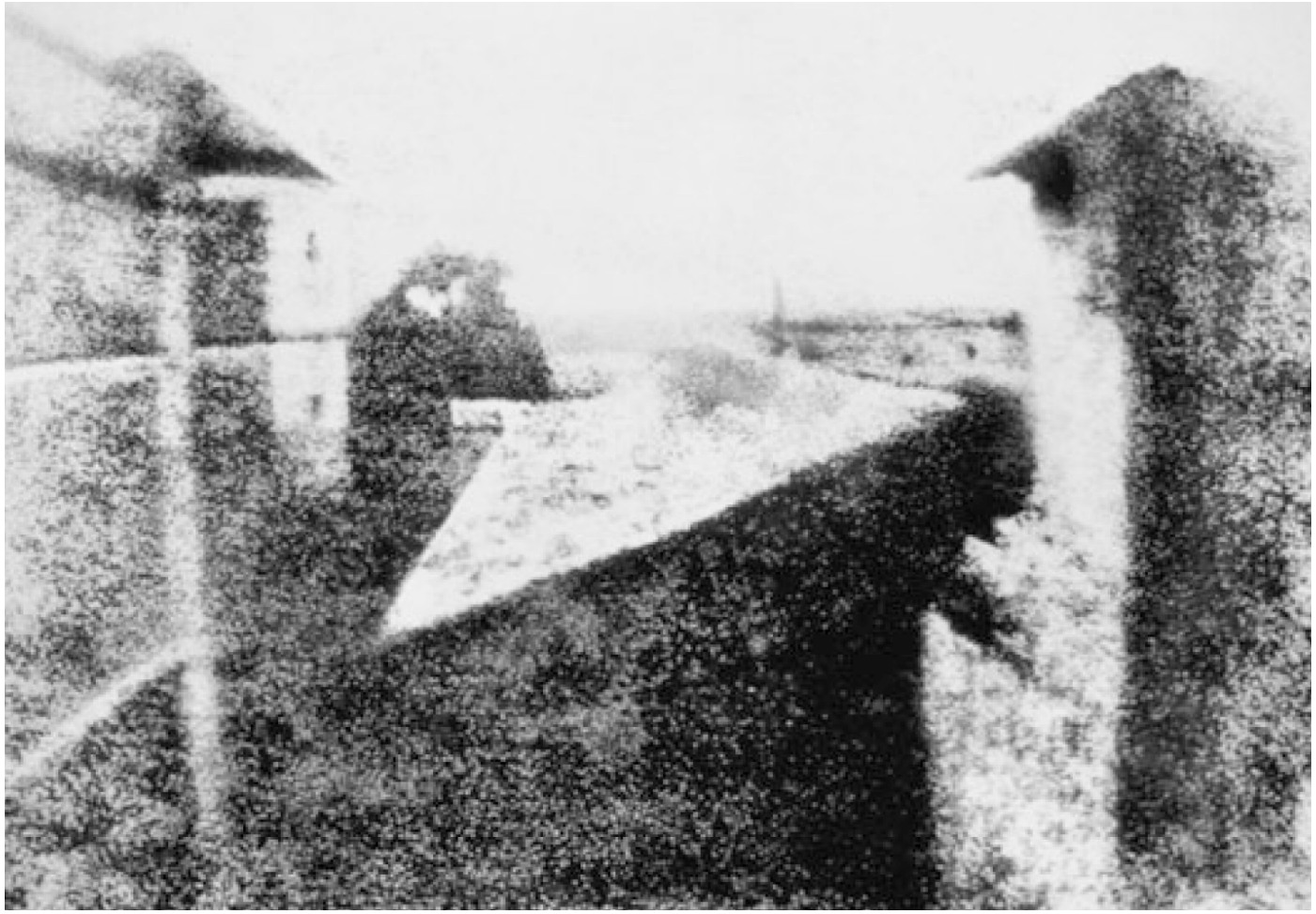
Stereo Vision and 3D Reconstruction

Wei Qi Yan

AUT, New Zealand

Table of Contents

The First Photograph (1826, France)



Stereo Camera

Terms

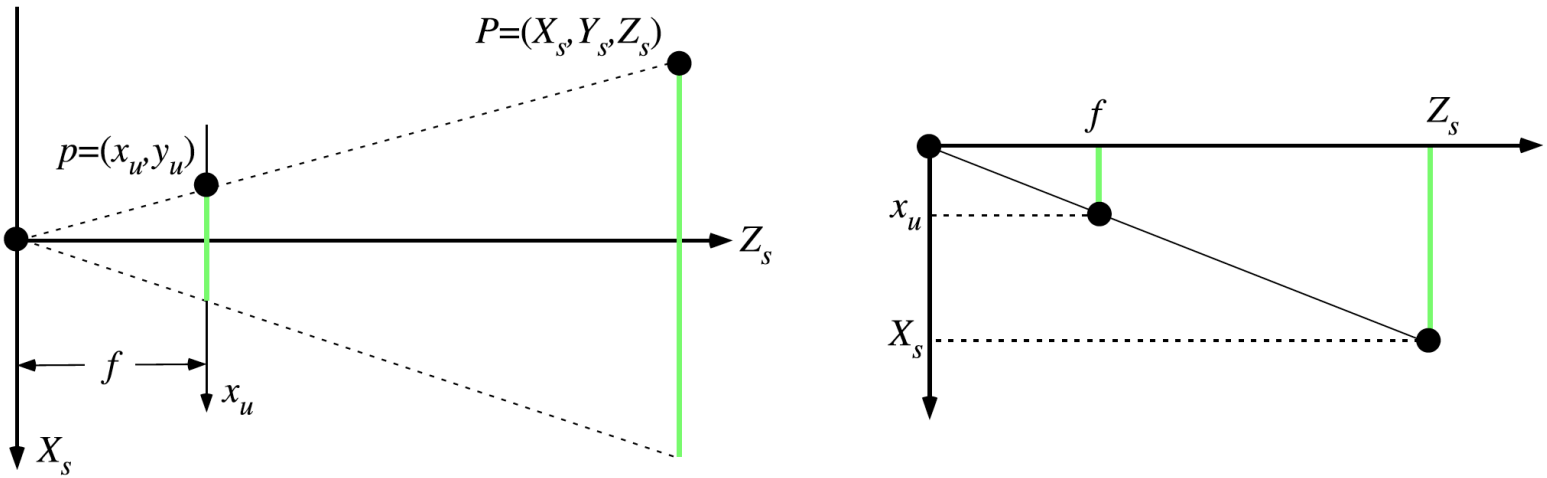
- Sony Mavica in 1981 (The first digital one)
- CCD: Charge-coupled device
- CMOS: Complementary metal-oxide semiconductor
- Color accuracy
- Lens distortion
- Aspect ratio
- Image resolution
- Bit depth
- Dynamic range
- Panning, tilting or zooming

R. Klette (2014) Concise Computer Vision, Springer

Central Projection

$$x_u = \frac{f \cdot X_s}{Z_s}, y_u = \frac{f \cdot Y_s}{Z_s}$$

where $P = (X_s, Y_s, Z_s)$ is a visible point in the world, $p = (x_u, y_u)$ is a pixel location, f is the focal length.

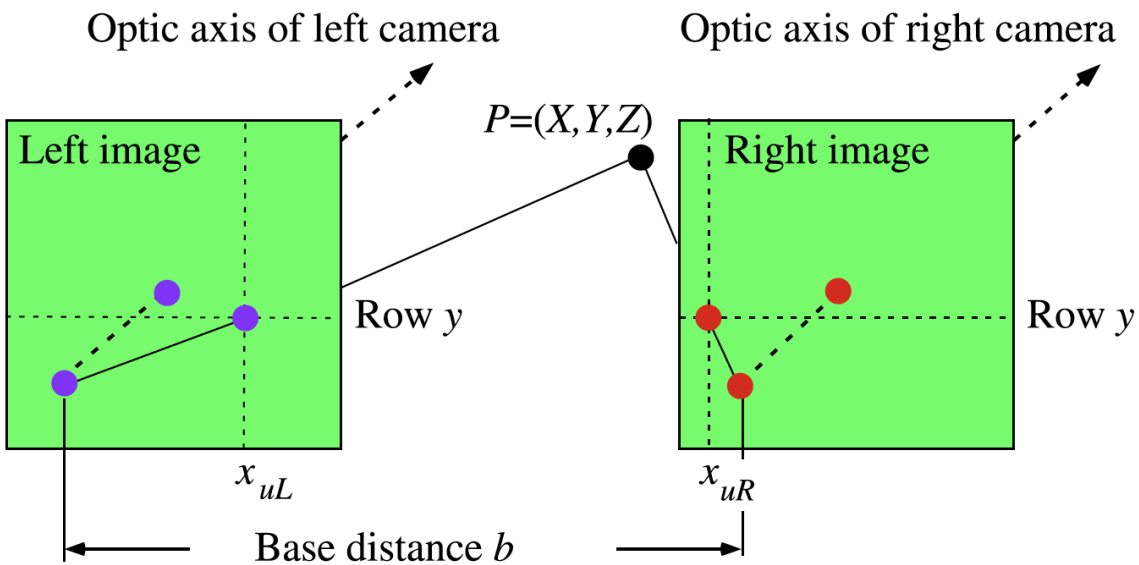


R. Klette (2014) Concise Computer Vision, Springer

Stereo Camera

Central Projection

- The two coplanar images have the identical size.
- Parallel optic axes
- An identical focal length
- Collinear image rows



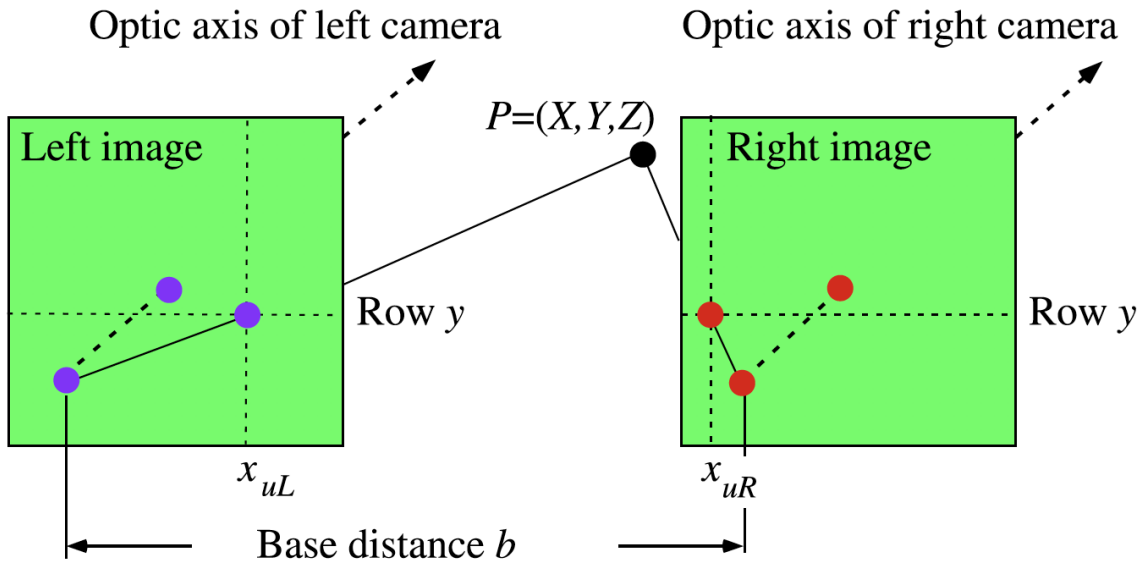
R. Klette (2014) Concise Computer Vision, Springer

Central Projection

$$p_{uL} = (x_{uL}, y_{uL}) = (\frac{f \cdot X_s}{Z_s}, \frac{f \cdot Y_s}{Z_s})$$

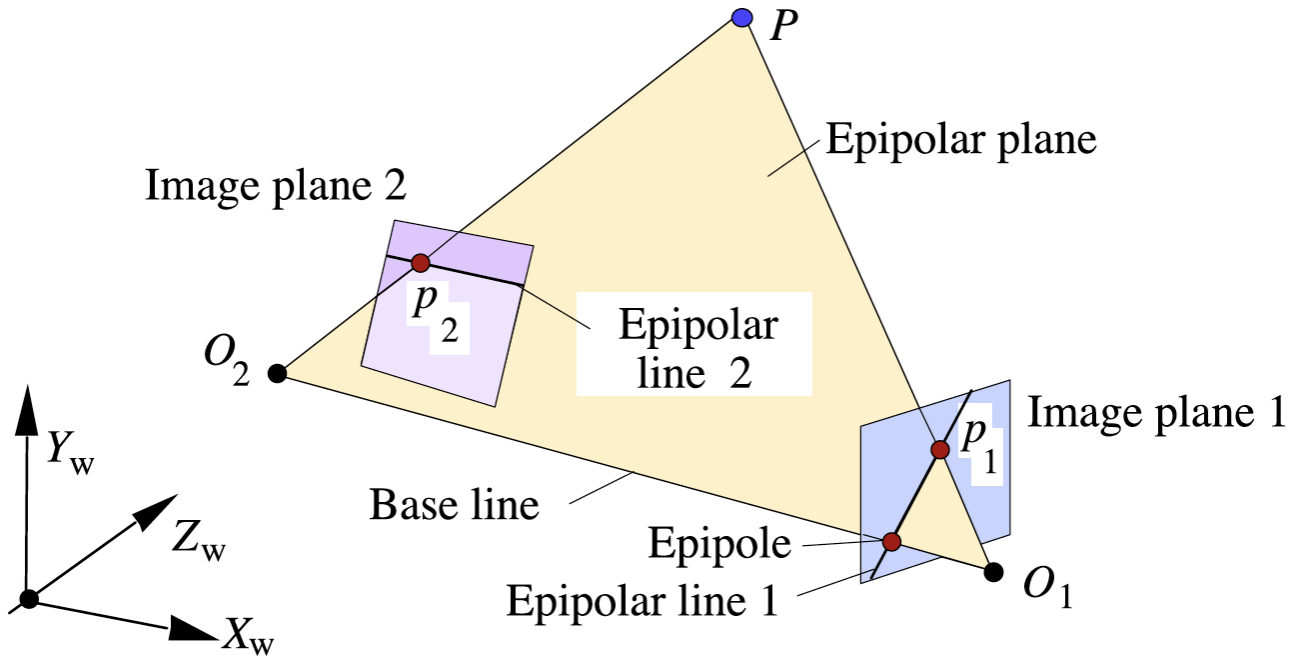
$$p_{uR} = (x_{uR}, y_{uR}) = (\frac{f \cdot (X_s - b)}{Z_s}, \frac{f \cdot Y_s}{Z_s})$$

where b is the base distance of the stereo system, p_{uL} for the left camera, p_{uR} for the right, $P = (X_s, Y_s, Z_s)$ is a visible point in the world, $p = (x_u, y_u)$ is a pixel location, f is the focal length.



Epipolar Geometry

Epipolar geometry shows the two cameras with associated coordinate frames and image planes.



Epipolar Geometry

- Epipolar geometry could represent the case of two cameras simultaneously by viewing the same scene.
- In epipolar plane, a world point is projected onto the image planes of the two cameras at two pixel coordinates respectively, known as conjugate points.
- Given a point in one image, its conjugate is constrained to lie along a line in the other image.

Questions?

Questions?

Regarding the stereo camera,

1. The two coplanar images have the identical size.
2. The stereo camera has two parallel optic axes.
3. The stereo camera has an identical focal length.
4. None of the given options.

Which answer is wrong:___.

Questions?

Stereo Vision

- Stereo vision is for estimating the 3D structure from two images from different viewpoints with approaches: Sparse and dense stereo.
- Sparse stereo is a natural extension of what we have learned about feature matching and recovers the world coordinate for each corresponding point pair.
- Dense stereo recovers the world coordinate for every pixel in the image.

Dense Stereo Matching

- A stereo pair is taken by two cameras, generally with parallel optical axes, and separated by using a known distance referred to as the camera baseline.
- For the parallel-axis camera geometry, the epipolar lines are parallel and horizontal, so conjugate points have the same vertical coordinate.
- The displacement along the horizontal epipolar line is called disparity.
- The epipolar constraint means that we only need to perform a 1D search for the corresponding point.
- The design of a stereo-vision system has three degrees of freedom: Baseline distance, disparity search range, and template size.

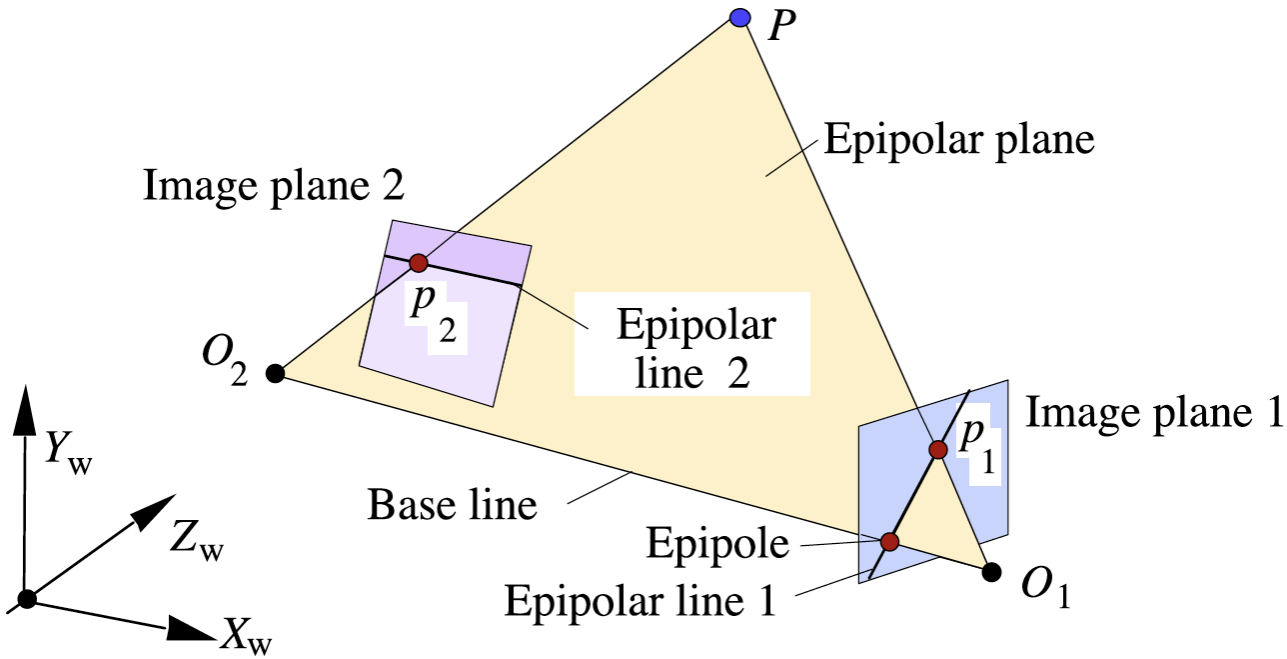
Anaglyphs

- Human stereo perception of depth works because each eye views the scene from a different viewpoint.
- The key in all 3D display to take the image from two cameras, with a similar baseline to the human eyes and present those images again to the corresponding eyes.
- The big advantage of anaglyphs is that they can be printed on paper or imaged onto ordinary movie film and viewed with simple and cheap glasses.
- Stereo cameras are built with precision to ensure that the optical axes of the cameras are parallel.

Structure and Motion

- The structure of scene is the 3D positions of points in the world.
- In a robotic scenario, the robot moves on a plane and a particular feature point lies on the ground or the top of a doorway.
- The magnitude of camera translational motion at each time step is treated as estimated from the essential matrix and the ground truth.

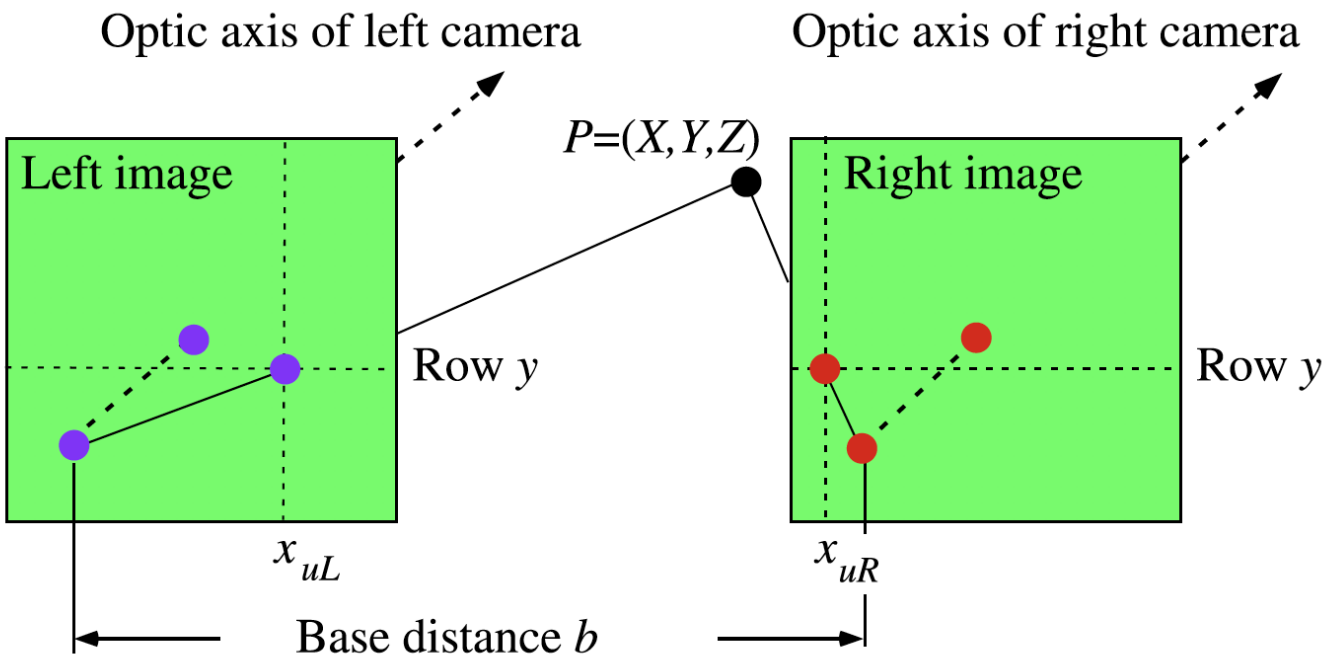
Geometry for Two Cameras



R. Klette (2014) Concise Computer Vision, Springer

Symmetric Cameras

In the camera coordinate system, we recover the unknown visible point $P = (X_s, Y_s, Z_s)$ using the undistorted image coordinates (x_{uL}, y_{uL}) and (x_{uR}, y_{uR}) as input, where we assume $y_{uL} = y_{uR} = y_u$ and $x_{uR} \leq x_{uL}$, the base line distance is $b > 0$, f is the unified focal length.



Symmetric Cameras

In the camera coordinate system, we recover the unknown visible point $P = (X_s, Y_s, Z_s)$ using the undistorted image coordinates (x_{uL}, y_{uL}) and (x_{uR}, y_{uR}) as input, where we assume $y_{uL} = y_{uR} = y_u$ and $x_{uR} \leq x_{uL}$, the base line distance is $b > 0$, f is the unified focal length.

$$\because Z_s = \frac{f \cdot X_s}{x_{uL}} = \frac{f \cdot (X_s - b)}{x_{uR}} \therefore X_s = \frac{b \cdot x_{uL}}{x_{uL} - x_{uR}}; Z_s = \frac{b \cdot f}{x_{uL} - x_{uR}}$$

$$P = (X_s, Y_s, Z_s) = (\frac{b \cdot x_{uL}}{x_{uL} - x_{uR}}, \frac{b \cdot y_u}{x_{uL} - x_{uR}}, \frac{b \cdot f}{x_{uL} - x_{uR}})$$

where $d = x_{uL} - x_{uR}$ is the disparity.

R. Klette (2014) Concise Computer Vision, Springer

Symmetric Cameras

In the camera coordinate system, we recover the unknown visible point:

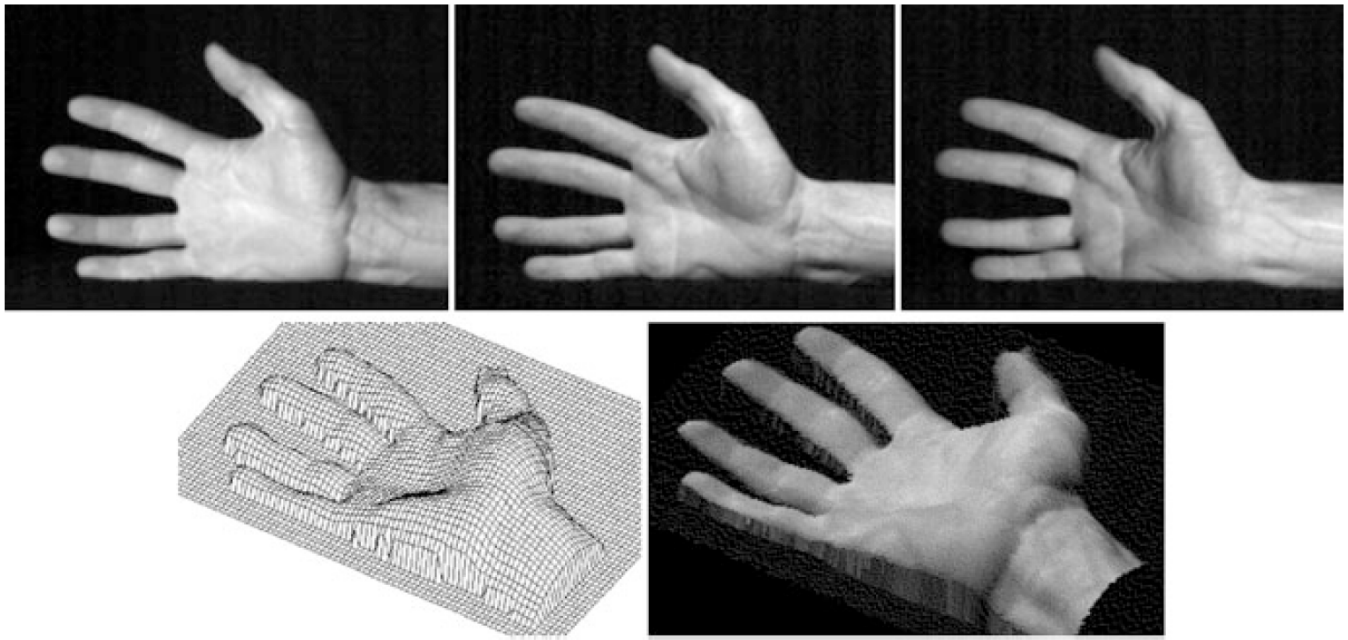
$$P = (X_s, Y_s, Z_s) = (\frac{b \cdot x_{uL}}{x_{uL} - x_{uR}}, \frac{b \cdot y_u}{x_{uL} - x_{uR}}, \frac{b \cdot f}{x_{uL} - x_{uR}})$$

Therefore:

- $d = x_{uL} - x_{uR} = 0$ means $P = (X, Y, Z)$ is at infinity(∞).
- Larger b and f support an increase in depth level but reduce the number of pixels that have corresponding pixels in the second image.
- An increase in image resolution is a way to improve the accuracy of depth levels.

R. Klette (2014) Concise Computer Vision, Springer

Stereo Vision



R. Klette (2014) Concise Computer Vision, Springer

Stereo Matching

We assume that stereo pairs are already geometrically rectified and pre-processed for reducing brightness issues. Corresponding pixels are expected to be in the left and right images in the same image row.

R. Klette (2014) Concise Computer Vision, Springer

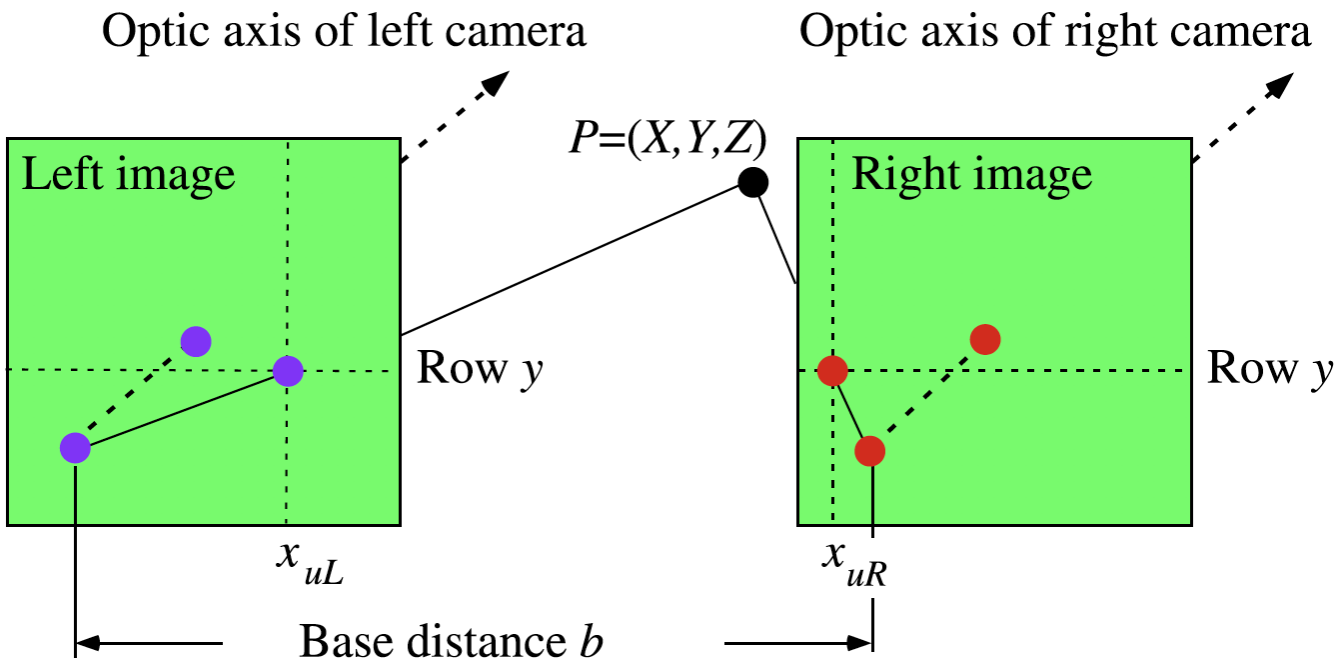
Disparity Matching

For a pixel $(x, y, B(x, y))$ in a **base image** B, we search for a corresponding pixel $(x + d, y, M(x + d, y))$ in the **match image** M, being on the same epipolar line identified by row y. The two pixels are corresponding if they are projections of the same point $P = (X, Y, Z)$, d is the disparity.

R. Klette (2014) Concise Computer Vision, Springer

Disparity Matching

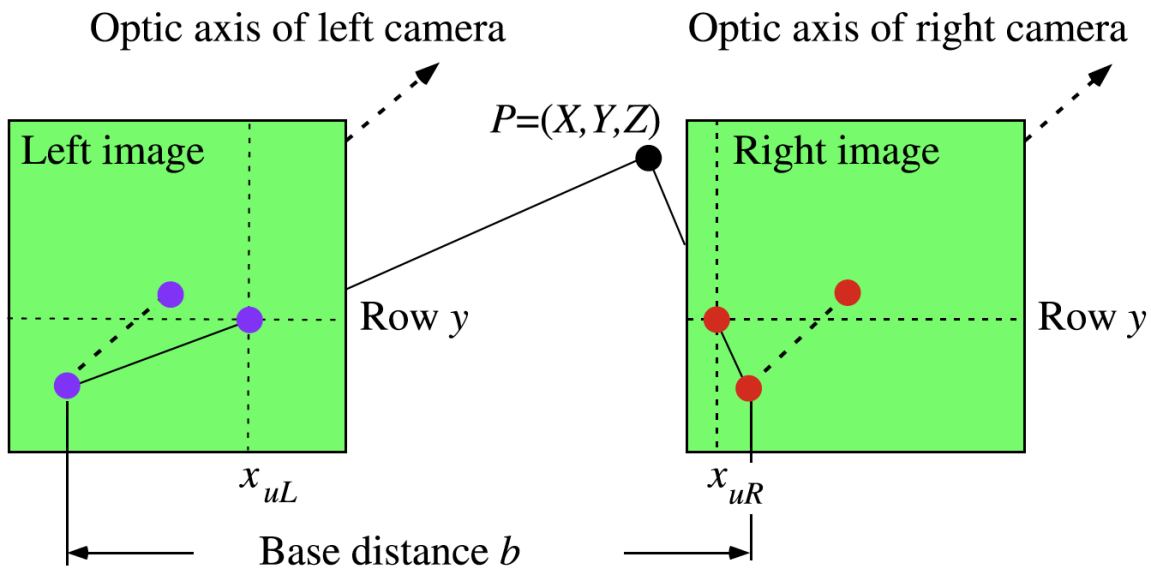
We initiate a search by selecting $p = (x, y)$ in B. This defines the search interval of points $q = (x + d, y)$ in M with $\max\{x - d_{max}, 1\} \leq x + d$.



R. Klette (2014) Concise Computer Vision, Springer

Disparity Matching

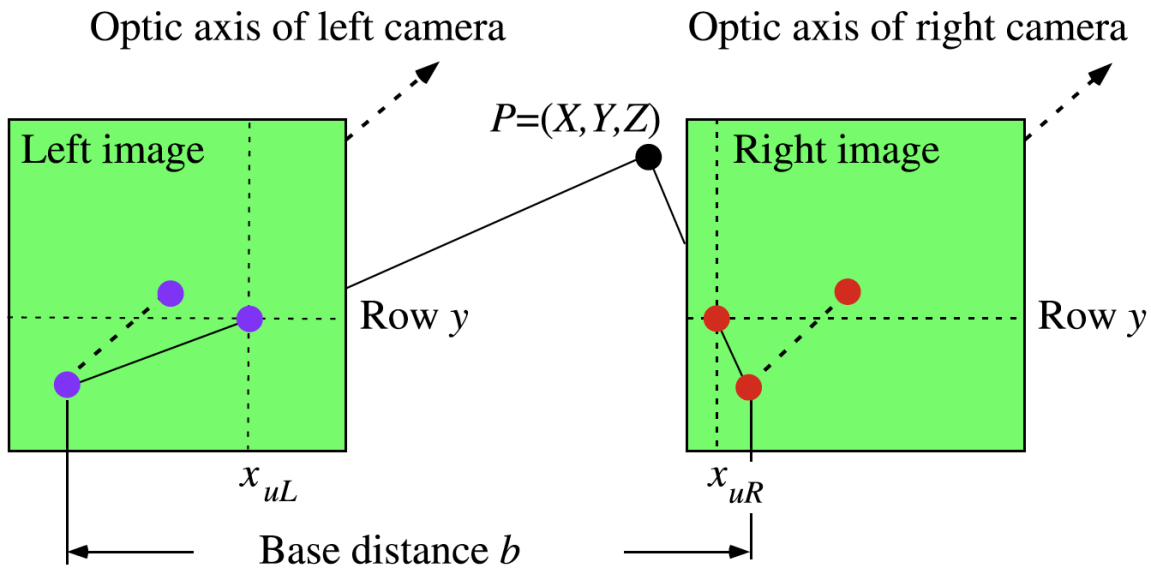
For identifying corresponding points, a straightforward idea is to compare neighbourhoods (rectangular windows for simplicity) around a pixel p in the image I.



R. Klette (2014) Concise Computer Vision, Springer

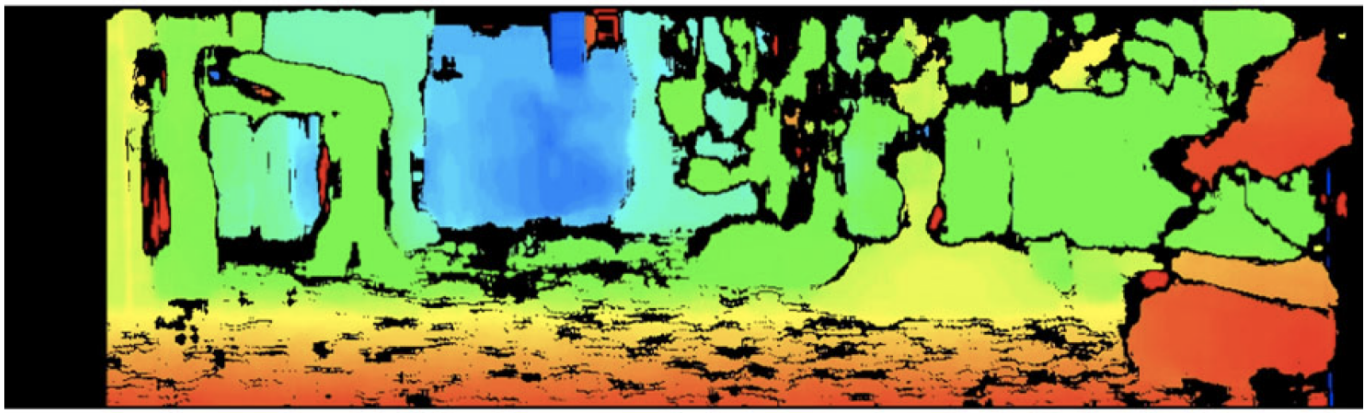
Disparity Matching

- Global Matching (GM): An area is approximated by less time-expensive control structure of a stereo matcher.
- Local Matching (LM): An area of influence is bounded by using fixed constant.
- Semi-Global Matching: We take more pixels into account than the local approach, but not yet as much as for a global approach.



Disparity Matching

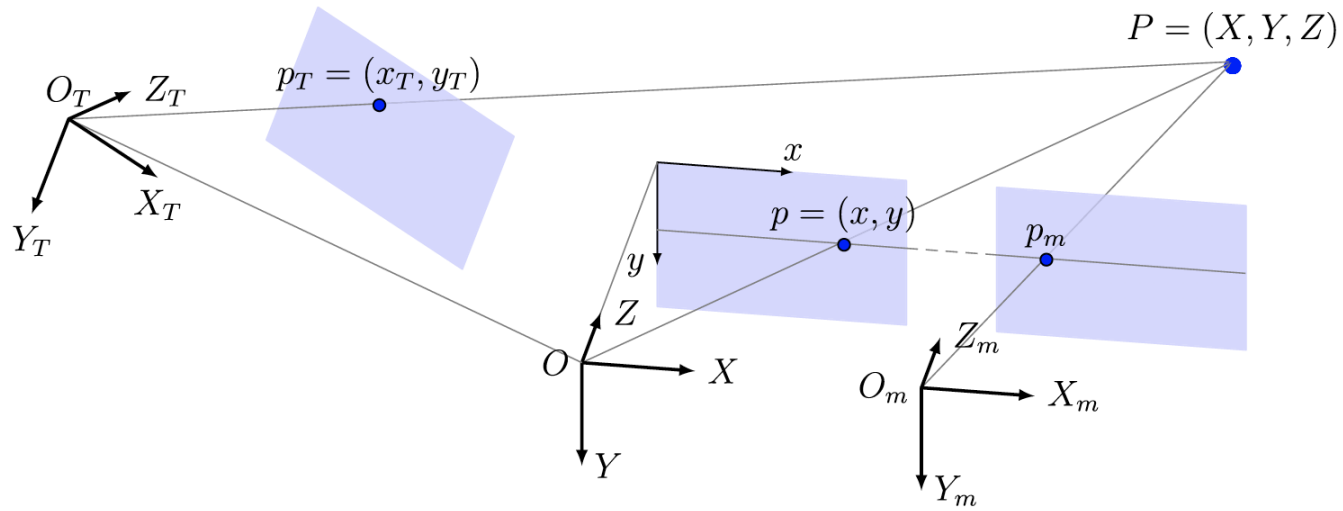
- Global Matching (GM): An area is approximated by less time-expensive control structure of a stereo matcher.
- Local Matching (LM): An area of influence is bounded by using fixed constant.
- Semi-Global Matching: We take more pixels into account than the local approach, but not yet as much as for a global approach.



R. Klette (2014) Concise Computer Vision, Springer

Third-Eye Technique

- Map a reference image of a pair of stereo camera into the pose of a third camera.
- Measure the similarity between created virtual image and the actually recorded third image.



Third-Eye Technique

Outline of the third-eye technology:

- Record stereo data with two cameras and calculate disparities.
- Have a third calibrated camera looking into the same space as the other two cameras.
- Use the calculated disparities for mapping the recorded image of the left camera into the image plane of the third camera, create a virtual image.
- Compare the virtual image with the image recorded by the third camera
- If the virtual and third images "basically coincide", then the stereo matcher provides "useful" disparities.

R. Klette (2014) Concise Computer Vision, Springer

Third-Eye Technique

A point $P = (X, Y, Z)$ mapped into a pixel (x, y) in the left image is mapped into a point (x_T, y_T) in the third image, and (x_T, y_T) is expressed in terms of (x, y) by using the calibrated translation (t_X, t_Y, t_Z). The base distance b, focal length f_T and the disparity d provided by the given stereo matcher:

$\because (X_T, Y_T, Z_T) = (X - t_X, Y - t_Y, Z - t_Z)$ and

$(x_T, y_T) = f_T \cdot (\frac{X_T}{Z_T}, \frac{Y_T}{Z_T})$,

$\therefore (x_T, y_T) = f_T \cdot (\frac{X - t_X}{Z - t_Z}, \frac{Y - t_Y}{Z - t_Z})$

If $X = \frac{b \cdot x}{d}$, $Y = \frac{b \cdot y}{d}$, $Z = f \cdot \frac{b}{d}$, then

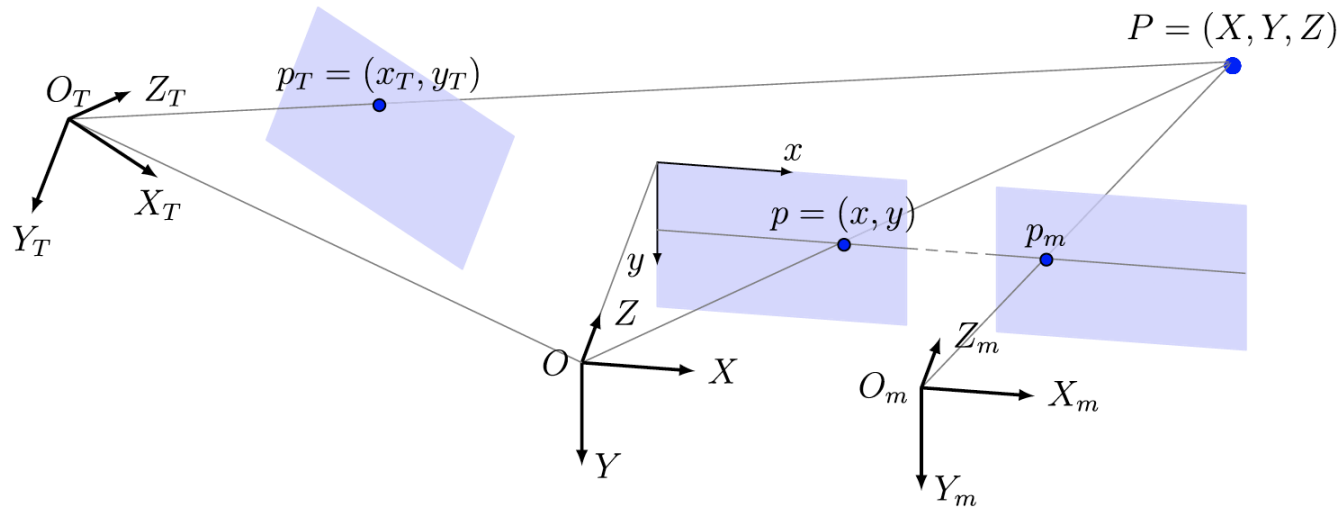
$x_T = f_T \cdot \frac{bx - dt_X}{fb - dt_Z}$ and $y_T = f_T \cdot \frac{by - dt_Y}{fb - dt_Z}$.

It maps the intensity value at pixel (x, y) in the reference image into the plane of the third image.

R. Klette (2014) Concise Computer Vision, Springer

Third-Eye Technique

- Map a reference image of a pair of stereo camera into the pose of a third camera.
- Measure the similarity between created virtual image and the actually recorded third image.



Similarity between Virtual Image and Third Image

Let Ω_t be the set of pixels that are used for the comparison for frames at time t. The means μ_V,μ_T and standard variations σ_V, σ_T for the virtual $V(p)$ and third image $T(p)$ at time t, respectively. The Normalized Cross-Correlation (NCC) is,

$$M_{NCC}(V,T) = \frac{1}{|\Omega_t|} \sum_{p \in \Omega_t} \frac{[T(p) - \mu_T][V(p) - \mu_V]}{\sigma_T \sigma_V}.$$

The NCC values is employed to compare the performance of stereo matches on very long sequences.

R. Klette (2014) Concise Computer Vision, Springer

Questions?

Questions?

Stereo cameras are built with precision to ensure that the optical axes of the cameras are,

1. parallel.
2. perpendicular.
3. with any directions.
4. none of the given options.

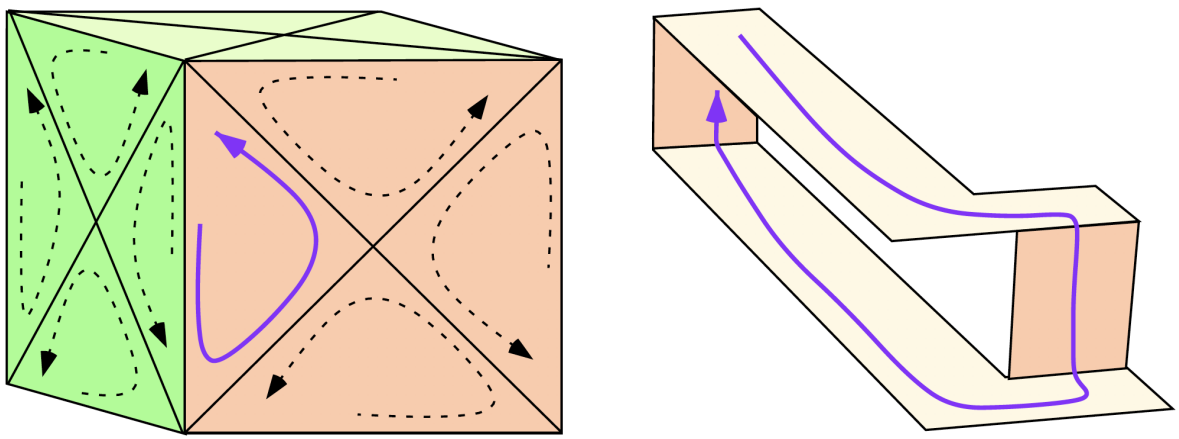
The right answer is:___.

Questions?

Surface

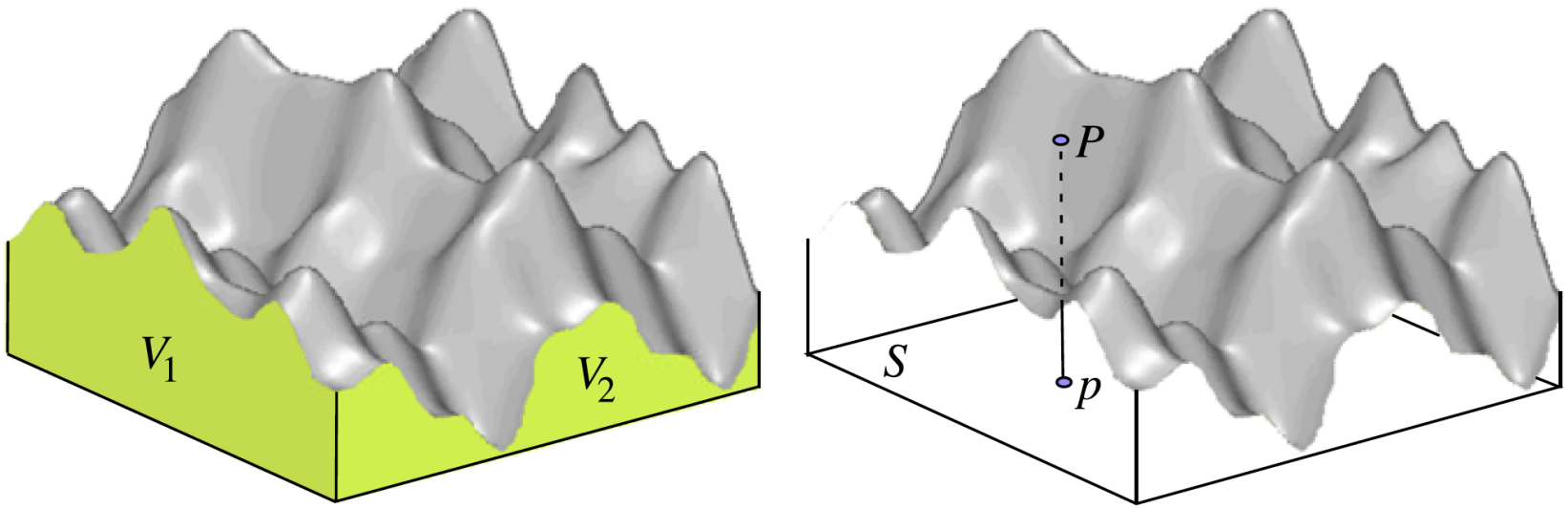
The surface S known as border or frontier of an existing 3D object in the real 3D world:

- Gap-free smooth surface:
  (1) Continuous derivatives exist
  (2) The existence of a neighbourhood in S
- Gap-free polyhedral surface:
  (1) Discontinuities at edges
  (2) The existence of a neighbourhood in S



Representation of a surface

- Explicit representation: $Z = F(X, Y)$
- Implicit representation: $F(X, Y, Z) = 0$



R. Klette (2014) Concise Computer Vision, Springer

Normal Vector

Gradient of a surface $Z = F_e(X, Y)$ is the vector given by:

$$\nabla Z = \mathbf{grad} Z = [\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}]^\top$$

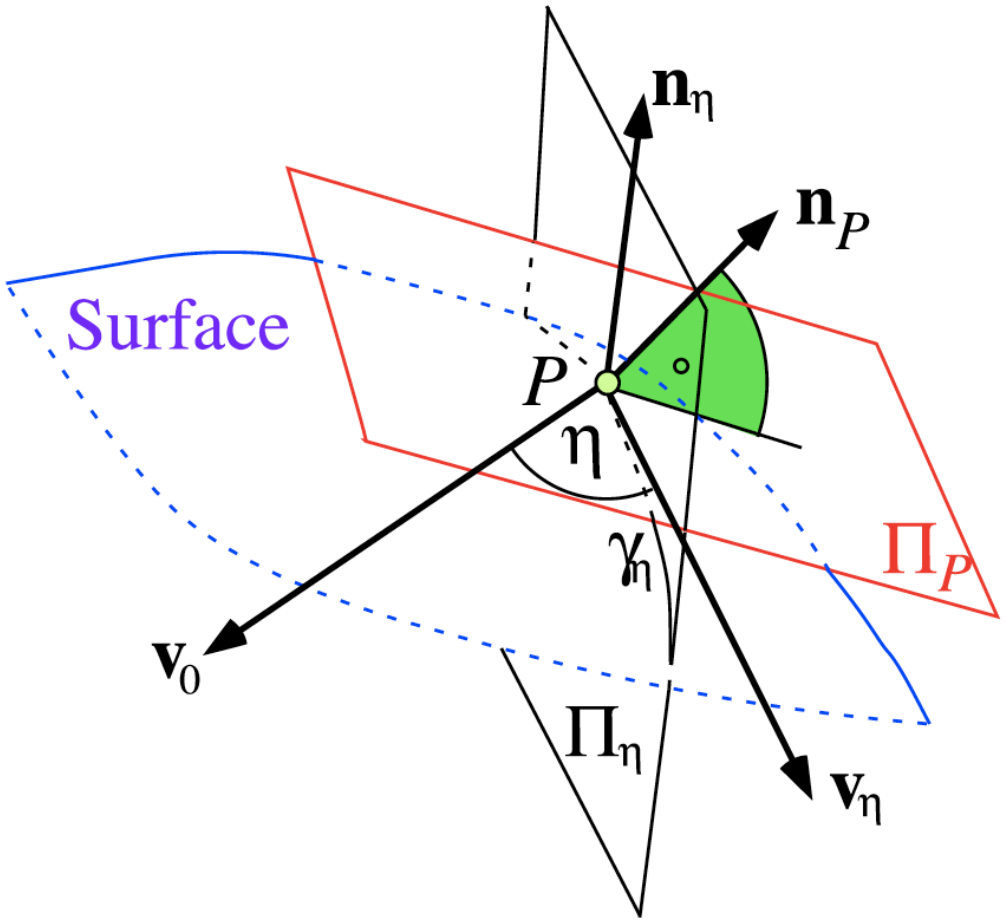
In the case of plane $aX + bY + Z = c$,

$$\mathbf{n} = [\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, 1]^\top = [a, b, 1]^\top$$

The unit normal vector is,

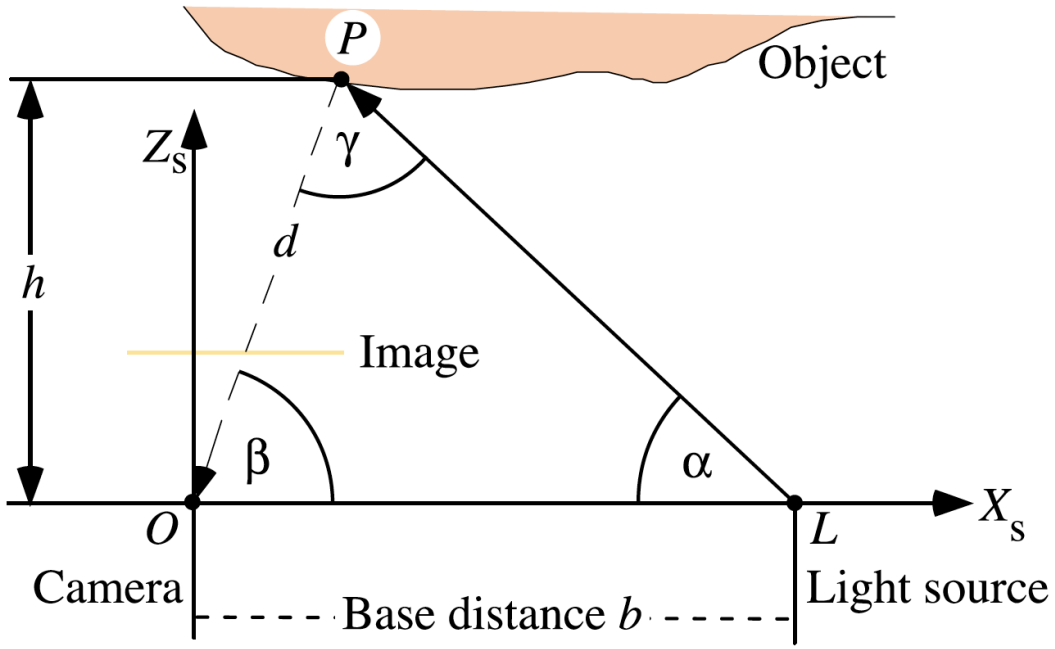
$$\mathbf{n}^\circ = [n_1, n_2, n_3]^\top = \frac{\mathbf{n}}{\|\mathbf{n}\|_2} = \frac{[a, b, 1]^\top}{\sqrt{a^2 + b^2 + 1}}$$

Tangent Plane



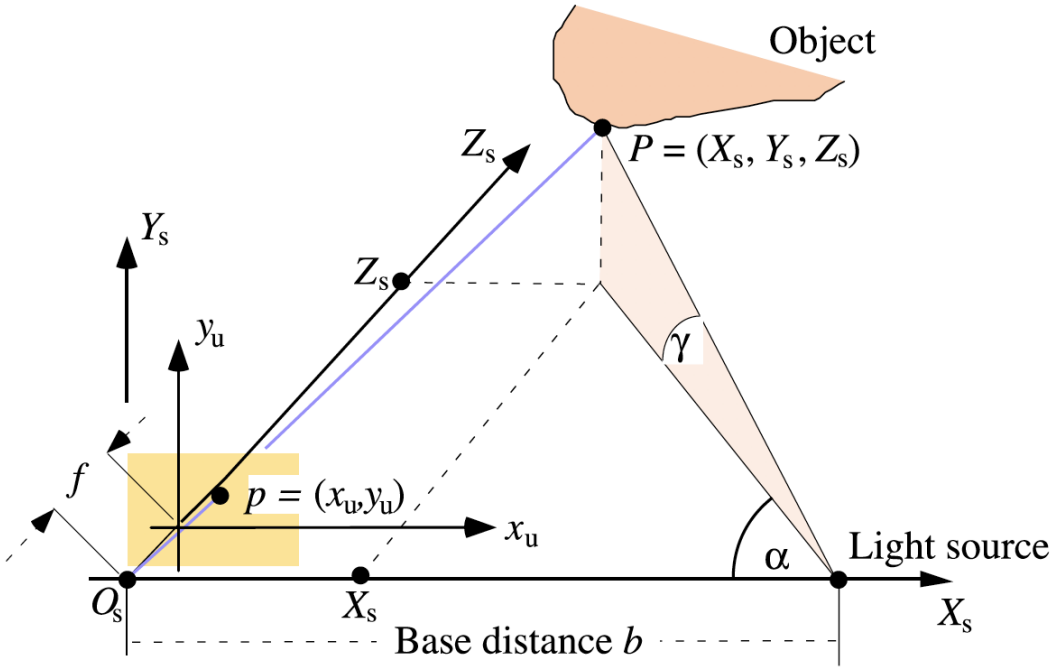
R. Klette (2014) Concise Computer Vision

Structured Lighting



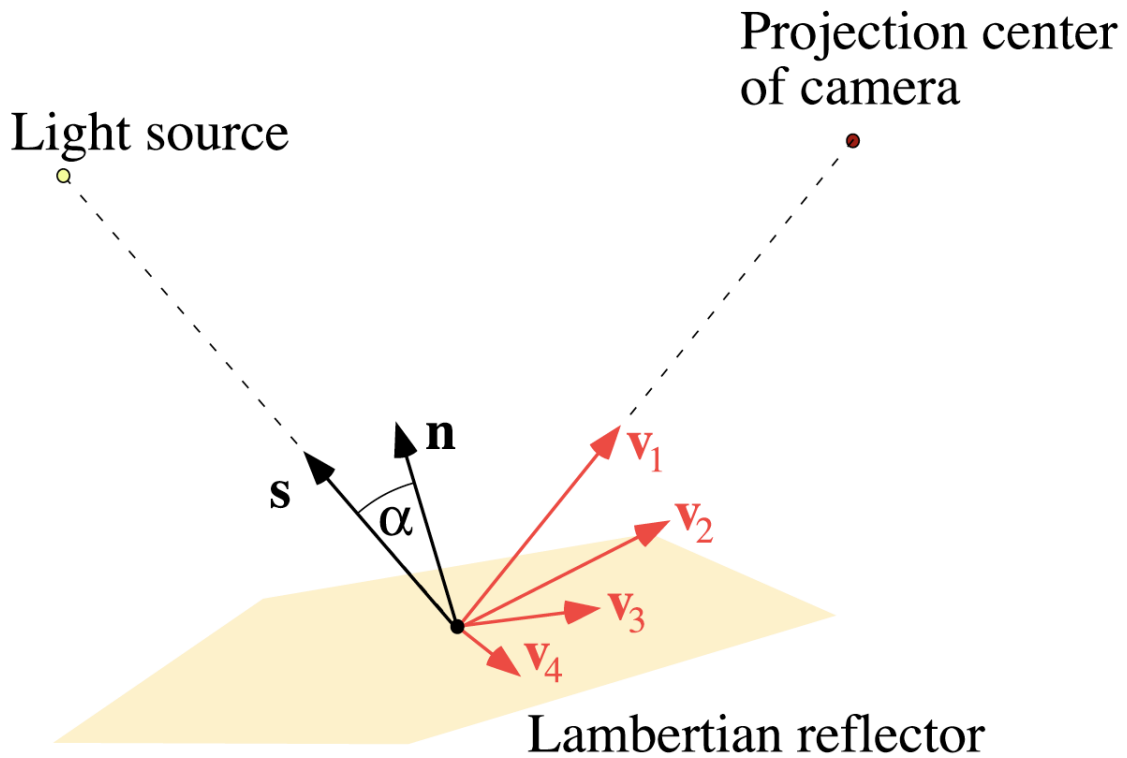
R. Klette (2014) Concise Computer Vision

Structured Lighting



R. Klette (2014) Concise Computer Vision, Springer

Lambert's Cosine Law



R. Klette (2014) Concise Computer Vision, Springer

Structured Lighting

Let $P = [a, b, 1]$ be the surface normal vector of a visible and illuminated surface at point P,

$$\cos\alpha = \frac{\mathbf{s}^\tau \mathbf{n}_p}{\|\mathbf{s}^\tau\|_2 \|\mathbf{n}_p\|_2}$$

The emitted light at the point P is scaled by

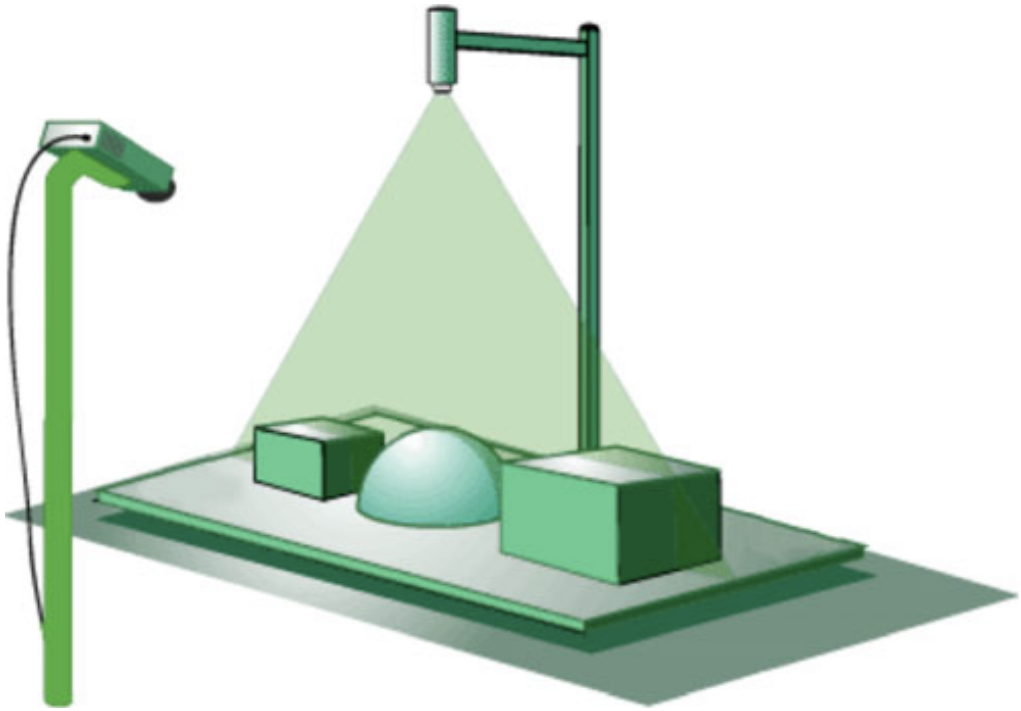
$$\eta(p) = \rho(p) \cdot \frac{E_l}{\pi}$$

where E_L was defined as a light source energy, which is reflected at P uniformly into all directions of a hemisphere.

$$R(p) = \eta(p) \frac{\mathbf{s}^\tau \mathbf{n}_p}{\|\mathbf{s}^\tau\|_2 \|\mathbf{n}_p\|_2}$$

where the reflectance $R(P) \geq 0$ is a second-order function.

Structured Lighting



R. Klette (2014) Concise Computer Vision, Springer

Questions?

Questions?

In 3D reconstruction, Lambert's Law is based on

1. cosine function
2. sine function
3. tanget function
4. none of the given options.

The right answer is:___.

Questions?

Learning Objectives

- Derive solutions for particular robotic vision and visual control tasks characterised by specifics of image data and deep learning algorithms.
- Critically evaluate the performance of robotic vision with deep learning algorithms, bench mark data, performance measures, and ways to define ground truth.