# Depth from Stereo Vision

ELECENG 3EY4: Electrical Systems Integration Project

Shahin Sirouspour

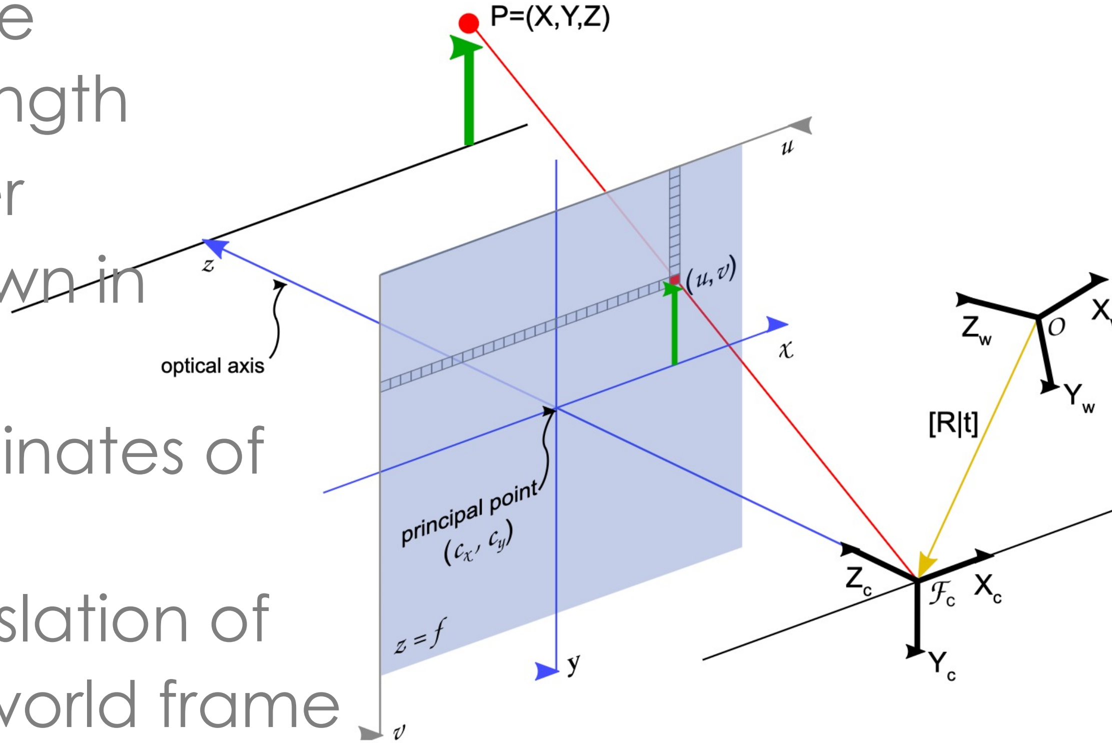
Winter 2024

# Pinhole Camera Model

- A point in 3D space is mapped to 2D image plane
- $X_cY_cZ_c$: Camera coordinate frame
- $X_wY_wZ_w$: World (fixed) coordinate frame
- $P$: Point in 3D space
- $f$: Camera focal length

$(c_x, c_y)$: Optical center

- Image plane is shown in blue
- $(u, v)$: Image coordinates of point in pixels
- $[R|t]$: Rotation/Translation of camera frame w.r.t. world frame




Pinhole camera model

(image source)

# Pinhole Camera Model

- $P_c = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$: 3D coordinates of point $P$ in camera frame
- $P_w = \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix}$: 3D coordinates of point $P$ in world frame
- $\begin{bmatrix} x_i \\ y_i \\ f \end{bmatrix}$: 3D coordinates of projection of $P$ in camera frame

$$\begin{bmatrix} x_i \\ y_i \\ \lambda \end{bmatrix} = k \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix}$$

# Pinhole Camera Model

- From third equation:

$$k = \frac{f}{z}$$

Therefore,

$$z\begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} fx \\ fy \\ z \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

- Note all the points on the ray originating from center of camera frame and passing through $P$ would have the same projection on image plane, hence creating a depth ambiguity

# Pinhole Camera Model

- The coordinates of the point in the camera frame are linked to those in the world frame through a homogenous transformation

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = [R_w^c \quad | \quad t_w^c] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}$$

$$R_w^c = R^T, \qquad t_w^c = -R^T t$$

- Coordinates of the projection of the point in the image plane are related to its pixel coordinates:

$$u - c_x = \frac{x_i}{s_x}, \qquad v - c_y = \frac{y_i}{s_y}$$

- $s_x, s_y$: pixel size along $x$ and $y$ directions.

# Pinhole Camera Model

- Combining the previous equations yields:

$$z\begin{bmatrix}u\\v\\1\end{bmatrix} = K\begin{bmatrix}x\\y\\z\end{bmatrix}$$

$$z\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}\frac{f}{s_x} & 0 & c_x\\0 & \frac{f}{s_y} & c_y\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}\frac{f}{s_x} & 0 & c_x\\0 & \frac{f}{s_y} & c_y\\0 & 0 & 1\end{bmatrix}[R_w^c \quad | \quad t_w^c]\begin{bmatrix}x_w\\y_w\\z_w\\1\end{bmatrix}$$

$$z\begin{bmatrix}u\\v\\1\end{bmatrix} = KT_{\mathrm{w}}^{\mathrm{c}}\begin{bmatrix}x_w\\y_w\\z_w\\1\end{bmatrix}$$

# Pinhole Camera Model

- Intrinsic camera parameters:

$$K \triangleq \begin{bmatrix} \frac{f}{s_x} & 0 & c_x \\ 0 & \frac{f}{s_y} & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

- Extrinsic camera parameters:

$$T_w^c \triangleq [R_w^c \quad | \quad t_w^c]$$

# Pinhole Camera Model

$$z\begin{bmatrix}u\\v\\1\end{bmatrix} = KT_{\mathrm{w}}^{\mathrm{c}}\begin{bmatrix}x_w\\y_w\\z_w\\1\end{bmatrix}$$

- This model relates the 3D coordinates of the point $P$ in the world frame to 2D coordinates of its projection in image plane in pixel unit
- The mapping from from 3D space to 2D space is not reversable!
  - We can determine $(u, v)$ from $(x_w, y_w, z_w)$ if we know the camera intrinsic and extrinsic parameters
  - However, there is no unique answer for $(x_w, y_w, z_w)$ given an image point $(u, v)$

# Determining Camera Position/Orientation

- Can we determine the homogenous transformation $T_w^c$ from imaging a few points in 3D space with known position coordinates?
- The coordinates of $k$'th corresponding pair in 2D-3D space are related by:

$$z^k \begin{bmatrix} u^k \\ v^k \\ 1 \end{bmatrix} = K T_{\mathrm{w}}^{\mathrm{c}} \begin{bmatrix} x_w^k \\ y_w^k \\ z_w^k \\ 1 \end{bmatrix}$$

- Intrinsic parameters $K$, image coordinates $\begin{bmatrix} u^k \\ v^k \end{bmatrix}$, and 3D position of point $\begin{bmatrix} x_w^k \\ y_w^k \\ z_w^k \end{bmatrix}$ are assumed to be *known*

# Determining Camera Position/Orientation

- The previous equation can also be written as:

$$z^k \begin{bmatrix} u^k \\ v^k \\ 1 \end{bmatrix} = K \left( R_w^c \begin{bmatrix} x_w^k \\ y_w^k \\ z_w^k \end{bmatrix} + t_w^c \right)$$

$$P_i^k = K\left(R_w^c P_w^k + t_w^c\right)$$

- An *optimization* problem can be formulated and solved to determine the unknown homogenous coordinate transformation

$$\min_{R_w^c \in S(O_3),\ t_w^c \in \mathbb{R}^3,\, z^k} \Sigma_{k=1}^{n} \left\| K\left(R_w^c P_w^k + t_w^c\right) - P_i^k \right\|^2$$

# Determining Camera Position/Orientation

- How many corresponding pairs of points do we need to solve this problem?

$$\min_{R_w^c \in S(O_3),\ t_w^c \in \mathbb{R}^3,\ z^k} \Sigma_{k=1}^{n} \left\| K\left(R_w^c P_w^k + t_w^c\right) - P_i^k \right\|^2$$

# Multi-view Vision

- Recall that the relationship between the coordinates of a point in camera frame to those in the image plane is given by:

$$z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

- It is impossible to to go from image coordinates $(u, v)$ to 3D camera frame coordinates $(x, y, z)$; this is depth ambiguity in single-view vision

# Multi-View Vision

- Looking at the same point from two different viewpoints

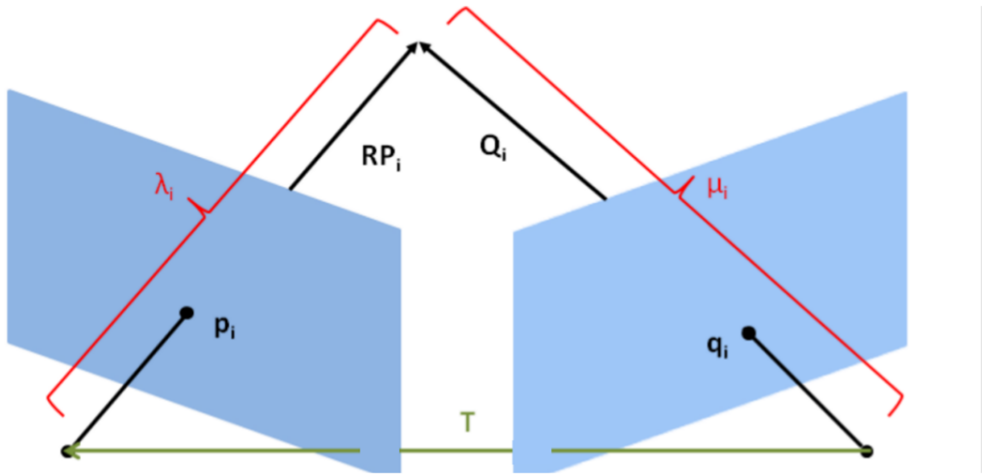



Image source

- *T*: translation of left camera frame with respect to right camera frame
- *R*: The rotation matrix of the left camera frame with respect to right camera frame

# Multi-View Vision

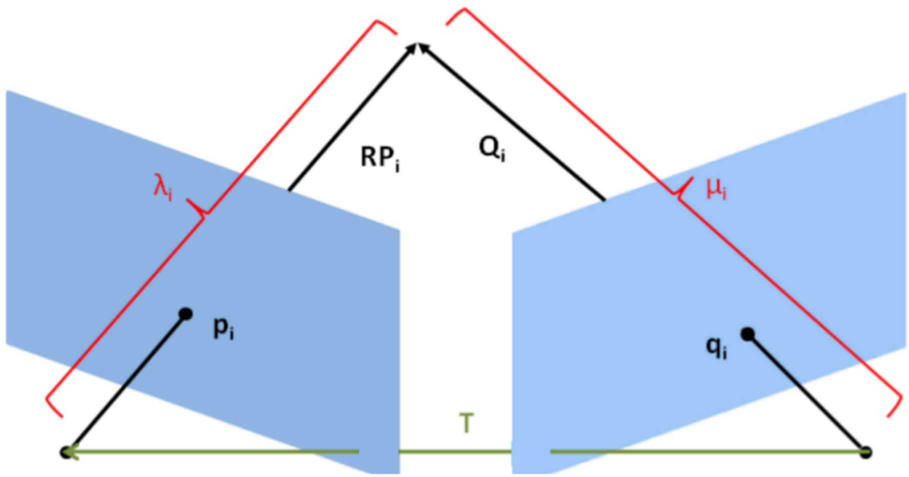
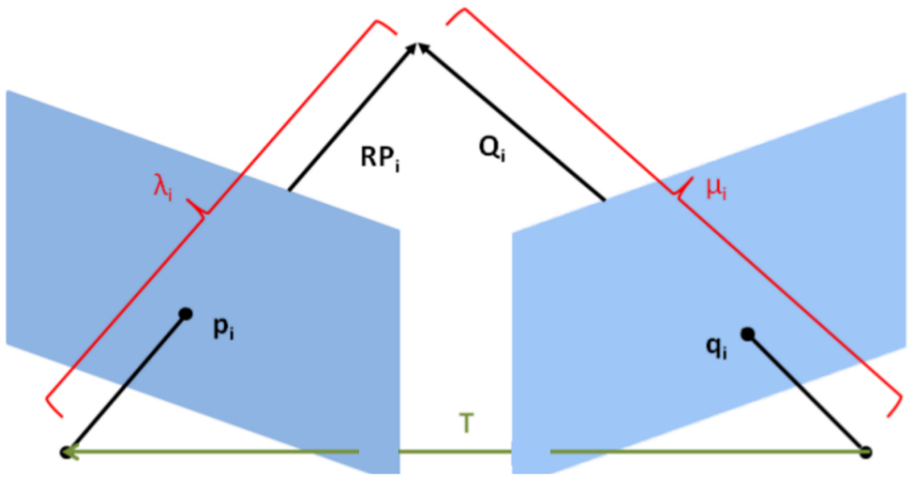



Image source

- $p, P$: Image and camera frame coordinates of point in left camera frame
- $q, Q$: Image and camera frame coordinates of point in right camera frame
- $\lambda, \mu$: unknown depths of the point in left and right coordinate frames

$$Q = \mu q,\ P = \lambda p$$

# Multi—View Vision




$$\mu q = R\lambda p + T$$

- Note that the following epipolar constraint holds:
$$q^T(T \times Rp) = 0$$
- Unknown depth information has been eliminated from the above equation

# Multi-View Vision

- Epipolar plane:

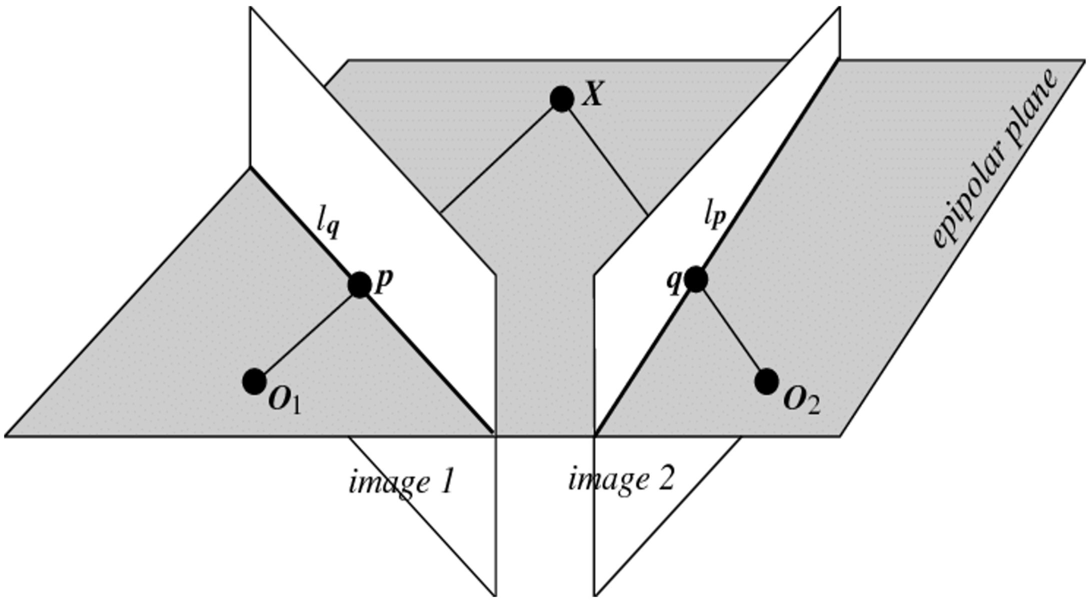



Image source

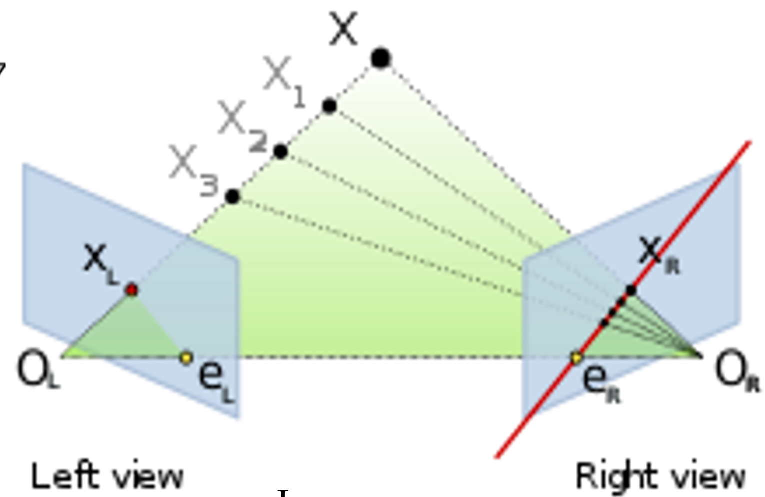



Image source

- $l_q$: a line in Image 1 where all potential corresponding points to point $q$ in Image 2 lie
- $l_p$: a line in Image 2 where all potential corresponding points to point $p$ in Image 2 lie

# Multi-view Vision

- Epipolar constraint:

$$q^T(T \times Rp) = 0$$
$$q^T(T \times R)p = 0$$

- $T \times R$ is a 3×3 matrix.

- Let $T = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$, then $T\times$ is a 3×3 skew-symmetric matrix,

$$T\times = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}$$
$$(T\times)^T = -T\times$$

# Stereo Vision

- Stereo vision is a special case of the more general multi-view vision where the two image planes are *parallel* and laterally displaced along the *x-axis*

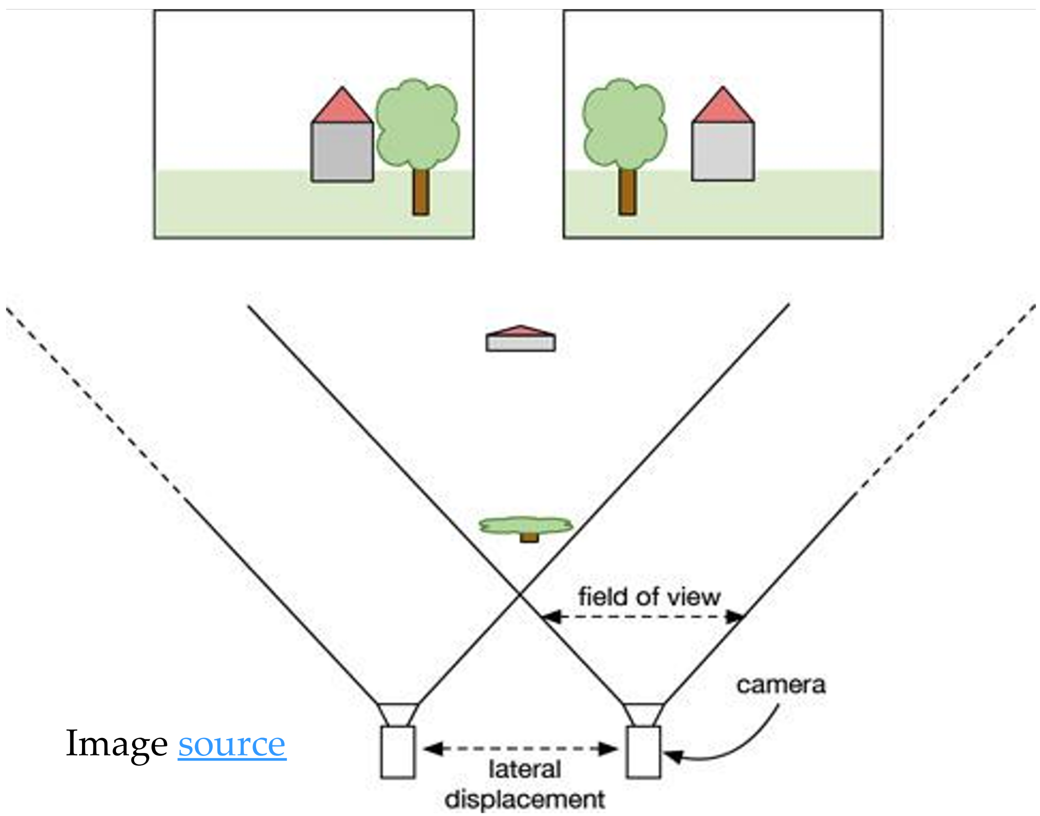



Image source

# Stereo Vision

- Recall the epipolar constraint in multi-view vision

$$q^T(T \times R)p = 0$$

- In the case of stereo vision

$$R = I_{3\times3}, \quad T = \begin{bmatrix} -t_x \\ 0 \\ 0 \end{bmatrix}$$

$$T \times R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & t_x \\ 0 & -t_x & 0 \end{bmatrix}, p = \begin{bmatrix} p_x \\ p_y \\ f \end{bmatrix}, q = \begin{bmatrix} q_x \\ q_y \\ f \end{bmatrix}$$

$$\begin{bmatrix} q_x & q_y & f \end{bmatrix} \begin{bmatrix} 0 \\ f t_x \\ -t_x p_y \end{bmatrix} = 0 \Longrightarrow p_y = q_y$$

Corresponding points are in the same rows of left and right images!

# Stereo Vision

- In this case, $\mu q = R\lambda p + T$ reduces to

$$\mu \begin{bmatrix} q_x \\ q_y \\ f \end{bmatrix} = \lambda \begin{bmatrix} p_x \\ p_y \\ f \end{bmatrix} + \begin{bmatrix} -t_x \\ 0 \\ 0 \end{bmatrix}$$

- which yields

$$\mu = \lambda$$
$$q_y = p_y$$
$$\mu = \lambda = \frac{t_x}{p_x - q_x}$$

$$disparity: p_x - q_x$$

# Stereo Vision

- The depth information can be recovered from the disparity of the corresponding image points along the $x$-axis!

$$z = \mu f = \lambda f = \left(\frac{1}{p_x - q_x}\right) t_x f$$

- $t_x$: stereo camera baseline
- 3D positions of the point in the right and left camera frames are given by

$$\mu q = \begin{bmatrix} \frac{t_x}{p_x - q_x} q_x \\ \frac{t_x}{p_x - q_x} q_y \\ \frac{t_x}{p_x - q_x} f \end{bmatrix}, \qquad \lambda p = \begin{bmatrix} \frac{t_x}{p_x - q_x} p_x \\ \frac{t_x}{p_x - q_x} p_y \\ \frac{t_x}{p_x - q_x} f \end{bmatrix}$$

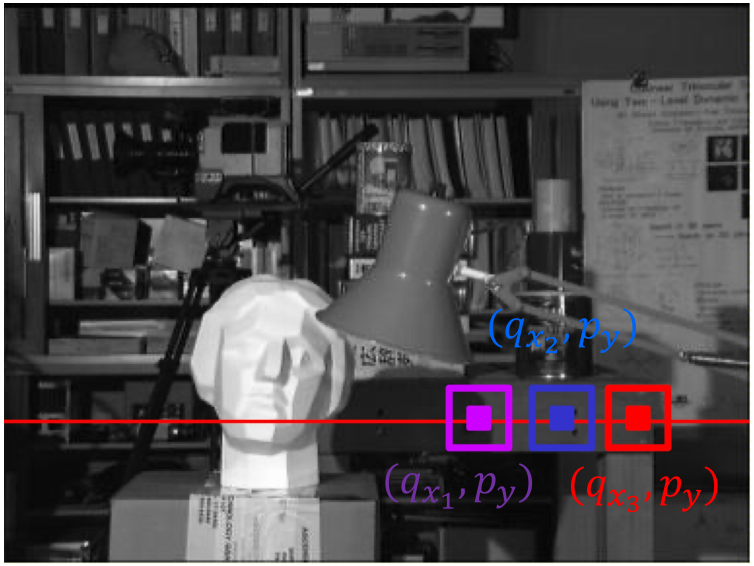
# Stere Vision

- The problem now is how to find *matching* pixels in the left and right images
- Recall that in stereo vision with parallel image planes, corresponding points in the left and right image are in the same row

$$p_y = q_y$$

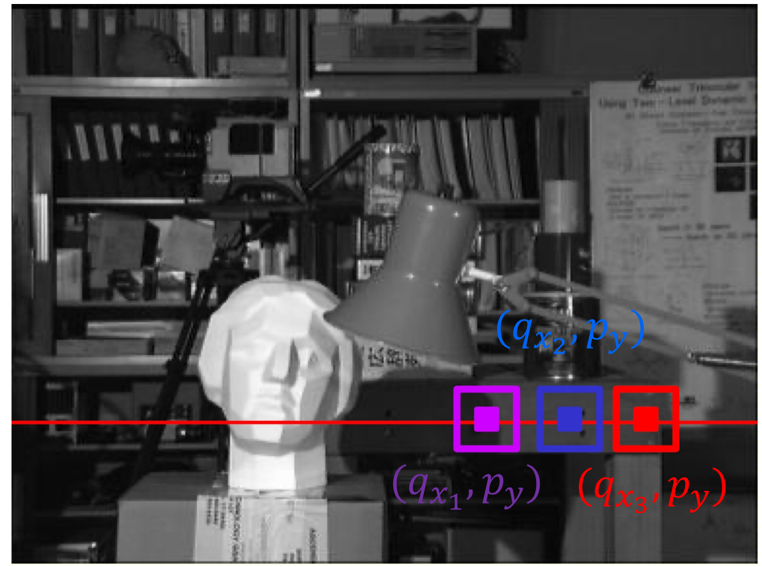
- So for each point $p$ in the left image, we must search for the best matching point in right image on the line $q_y = p_y$
- Only points with $q_x \leq p_x$ need to be checked (Why?)

# Stereo Vision

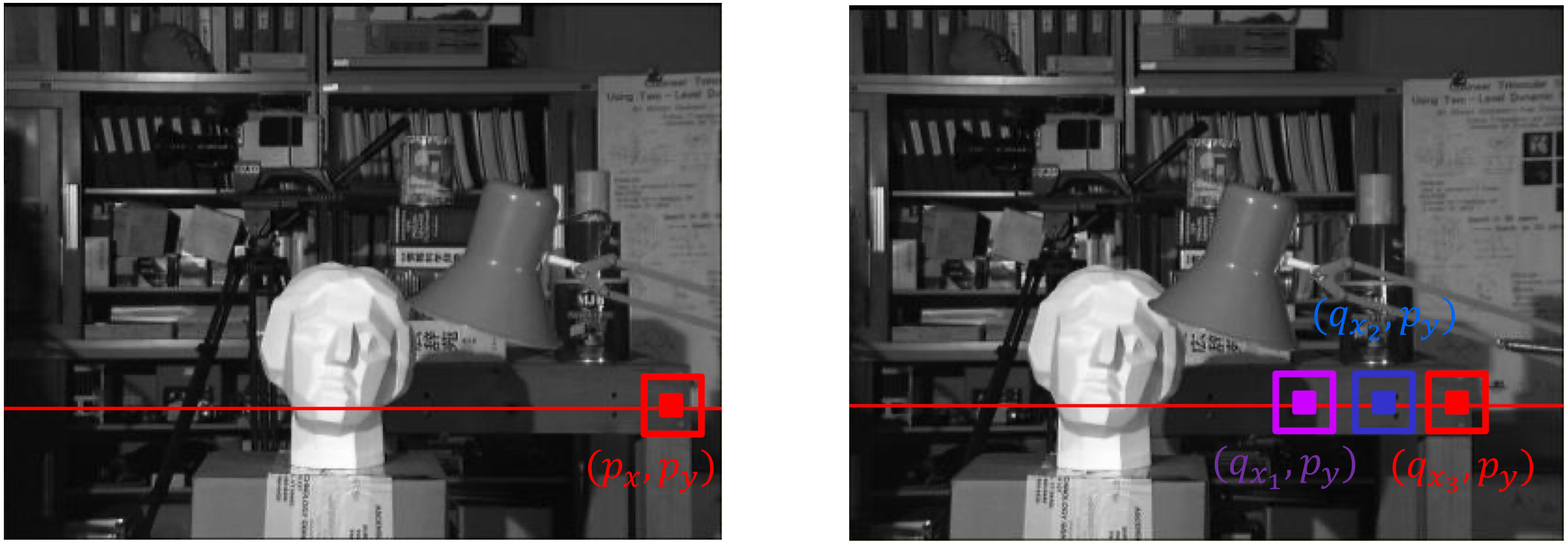

- Points in the right image that can potentially correspond to the point in left image $(p_x, p_y)$ are all on a horizontal line. Three candidates are shown in the right image

# Stereo Vision

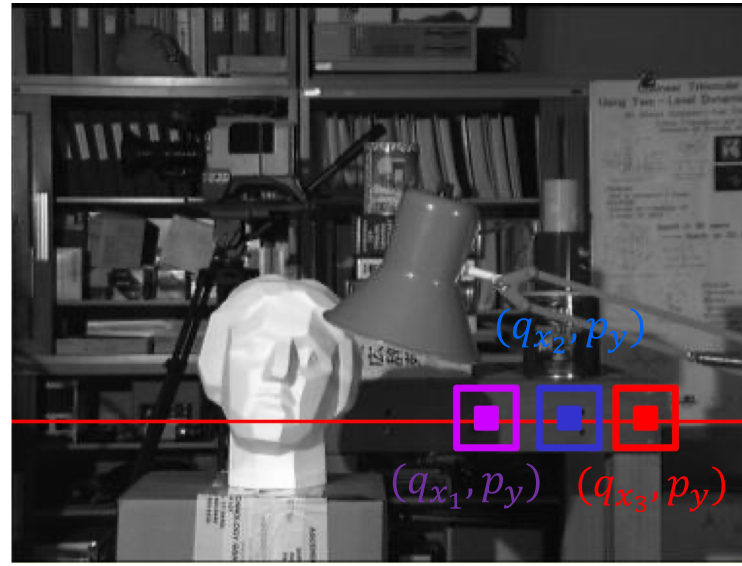

- What is the best match in the right image for the selected point in the left image?
- We need to define a metric of *closeness* of the match between two points

# Stereo Vision




- We can compare two patches of images (colored boxes in the images above) centered around the potential matching points
- A popular criteria for comparison is Sum of Squared Differences (SSD) of the image intensities in the two image patches

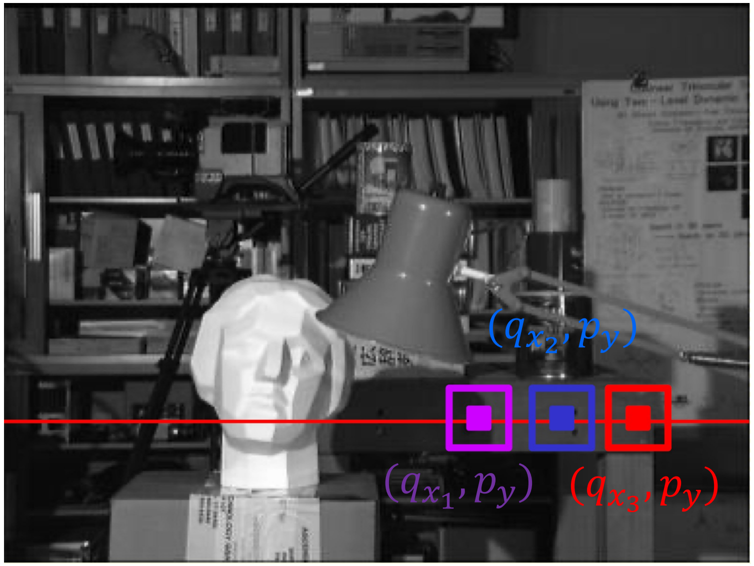
# Stereo Vision


$$d(p_x, p_y) = \arg\ \min \sum_{x=-\delta}^{x=\delta} \sum_{y=-\delta}^{y=\delta} \Big(I_l(p_x + x, p_y + y) - I_r(p_x - d + x, p_y + y)\Big)^2$$

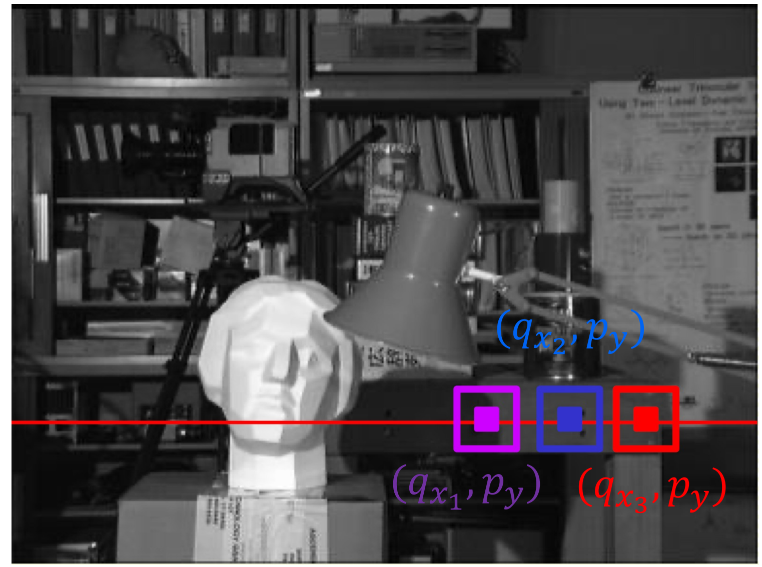
# Stereo Vision

McMaster University


- $d(p_x, p_y)$ is the disparity value at the point $(p_x, p_y)$ in the left image. The 3D position of the point in the left camera frame is given by

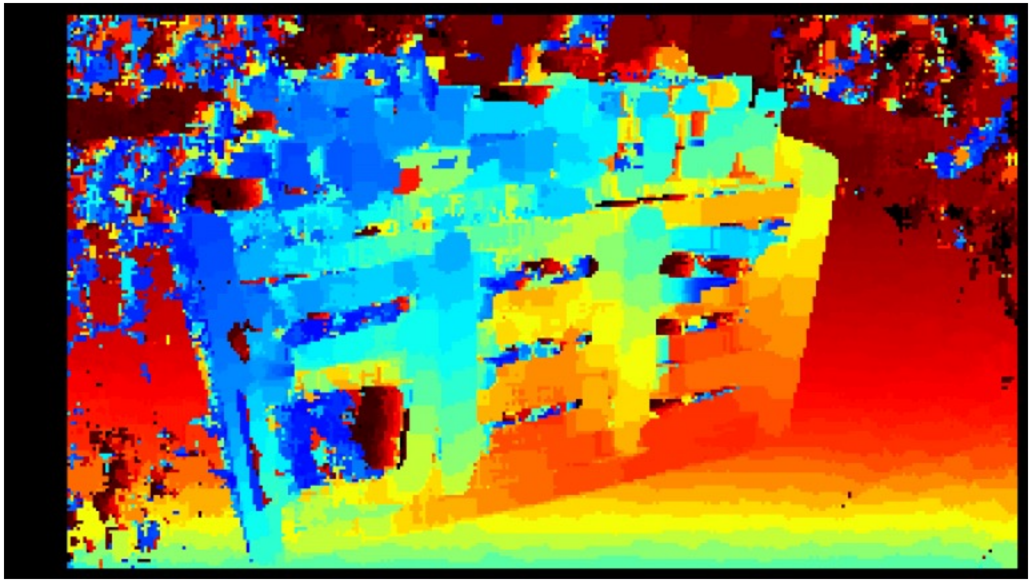
# Stereo Vision


$$\lambda p = \begin{bmatrix} \frac{t_x}{d(p_x, p_y)} p_x \\ \frac{t_x}{d(p_x, p_y)} p_y \\ \frac{t_x}{d(p_x, p_y)} f \end{bmatrix}$$

# Stereo Vision

Left and right images (image source)



Depth map produced from disparity map (image source)