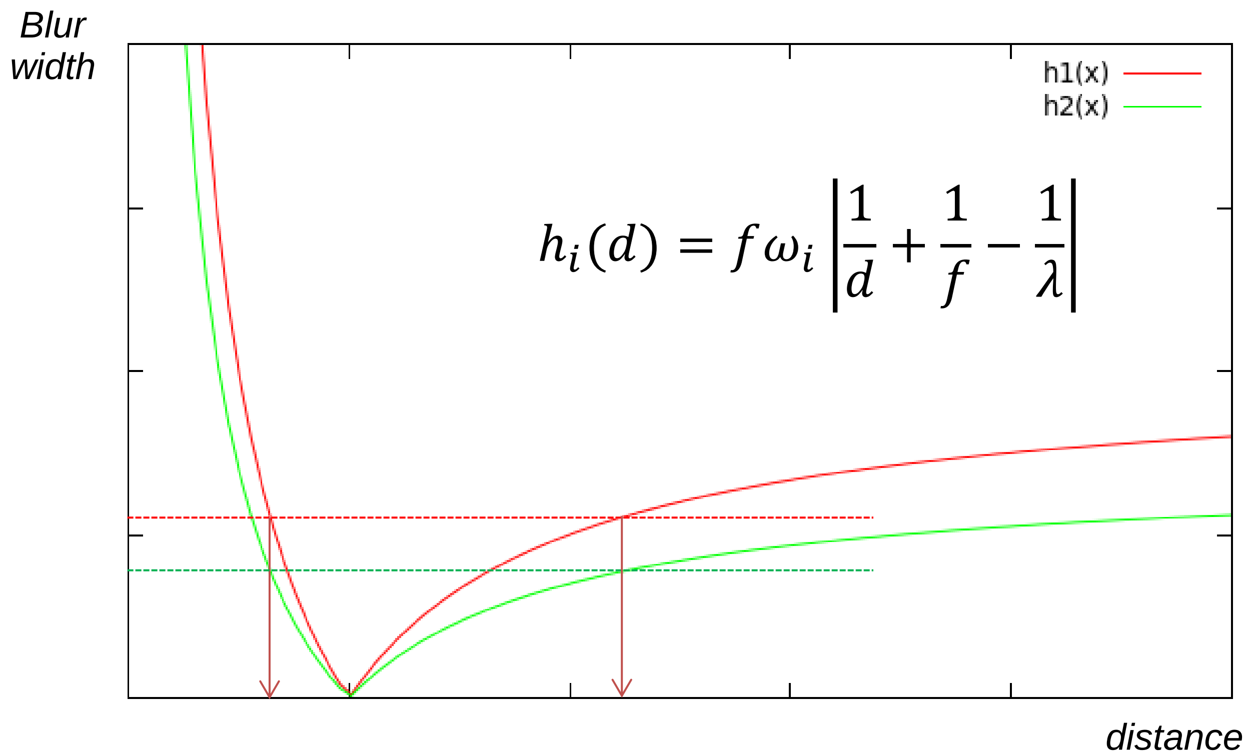
Mesuring the blur width along the contours

***[Pentland 1987]***

<u>Idea</u>: repeat the measure while varying the aperture $\omega$ and/or the image focal $f$ ?

# PASSIVE 3D: DEPTH FROM (DE)FOCUS

The blur width depends linearly of the aperture, then using different apertures only does not disambiguate the distance from the sharpness plane:





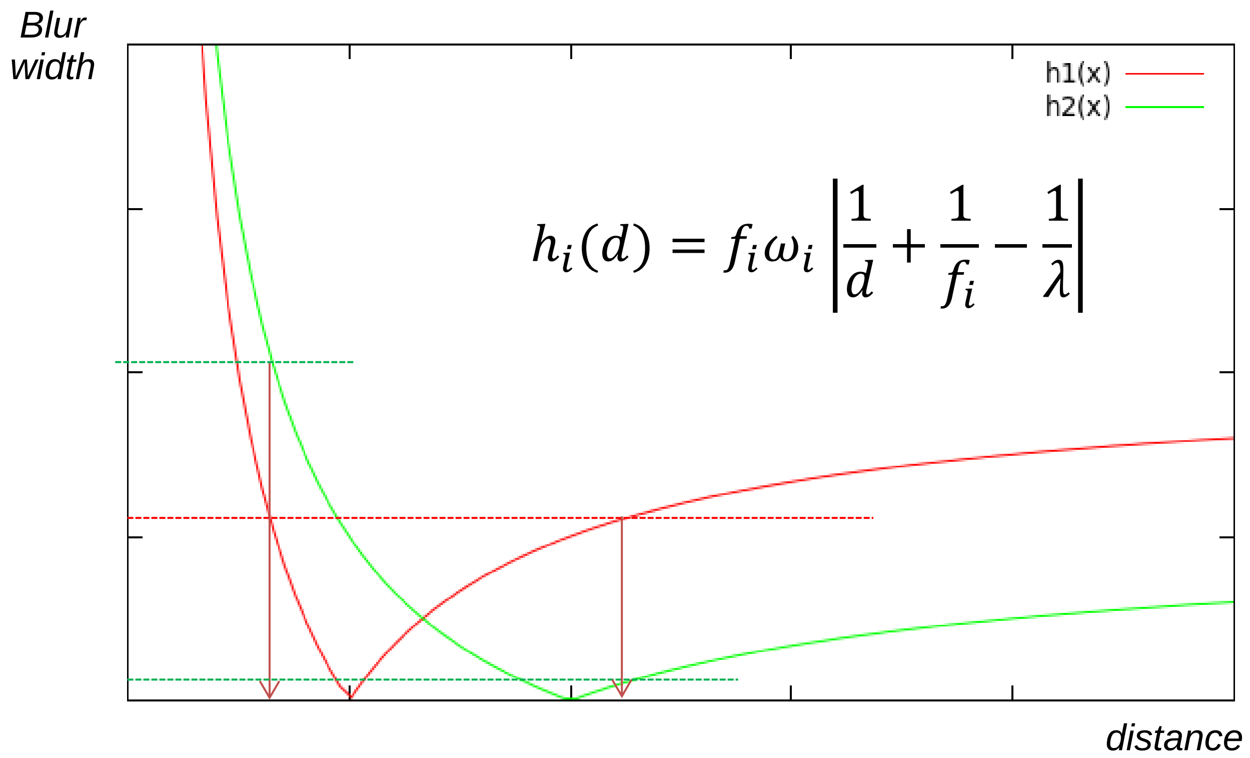
*Constant focal, variable aperture.*

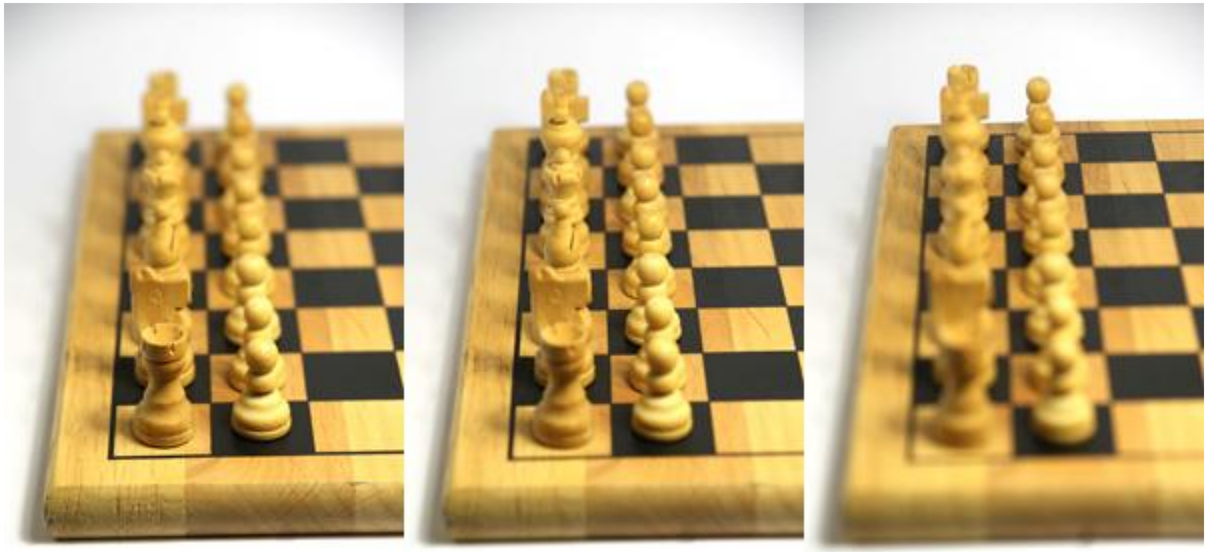
# PASSIVE 3D: DEPTH FROM (DE)FOCUS

In contrast, using several couples (aperture, image focal) allows to deduce the distance from the blur width in an absolute manner.

(Figure: product $f_i \omega_i$ constant)

*[Pentland 1987]*



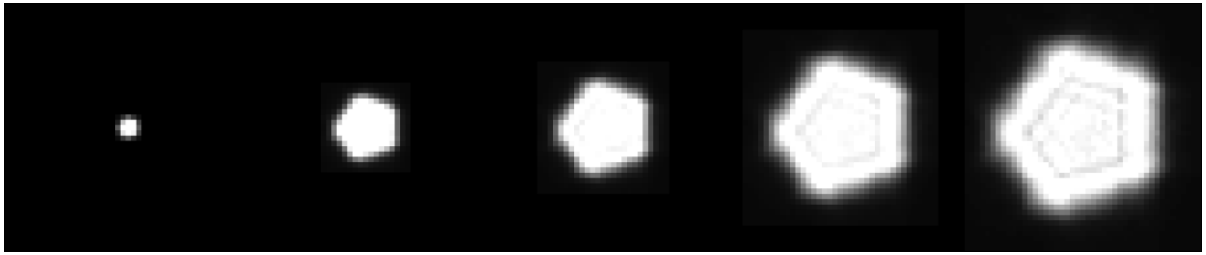
*Aperture constant, variable focal.*

# BLUR MODEL *VS* APERTURE CALIBRATION

The Gaussian kernel is considered a better blur model than the gate function because the blur is actually the combination of several phenomena: diffraction, chromatic aberrations, discretisation, that lead to the composition of several convolutions.

However a better alternative to blur models is to perform an aperture calibration of the camera by recording the different images formed by one point for different focalisation distances (point spread functions of the convolution kernels).

***[Levin 2007]***

*Traditional 5-blade diaphragm and the family* $\{g_d\}_{d \in D}$ *of calibrated kernels.*

Estimating the right distance is then equivalent to finding the kernel $g_d$ which best corresponds to the local observation.

The « direct » estimation being only possible on contours, indirect estimation is used instead, using deconvolution…

# BLUR ESTIMATION BY DECONVOLUTION

$I$ the observed image

$\{g_d\}_{d \in D}$ the family of calibrated convolution kernels, indexed by distance

$J_d$ the deconvolution of $I$ by $g_d$

The reconstruction error $\varepsilon_d(x)$ at pixel $x$ and distance $d$ is defined as:

$$\varepsilon_d(x) = \sum_{y \in W_x} \|I - J_d \star g_d\|^2$$

where $W_x$ is a spatial neighbourhood of $x$.

Distance estimation is then performed as follows:

$$d_{opt}(x) = \arg\min_{d \in D} \varepsilon_d(x)$$

# DECONVOLUTION: INVERSE AND WIENER FILTERING

The problem is now equivalent to image deconvolution (restoration), the convolution kernel at the origin of the blur being known (non-blind).

Quick sketch of non-blind deconvolution:

$$F = I \star g_d \xrightarrow{\text{Fourier transform}} \tilde{F} = \tilde{I} \times \widetilde{g_d} \xrightarrow{\text{Inverse filter}} \widetilde{J_d} = \frac{\widetilde{F}}{\widetilde{g_d}} \xrightarrow{\text{Inverse Fourier transform}} J_d$$

Not usable because of the zeros of $\widetilde{g_d}$ and additive noise!!!

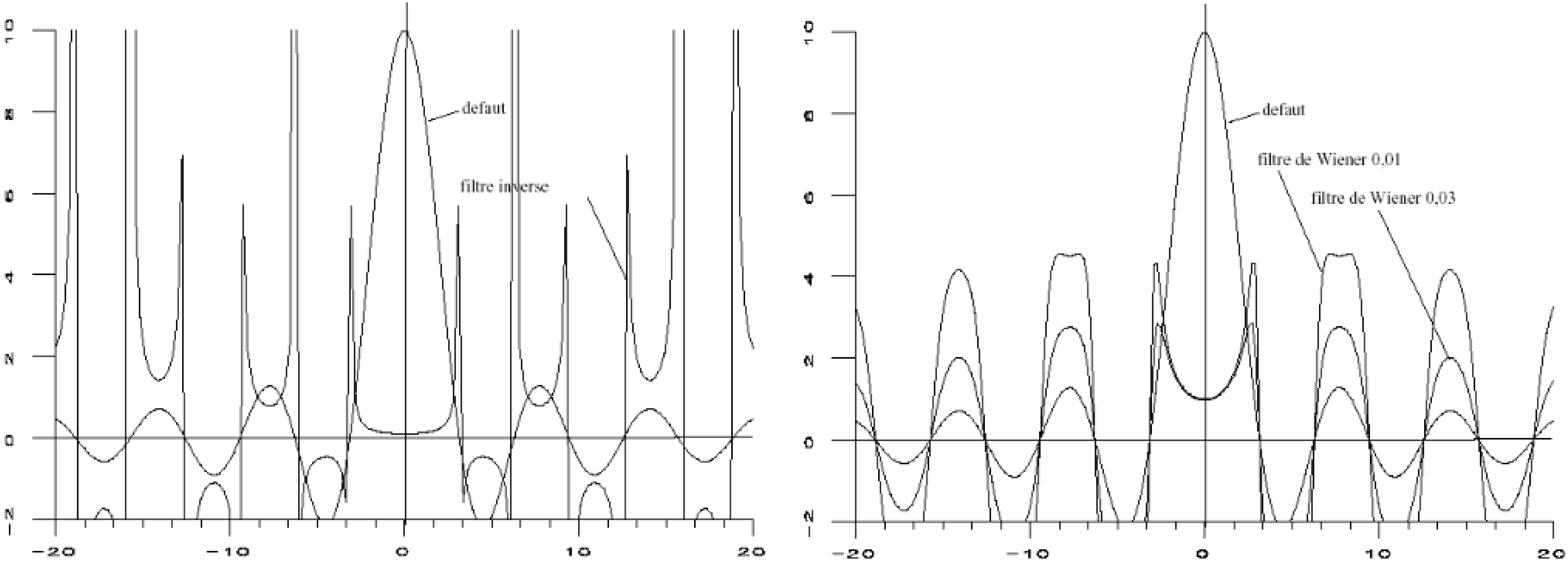
$$F = I \star g_d + b \xrightarrow{\text{Fourier transform}} \tilde{F} = \tilde{I} \times \widetilde{g_d} + \tilde{b} \xrightarrow{\text{Wiener filter}} \boxed{\widetilde{J_d} = \frac{\widetilde{g_d}' \times \tilde{F}}{\widetilde{g_d}\widetilde{g_d}' + \alpha}} \xrightarrow{\text{Inverse Fourier transform}} J_d$$

$$\alpha \approx \frac{\left\langle \left| \tilde{b}(u) \right|^2 \right\rangle}{\left\langle \left| \tilde{I}(u) \right|^2 \right\rangle}$$

$\alpha$ is a regularisation term, which depends on the relative power of noise *b* with respect to image signal *I*. It can be set as constant or depend on frequencies: $\alpha(u)$. Wiener filtering thus performs a trade-off between deconvolution and regularisation.

**In any case, the reconstruction error $\boldsymbol{\varepsilon_d}$ strongly depends on the zeros of the convolution filter in the frequency domain ($\widetilde{g_d}$ ).**

# DECONVOLUTION: INVERSE AND WIENER FILTERING




Left: a (constant speed) motion blur in the frequency domain (cardinal sine), and the corresponding inverse filter.
Right: the same default and the correcting Wiener filters for two different values of $\alpha$ assumed constant.

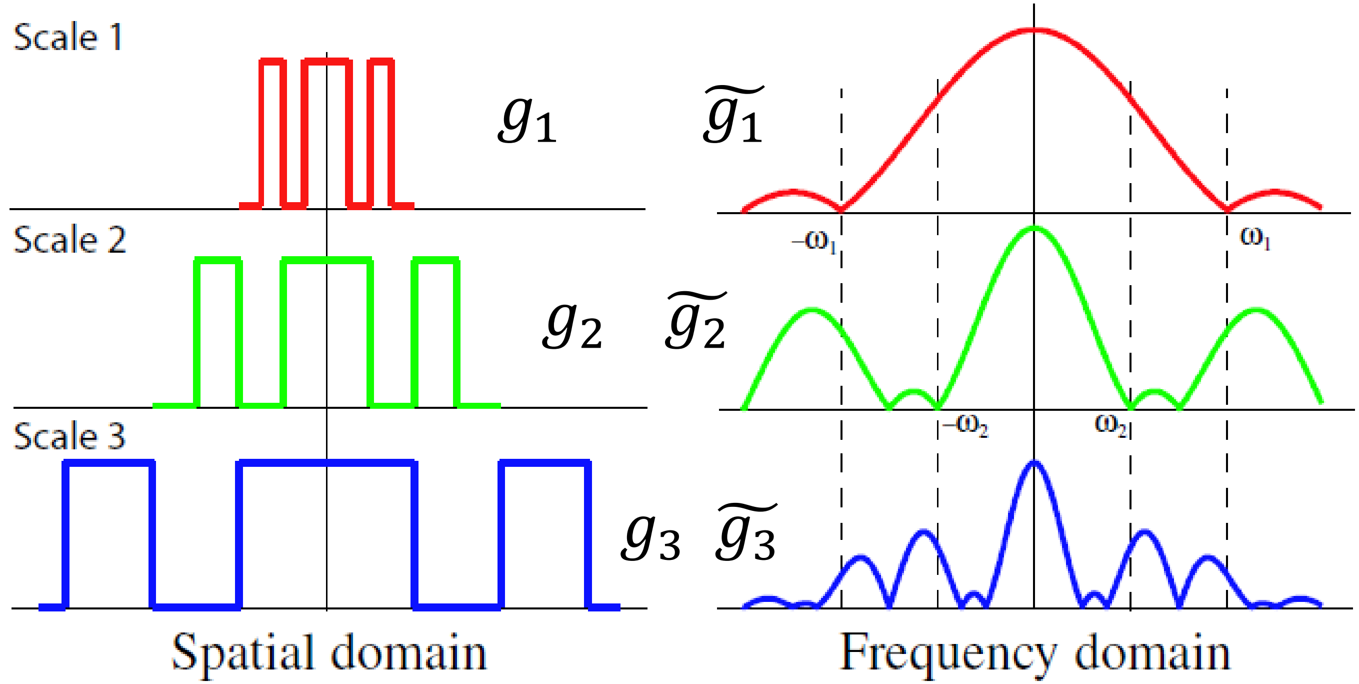
*[Figure: Maître 2003]*

# PASSIVE 3D: CODED APERTURE

In deconvolution techniques, the zeros of the filter in the frequency domain are those that mainly contribute to the reconstruction errors.

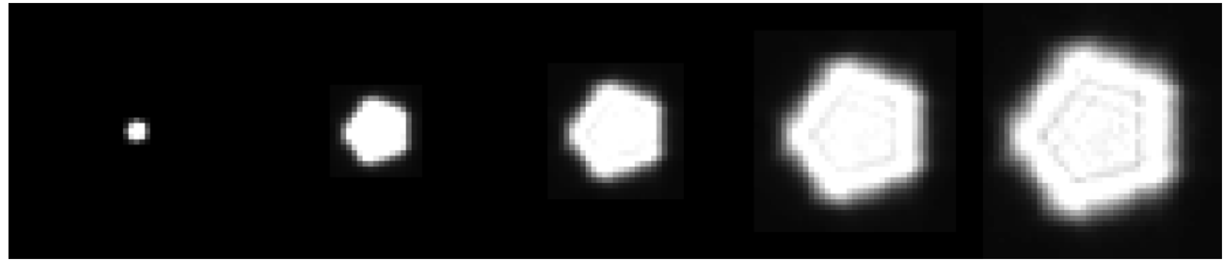
As a consequence, if the different convolution kernel candidates $\{g_d\}_{d \in D}$ have their zeros located at the same frequencies in the Fourier domain, it is much more difficult to distinguish their effects on the image (by deconvolution) than if their zeros appear at different locations.

The principle of coded aperture is to choose the shape of the aperture in such a way that the zeros of the different filters $\{g_d\}_{d \in D}$ appear, depending on *d*, at different location of the frequency domain:

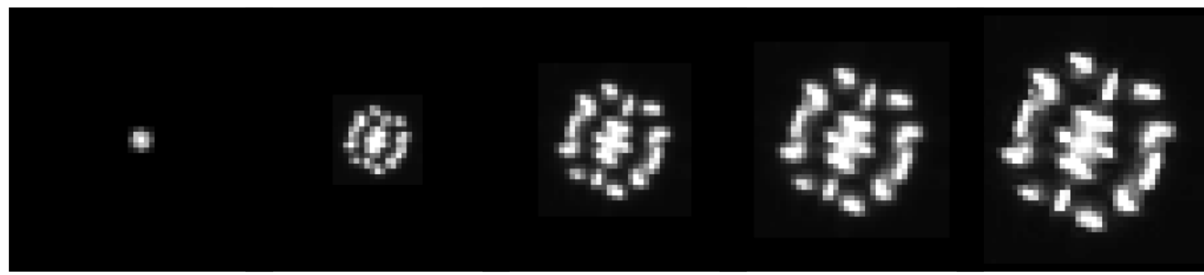



[Levin 2007]

# PASSIVE 3D: CODED APERTURE

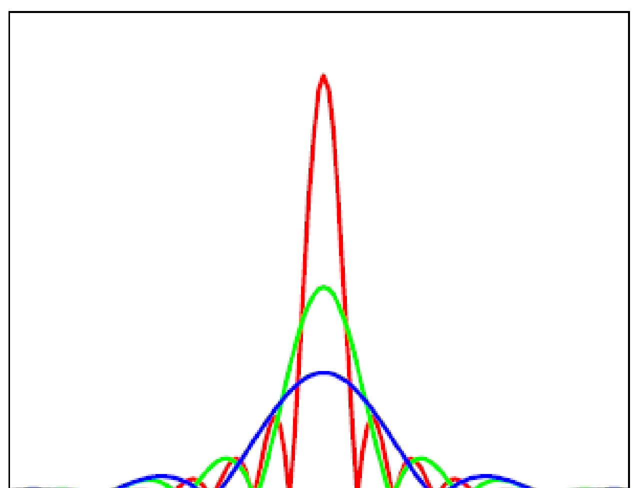


*Traditional 5-blade diaphragm and the family $\{g_d\}_{d\in D}$ of kernels.*

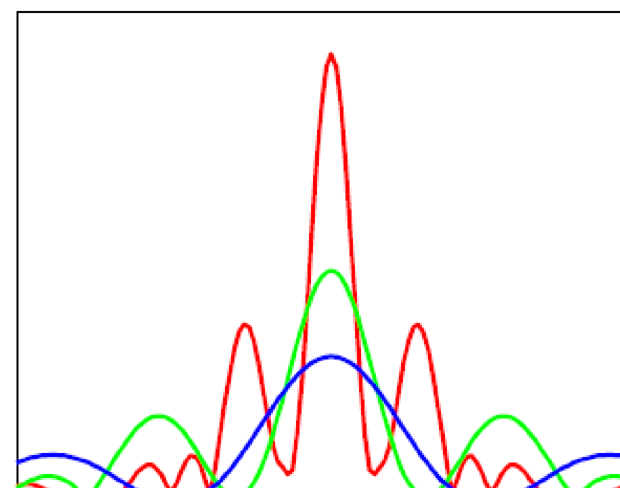


*Coded aperture and the family $\{g_d\}_{d\in D}$ of kernels.*

Comparing the kernels in frequency domain $\{\widetilde{g_d}\}_{d\in D}$ between classic and coded apertures (note the location of the zeros):



Conventional aperture



Coded aperture

[Levin 2007]

# PASSIVE 3D: CODED APERTURE

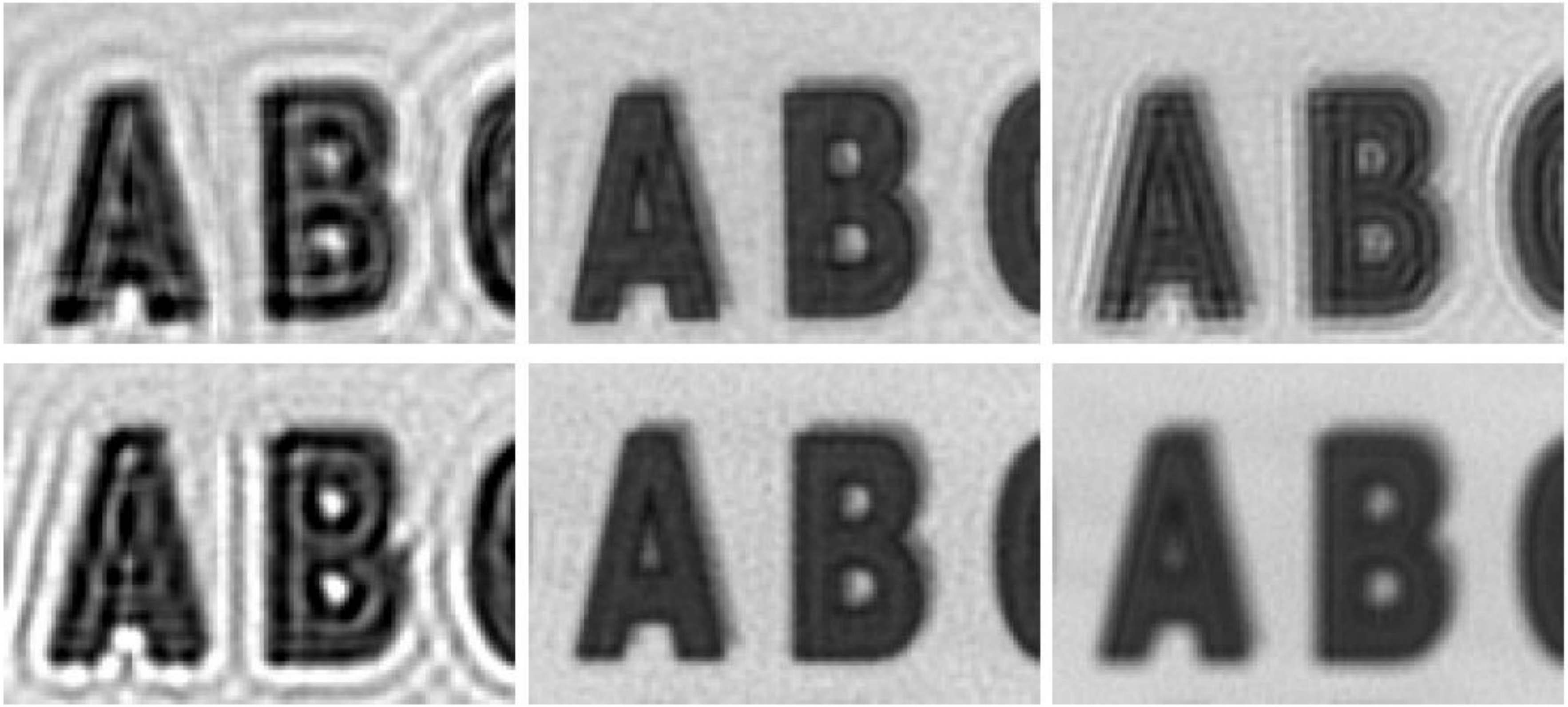
Images obtained by deconvolution with coded aperture allow to better discriminate the right scales (distances):

[Levin 2007]

$d > d_{opt}$ $\quad$ $d \simeq d_{opt}$ $\quad$ $d < d_{opt}$

***Coded aperture***

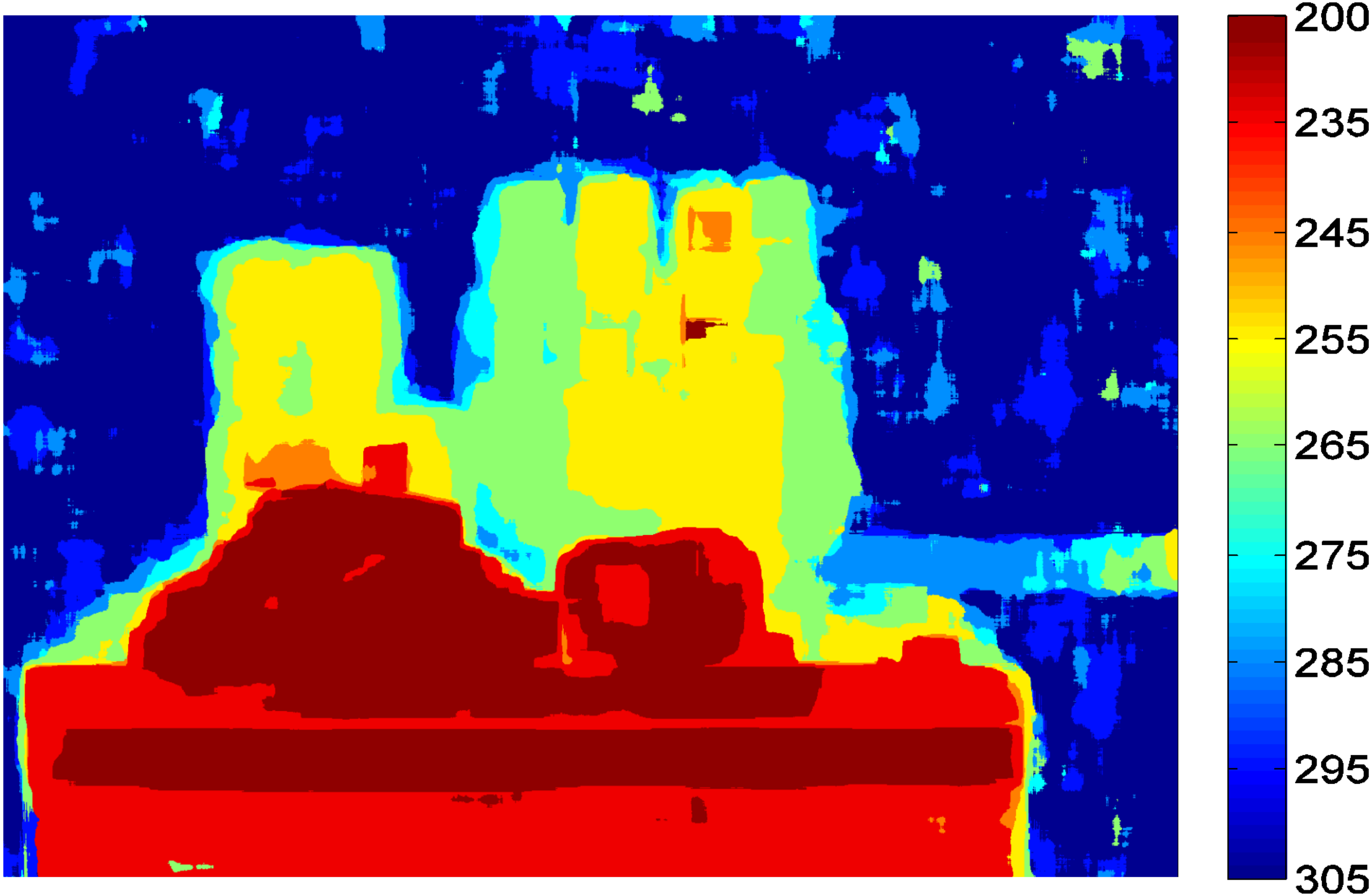
***Classic aperture***



# RANGE TEST IMAGE


[Levin 2007]

# RANGE IMAGE: CODED APERTURE RAW RESULT




[Levin 2007]

# RANGE IMAGE: RESULT AFTER POST-PROCESSING


*[Levin 2007]*

# CO-DESIGN FOR 3D: CONCLUSIONS

Co-design techniques aim at globally optimising a vision system by an opportunistic approach which makes the most of the different components and tries to combine them in a more intimate way: optics, mechanics, electronics, digital processing…

This lecture focused on 3d perception, but camera co-design is also much investigated for improving or “augmenting” digital images (“computational photography”).

An important feature of co-designed system is the balance between, on the one hand, hardware complexity and intrusive nature (lighting) of the system, and on the other hand, the software complexity. However, the weight of the software remains strong in most of the presented systems.

Another important point about passive systems, is the difficulty (if not impossibility) to process regions without structures (homogeneous).

The principles of the presented techniques are generally old, but the corresponding technology maturation is very recent, with several off-the-shelf products available now.

Finally, many depth visual cues remain unexploited today: there is still plenty of research and development works needed to contribute to the on-going expansion of co-design.

# REFERENCES (Part 1)

*[Giese 2006]* Martin A. Giese, *Visual perception*, lecture material, University of Tübingen, 2006.

*[Tautz 2008]* Jürgen Tautz, *The Buzz about Bees: Biology of a Superorganism*, Springer, 2008.

*[Krapp et Wicklein 2008]* H.G. Krapp, M. Wicklein *Central processing of visual information in insects*. In: The Senses: A Comprehensive Reference, ed. by A. Basbaum, A. Kaneko, G.M. Shepherd, G. Westheimer, Academic Press, 2008, p. 131–204.

*[Hoffman 2008]* D.M. Hoffman, A.R. Girshick, Kurt Akeley, M.S. Banks, *Vergence–accommodation conflicts hinder visual performance and cause visual fatigue*, Journal of Vision, 2008 vol. 8 no. 3 article 33.

*[Steele 2014]* Kenneth M. Steele, *Psychology of Perception*, Lecture material, Appalachian State University, Fall 2014 http://www1.appstate.edu/~kms/

# REFERENCES (Part 2)


***[Chiabrando 2009]*** F. Chiabrando, R. Chiabrando, D. Piatti, F. Rinaudo, *Sensors for 3D Imaging: Metric Evaluation and Calibration of a CCD/CMOS Time-of-Flight Camera*, Sensors, vol. *9*, 10080-10096, 2009.

***[Geng 2011]*** Jason Geng, *Structured-light 3D surface imaging: a tutorial*, Advances in Optics and Photonics, vol. *3*, 128-160, 2011.

***[Posdamer 1982]*** J. L. Posdamer and M. D. Altschuler, *Surface measurement by space-encoded projected beam systems*, Comput. Graph. Image Processing 18, (1), 1–17 1982.

***[Narasimhan 2006]*** S. Narasimhan, *Computer Vision: Spring 2006, lecture n.17*, Carnegie Mellon University.

***[Zhang 2002]*** L. Zhang, B. Curless, S. M. Seitz, *Rapid shape acquisition using color structured light and multi-pass dynamic programming*, IEEE Int. Symp. on 3D Data Processing Visualization and Transmission, pp. 24–36, 2002.

# REFERENCES (Part 3)


*[Adelson 1992]* E. H. Adelson, J. Y. A. Wang, *Single Lens Stereo with a Plenoptic Camera*, IEEE Trans. Pattern Analysis and Machine Intelligence 14(2): 99-106, 1992.

*[Ng 2005]* R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. *Light Field Photography with a Hand-Held Plenoptic Camera,* Stanford University Computer Science Tech Report CSTR 2005-02, April 2005.

*[Pentland 1987]* Alex P. Pentland, *A new sense for depth of field*, IEEE Trans. Pattern Analysis and Machine Intelligence 9(4): 523-531, 1987.

*[Maître 2003]* Henri Maître (ss la direction de), *Le Traitement des Images,* Chapitre 5 : Restauration, Hermès – Lavoisier, Série I2C, 2003.

*[Levin 2007]* A. Levin, R. Fergus, F. Durand, W.T. Freeman, *Image and depth from a conventional camera with a coded aperture,* ACM Transactions on Graphics 26 (3): 70-78, 2007.