Computer Vision

Forsyth & Ponce

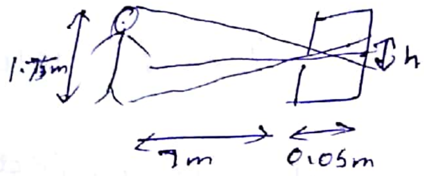
Hartley & Zisserman

Rick Szeliski

2

Height of person = 1.75m

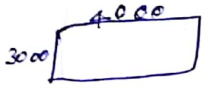
Distance " from camera = 7m



f = 50mm

sensor height = 3cm

resolution = 4000×3000 ⇒



, height of person in pixels in image = ?

→ 3cm 3000p

$h = 0.05 \times \frac{1.75}{7} = 1.25\text{ cm}$

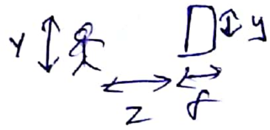
1.25cm 1250p

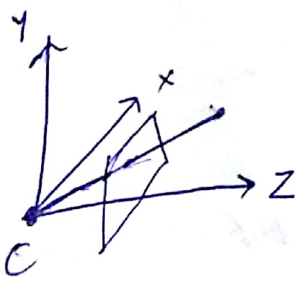
e

, camera raised by 1m, how much does the person move in the sensor?

$\frac{4}{7} \times 1250 \simeq 800$

$\Delta y = f \frac{\Delta Y}{Z}$





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

These are non-linear.

↓ homogeneous coordinates

linear

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} = \underbrace{\begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{P} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

$\Rightarrow \bar{x} = P\bar{X}$

$\bar{x}$ (image coordinates)

Camera (3×4 matrix)

$\bar{X}_w$ (world coordinates)

↓

Here it is $\bar{X}_c$ (camera coordinates)

$$\bar{x} = P\bar{X}_c = \begin{pmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ 0 & 0 & 1 \end{pmatrix} [I|0]\, \bar{X}_c = K[I|0]\, \bar{X}_c$$

Here it is $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

$K \to$ internal camera calibration matrix

Principal point $\to$ where the principal axis and image plane intersect

f_x, f_y measured in no. of pixels.

If the object is shifted, skewed (s)

$$K = \begin{pmatrix} f_x & s & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{pmatrix} \to \text{5 degrees of freedom in general}$$

If camera is moved (to c) and rotated by R,

$$X_c = \begin{pmatrix} R & -RC \\ 0 & 1 \end{pmatrix} X_w$$

$$x = K[I|0]\, X_c = K[R|-RC]\, X_w$$

$x = PX_w$, where $P = [\underbrace{KR}_{3\times 3} | \underbrace{-KRC}_{3\times 1}]_{3\times 4} = [M|\cancel{p_4}]$

$P = [p_1\ p_2\ p_3\ p_4] = [p^1\ p^2\ p^3]$

$M = KR,\ p_4 = -KRC$

For othographic projection, left submatrix is singular

Harris corner detector

$$A(x,y) = \sum_{w(x,y)} \left(I(x,y) - I(x+\Delta x, y+\Delta y)\right)^2$$

From Taylor series, approx. to

$$A(x,y) = \sum_{w} \left(I(x,y) - I(x,y) - I_x(x,y)\,\Delta x - I_y(x,y)\,\Delta y\right)^2$$

$$= \sum_{w} (I_x \Delta x)^2 + (I_y \Delta y)^2 + 2 I_x I_y \Delta x \Delta y$$

$$= (\Delta x \quad \Delta y) \left(\sum_{w} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \right) \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

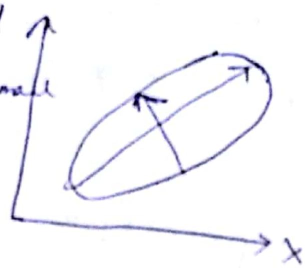
$$= (\Delta x \quad \Delta y)\, M \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$ → Equation of ellipse

<u>Eigen values of M</u>

$\lambda_1 >> \lambda_2$ ① Ellipse → edge $R<0$

$\lambda_1 \approx \lambda_2 \approx 0$ ② small circle → flat region, $|R|$ small

$\lambda_1 \approx \lambda_2 \to$ large ③ Big circle (ellipse) → corner $R>0$



PCA

$$M = \begin{pmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{pmatrix}$$ → covariance matrix

Corner response, $R = \det(M) - \alpha(\mathrm{trace}(M))^2$

$$= \lambda_1 \lambda_2 - \alpha(\lambda_1 + \lambda_2)^2$$

Alternate formulation, $f = \dfrac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} = \dfrac{\det(M)}{\mathrm{trace}(M)}$

Non-maximal subtraction → Local maximum of R

Invariance - rotation, intensity scaling (partially)

SIFT

① Take gaussian ~~at different σ~~ (within an octave) and sub sample (for multiple octaves)

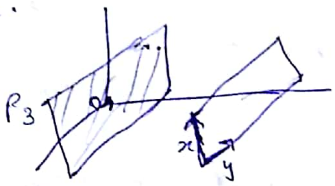
② LoG or DoG $\rightarrow \frac{1}{2\pi\sigma_1^2} e^{-\frac{x^2+y^2}{2\sigma_1^2}} - \frac{1}{2\pi\sigma_2} e^{-\frac{x^2+y^2}{2\sigma_2^2}}$

③ Take maxima in the DoG (26 neighbors) (~~At~~ In various octaves)

$\begin{bmatrix} 3 \\ 4 \\ 5 \\ 0 \end{bmatrix}$ $\rightarrow$ Points at ∞
OR
Vanishing points

P_1, P_2, P_3 are the images of vanishing points of the world x, y, z ~~coordinates~~ directions.

p^3 is the principal ~~axis~~ plane. The points on the plane are parallel to image plane.



HW

1. The principal point (image point) is given by $x_0 = Mm_3$, m_3 is the third row of M

2. $\det(M)\, m_3$ gives principal axis as a vector from the camera center through the principal point to the front of the camera.

Camera Calibration

1. 3D reference object based
2. From a precisely moving plane
3. Using a plane with unknown motion
4. From a set of collinear points that moves such that lines pass through a fixed point
5. Self calibration

1. 3D reference object based

$$\begin{pmatrix} x_i \\ y_i \\ w_i \end{pmatrix} = P \underbrace{\begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix}}_{X}$$

$$u_i = \frac{x_i}{w_i} = \frac{p^1 X}{p^3 X}$$

$$v_i = \frac{y_i}{w_i} = \frac{p^2 X}{p^3 X}$$

} Two equations from 1 point

12 parameters → require 6 points

Decomposing P,

$$P = K[R|t]$$

$KR \rightarrow 3\times3$ submatrix from $P_{3\times4}$

To find K,

$$(KR)(KR)^T = KK^T \quad (\because RR^T = I)$$

$$K = \begin{pmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix}$$

* Any non-singular K is a camera

$$Kt = P_4$$

$t = K^{-1}P_4 \Rightarrow$ K should be non singular

Refining P, (non-linear optimization)

$$\min_{P} \sum_{i} \| x_i - \phi[\ \ ) \|^2$$

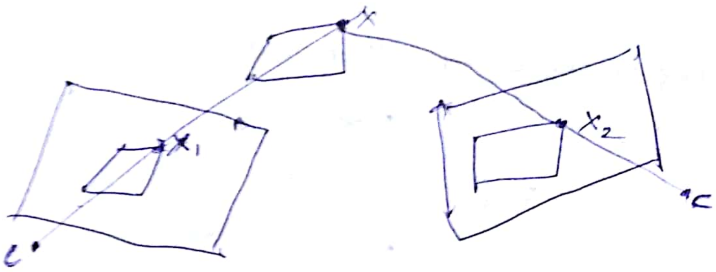
Radial distortion, (non-linear)

$$r_c^2 = x_c^2 + y_c^2$$

- Structure from motion

1. <u>Planar World</u>

$z = 0$



$$\bar{x}_1 = K_{3\times3}[R\ t]_{3\times4}\bar{X}$$

$$= K(r_1\ r_2\ r_3\ t)\begin{pmatrix} X \\ Y \\ 0 \\ 1 \end{pmatrix}$$

$$= \underbrace{K_{3\times3}(r_1\ r_2\ t)_{3\times3}}\begin{pmatrix} X \\ Y \\ 1 \end{pmatrix}$$

$$\bar{x}_1 = H_1 \bar{X}_{3\times3}$$

$\bar{x} = P\bar{X}$

* ~~If~~ P is not invertible, info is lost.

H_1, H_2 are invertible.

$$\bar{x}_2 = H_2\bar{X}$$

Eg: Camscanner

$$\Rightarrow \bar{x}_1 = H_1 H_2^{-1}\bar{x}_2$$

$$\left.\begin{aligned} \bar{x}_1 &= H_{12}\bar{x}_2 \\ \bar{x}_2 &= H_{21}\bar{x}_1 \end{aligned}\right\} \Rightarrow$$ Every point in the 1st image maps to a point (homography) in the 2nd image

Assumption: ~~the~~ ~~take~~ Some part of the world is captured since FoV ~~of~~ camera is limited.

ɣ a homography between the two images

the world and the image

//

2. Same Camera Center

\- 3D world

we lost the info about the 3rd dimension.

The line joining X and C intersects the two image planes where the images of X are formed on the planes.

$$\bar{x}_1 = K_1 R_1 (I \ -c)\bar{X}$$

Eg: Panarama

$$\bar{x}_2 = K_2 R_2 (I \ -c)\bar{X}$$

$$= K_2 R_2 (K_1 R_1)^{-1} (K_1 R_1)(I \ -c)\bar{X}$$

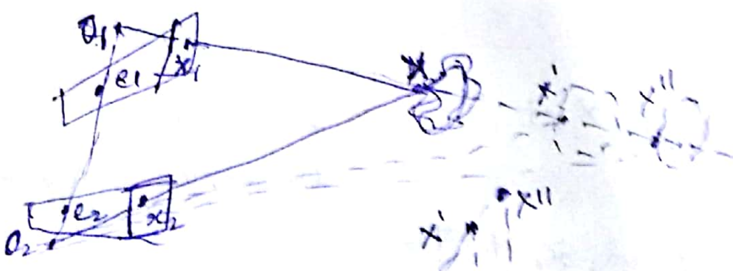
* Even if K_1, K_2 are different, it is homography

$$\bar{x}_2 = (K_2 R_2)(K_1 R_1)^{-1}\bar{x}_1$$

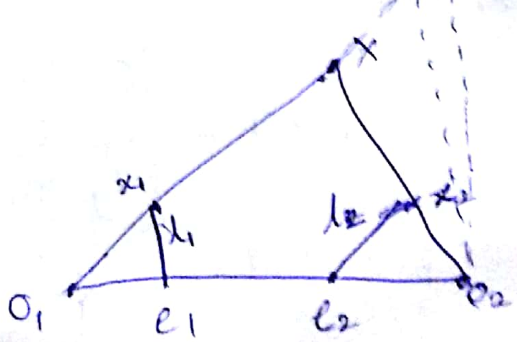
$$\boxed{\bar{x}_2 = \underset{3\times 3}{H_{21}}\,\bar{x}_1}$$ ⇒ homography → info is not lost

3. Generic world and cameras

we don't know $\bar{X}$



→ $\bar{X}$ lies on this ray only



→ plane

l_1 & l_2 lie on the plane

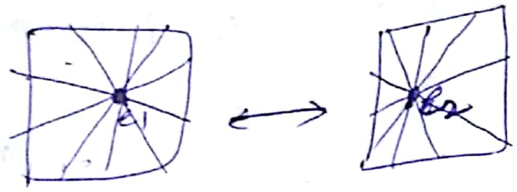
All the points on l_1 lies on l_2 and vice versa.

~~$e_1 e_2$ are~~ $\overline{O_1O_2}$ remain same.

The plane moves up/down such that $\overline{O_1O_2}$ doesn't change.

$\Rightarrow$ e_1 & e_2 ~~do~~ don't change.

$\Rightarrow$ l_1 & l_2 are set of lines passing through e_1 & e_2 respectively.



one-one mapping b/w the lines in 1st & 2nd planes.

* <u>Epipolar Geometry</u>

$l_1, l_2 \rightarrow$ epipolar lines

$e_1, e_2 \rightarrow$ epipoles

The plane is called epipolar plane

If O_1 is the origin,

$$\lambda_1 \bar{x}_1 = \bar{X}$$

$$\lambda_2 \bar{x}_2 = RX + T \quad \rightarrow \text{such that } O_2 \text{ becomes origin}$$

$$\lambda_2 \bar{x}_2 = R \lambda_1 \bar{x}_1 + T$$

Cross product with T

$$\lambda_2 \hat{T} \bar{x}_2 = \lambda_1 \hat{T} R \bar{x}_1 + 0 \qquad T \times T = \underbrace{\begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ T_y & T_x & 0 \end{bmatrix}}_{\hat{T}} \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix}$$

multiply by $\bar{x}_2^T$

But $T \times T = 0$

$$\Rightarrow \lambda_2 \underbrace{\bar{x}_2^T \underbrace{\hat{T} \bar{x}_2}_{\text{cross product}}}_{\text{dot product}} = \lambda_1 \bar{x}_2^T \hat{T} R x_1$$

$$\Rightarrow 0 = \lambda_1 \bar{x}_2^T \hat{T} R \bar{x}_1$$

$$\Rightarrow \bar{x}_2^T \hat{T} R x_1 = 0$$

$E = \hat{T} R \rightarrow$ symmetric

$$\Rightarrow \boxed{\bar{x}_2^T E_{3\times3} x_1 = 0}$$

$$\frac{x_2^T}{\lambda_2} \frac{E x_1}{\lambda_1} = 0$$

$$\boxed{\bar{x}_1^T E x_2 = 0}$$

$E \rightarrow$ Essential matrix

Here, K is assumed to be I (strongly caliberated)

Otherwise, (weakly caliberated)

$\lambda_1 x_1 = K_1 X$

$\lambda_2 x_2 = k_2 (RX + T)$

$$\Rightarrow x_2^T K_2^{-T} \hat{T} R K_1^{-1} x_1 = 0$$

$x_2^T F_{3\times3} x_1 = 0 \rightarrow x_2$ falls on the line $Fx_1 = 0$

$F \rightarrow$ Fundamental matrix

Scanned by CamScanner

RANSAC

Increases the no. of inliers.

The no. of points lying within the two lines - inliers

Total no. of points = 42

Points required to find k = 6

Project the remaining 36 poi

Find MST (reprojection error)

Repeat this with differer

Use the k that gives m

Homography

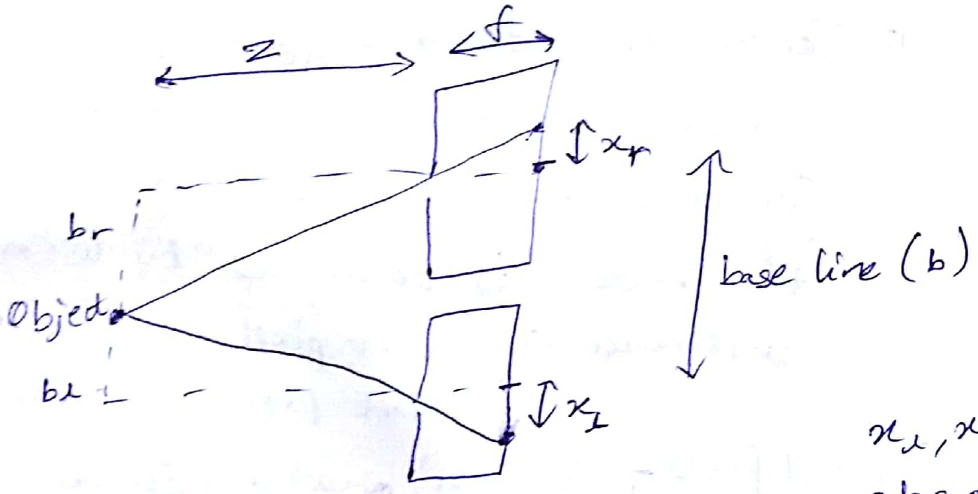
$$\underset{A}{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{18} \\ a'_{11} & a'_{12} & \cdots & a'_{18} \\ & & & \\ & & & \end{bmatrix}_{2n \times 8}} \underset{h}{\begin{bmatrix} h_{11} \\ h_{12} \\ \vdots \\ h_{32} \end{bmatrix}_{8 \times 1}} = \phi$$

(For n points)

$a_{ij} \rightarrow$ coefficient of h_{ij} obtained from the equations in ①

Stereo

shift $\propto \dfrac{1}{\text{distance to the object}}$



x_l, x_r are absolute values here.

$$\frac{x_l}{f} = \frac{b_l}{z}, \quad \frac{x_r}{f} = \frac{b_r}{z}$$

At $z = \infty$, both cameras see the same view

$$b_r + b_l = b$$

$$\frac{b_r}{z} + \frac{b_l}{z} = \frac{b}{z}$$

$$\frac{x_r + x_l}{f} = \frac{b}{z} \Rightarrow \boxed{z = \frac{fb}{x_l + x_r}}$$

Identifying common points

- Identifying similar windows

1. sum of square differences $||v_1 - v_2||_2$

2. " absolute " $||v_1 - v_2||_1$

3. Correlation → dot product → $\frac{v_1^T v_2}{\sqrt{v_1^T v_1}\sqrt{v_2^T v_2}}$

Normalized correlation

$$\rightarrow \frac{\bar{v}_1^T v_2}{\sqrt{\bar{v}_1^T v_1}\sqrt{\bar{v}_2^T v_2}}$$ Range $[-1, 1]$

4. Learn the function with deep network

5. Census transform - Go around the window in a circular manner (for speed)
Compare the two bit vectors.
(~~the~~ vector is obtained by using:
~~difference~~ next pixel is higher than prev → 1
next pixel is smaller than prev → 0)

6. Birchfield - Tomasi match (for accuracy)

$$z = \frac{fb}{x_l + x_r}$$

If there is an error in x_l or x_r,
z changes drastically (for small x_l, x_r)

⇒ For closer objects, we get a high resolution in depth (∵ z changes only by a small value)

Instead of checking for a window 1 pixel away, use $\frac{1}{2}$ or $\frac{1}{3}$ pixel away (can use the surface function to find the distance)

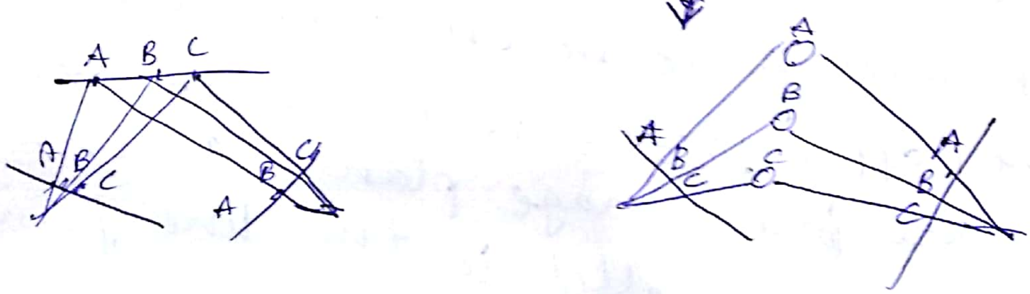
Constraints

1. Match lies on epipolar line of the pixel
   (instead of searching for the whole image)

2. Colour consistency

3. Uniqueness - a point in left image should match with only one in the right

4. Ordering/monotonicity (A, B are any two points)
   - If A is to the left of B in ~~left~~ image, it is left to B in the right image also
   ⇒ we search in only one direction on the epipolar line
   - violated if great difference in depth



Uniqueness + ordering ⇒ DP can be used to find the similar sequence.
$O(n^2)$

5. Continuity - Disparity values vary smoothly
   - violated at occlusion boundaries (depth doesn't change abruptly)

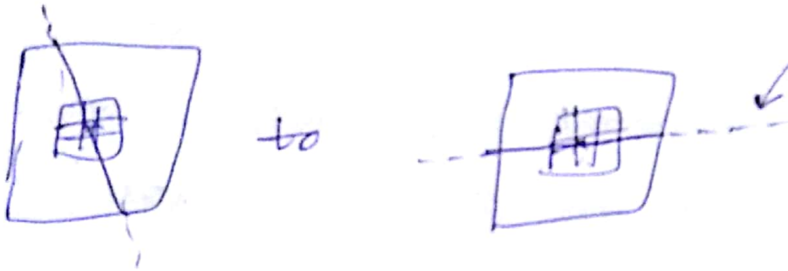
6. Sparse correspondence - good feature points
   Eg: Harris corner detection

7. Dense correspondence

Reduced Search & Rectification : Epipolar

If left & right cameras have same image plane and pure x-translation b/w them:

- Matches lie on same scan line



Rectification - Apply homography to 2nd image so that the image planes of the two cameras become parallel

- For matches to be on same scan line, apply stereo rectification, i.e., rotate both image planes so that they are parallel to the line joining the cameras

DP

- Now can be used only on scan lines