1.1 Epipolar Geometry

a) Assume you have two cameras, both with intrinsic parameters and rotations $\mathbf{K} = \mathbf{R} = \mathbf{I}$. Then for each of the three translation vectors $\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{t}_3$ given below, compute the essential matrix, describe the orientation of the epipolar lines and determine location of the epipoles for the resulting camera configurations.

$$\mathbf{t}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{t}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{t}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

$E = [t]_\times R$.

$$[t_i]_\times = \begin{bmatrix} 0 & -t_i^z & t_i^y \\ t_i^z & 0 & -t_i^x \\ -t_i^y & t_i^x & 0 \end{bmatrix}$$

We have

$$\tilde{E}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\tilde{E}_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}. \quad \tilde{E}_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

For the corresponding epipolar lines,

$\tilde{l}_1 = \tilde{E}_1 \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ y \end{pmatrix}$ horizontal

$\tilde{l}_2 = \tilde{E}_2 \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -x \end{pmatrix}$ vertical

$\tilde{l}_3 = \tilde{E}_3 \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} -y \\ x \\ 0 \end{pmatrix}$ 45° slanted

For the first two cases, the epipoles are ideal points at infinity.

Consider

$$\tilde{l}_2 = \begin{pmatrix} -y \\ x \\ 0 \end{pmatrix}$$

with $(1, 2)$ and $(1, 1)$, we can obtain the pipole as

$$\tilde{e}_2 = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} \times \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

b) In the third case of the previous problem, where does the baseline lie and what does that imply for the location of the epipoles?

Since we have pure translation in z-direction, the base line lies on the principal axis of the two cameras. Therefore the epipoles coincide with the principal point.

c) When is the fundamental matrix equal to the essential matrix? Discuss your reasoning.

Hint: *Think about the relationship between camera- and image coordinates.*

$F = K_2^{-T} \tilde{E} K_1^{-1} \overset{!}{=} \tilde{E}$

We obtain.

$K_2 = K_1 = I.$

Which equals $f_x = f_y = 1,\ s = 0.\ C_x = C_y = 0.$

1.2 Triangulation

a) Consider a system of two cameras with the following intrinsics and extrinsics:

$$\mathbf{K}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{K}_2 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{R}_2 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{t}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{t}_1 = \begin{pmatrix} -2 \\ 0 \\ 1 \end{pmatrix}$$

Furthermore, assume that you have observations for a point in both cameras:

$$\tilde{\mathbf{x}}_1^s = \begin{pmatrix} 1/4 \\ 1/2 \\ 1 \end{pmatrix}, \tilde{\mathbf{x}}_2^s = \begin{pmatrix} -1/5 \\ 1/5 \\ 1 \end{pmatrix}$$

For the given system, triangulate the 3D point $\tilde{\mathbf{x}}_w$ in world coordinates that corresponds to the observations. You can assume that the observations are exact.

$P = K\,[R\ \ t]$.

We have

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad P_2 = \begin{bmatrix} -2 & 0 & 1 & -3 \\ 0 & -2 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} x_i^s \tilde{p}_{i3}^T - \tilde{p}_{i1}^T \\ y_i^s \tilde{p}_{i3}^T - \tilde{p}_{i2}^T \end{bmatrix}}_{A_i} \tilde{x}_w = 0$$

Then.

$$A = \begin{bmatrix} -1 & 0 & \frac{1}{4} & 0 \\ 0 & -1 & \frac{1}{2} & 0 \\ 2 & 0 & -\frac{6}{5} & \frac{4}{5} \\ 0 & 2 & -\frac{2}{5} & -\frac{4}{5} \end{bmatrix}$$

Solving $A\tilde{x}_w = 0$

we obtain $\tilde{x}_w = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 1 \end{pmatrix}$

1.3 Stereo Vision

a) Show that for a stereo camera system the depth measurement error grows quadratically with depth.

$$z(d) = \frac{fb}{d}$$

Using first-order Taylor expansion,

$$z(d+\Delta d) = z(d) + \Delta d \frac{\partial}{\partial d}\left(\frac{bf}{d}\right)$$

$$= z(d) - \Delta d \frac{bf}{d^2} = z(d) - \Delta d \frac{z^2}{bf}$$

b) You can also think of this relationship as the depth resolution of the stereo camera system. How can we change the system setup to get a better depth resolution? What disadvantages might this have?

From the error term $\Delta d \frac{z^2}{bf}$ we can see that increasing the baseline b leads to a decreased error term and a better depth resolution. However, the wider the baseline is, the more difficult the matching problem becomes.

1.4 Block Matching

a) Consider two $K \times K$ windows of pixels flattened to vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{K^2}$. Show that the *Zero Normalized Cross-Correlation (ZNCC)* is invariant to changes in brightness in these windows. For this you can assume changes in brightness to be linear transformations of the form $\mathbf{w}_i' = \alpha_i \mathbf{w}_i + \mathbf{1}\beta_i$, where $\mathbf{1} \in \mathbb{R}^{K^2}$ is a vector of all ones.

$$ZNCC = \frac{(w_L(x,y) - \bar{w}_L(x,y))^T}{\|w_L(x,y) - \bar{w}_L(x,y)\|_2} \frac{(w_R(x-d,y) - \bar{w}_R(x-d,y))}{\|w_R(x-d,y) - \bar{w}_R(x-d,y)\|}$$

$$\frac{(w_1' - \bar{w}_1')^T}{\|w_1' - \bar{w}_1'\|_2} = \frac{(\alpha_1 w_1 + 1\beta_1 - \frac{1}{K^2}\sum_{k=1}^{K^2}(\alpha_1 w_1 + 1\beta_1))^T}{\|\alpha_1 w_1 + 1\beta_1 - \frac{1}{K^2}\sum_{k=1}^{K^2}(\alpha_1 w_1 + 1\beta_1)\|_2}$$

$$= \frac{(\alpha_1 w_1 + 1\beta_1 - \frac{1}{K^2}\sum_{k=1}^{K^2}\alpha_1 w_1 - 1\beta_1)^T}{\|\alpha_1 w_1 + 1\beta_1 - \frac{1}{K^2}\sum_{k=1}^{K^2}\alpha_1 w_1 - 1\beta_1\|_2}$$

$$= \frac{(\alpha_1(w_1 - \frac{1}{K^2}\sum_{k=1}^{K^2} w_1))^T}{\|\alpha_1(w_1 - \frac{1}{K^2}\sum_{k=1}^{K^2} w_1)\|_2}$$

$$= \frac{(w_1 - \frac{1}{K^2}\sum_{k=1}^{K^2} w_1)^T}{\|w_1 - \frac{1}{K^2}\sum_{k=1}^{K^2} w_1\|_2} = \frac{(w_1 - \bar{w}_1)^T}{\|w_1 - \bar{w}_1\|_2}$$

b) You are given the following pair of 5×7 stereo images, where part of the background is occluded by a thin structure (represented by the column of 10s). Using the *Sum of Squared Differences (SSD)* with a 3×3 window, determine the *Winner-Takes-All (WTA)* disparity $d \in \{0, 1, 2\}$ for the background pixel marked in Cyan. Discuss your result.

1	2	3	10	5	6	7
1	2	3	10	5	6	7
1	2	3	10	5	6	7
1	2	3	10	5	6	7
1	2	3	10	5	6	7

(a) Left image

2	10	4	5	6	7	8
2	10	4	5	6	7	8
2	10	4	5	6	7	8
2	10	4	5	6	7	8
2	10	4	5	6	7	8

(b) Right image

$d=0$, $SSD(5,3,0) = 3\times((10-5)^2+(6-6)^2+(7-6)^2) = 81$

$d=1$, $SSD(5,3,1) = 3\times((10-4)^2+(5-5)^2+(6-6)^2) = 108$

$d=2$, $SSD(5,3,2) = 3\times((10-10)^2+(5-4)^2+(6-5)^2) = 6$.

Hence, $d=2$ would be the property disparity.

c) Below we have the full disparity maps for the images from the previous exercise computed from the left- to the right image and from the right- to the left image, respectively. Perform a left-right consistency test for the pixels marked in Cyan, Green and Red. Which of the points pass the test? Is the test succesful in determining incorrect disparity estimates?

Remark: *The disparities were computed in the same way as in the previous exercise. To compute disparities along the image boundaries, the images were padded with zeros.*

0	1	2	2	2	1	0
0	1	2	2	2	1	0
0	1	2	2	2	1	0
0	1	2	2	2	1	0
0	1	2	2	2	1	0

(a) Left → Right

2	2	2	2	1	0	0
2	2	2	2	1	0	0
2	2	2	2	1	0	0
2	2	2	2	1	0	0
2	2	2	2	1	0	0

(b) Right → Left

Cyan: $d_{L\to R}(6,2) \to d_{R\to L}(6-1,2)$

$d_{R\to L}(5,2) \to d_{L\to R}(5+1,2)$ consistent

Green: $d_{L\to R}(4,3) \to d_{R\to L}(4-1,3)$

$d_{R\to L}(3,3) \to d_{L\to R}(3+1,3)$ consistent

Red: $d_{L\to R}(3,5) \to d_{R\to L}(3-1,5)$

$d_{R\to L}(2,5) \to d_{L\to R}(2+1,5)$. consistent

1.5 Learned Stereo and End-to-End Models

a) Recent approaches in end-to-end disparity estimation often build a disparity cost volume and apply 3D convolutions to it to estimate the final disparity map. However, 3D convolutions are computationally expensive, limiting both the resolution and maximum disparity that can be used.

Below we consider two sequences of two 2D- and 3D convolutions, respectively, applied to the same input tensor. We describe the layer configurations as ConvND(C_{in}, C_{out}, k), where input channels are denoted by C_{in}, output channels by C_{out} and k is the kernel size. For simplicity you can assume square kernels, as well as appropriate padding and a stride of one, such that the spatial dimensions remain unchanged. Shapes are specified by the number of channels followed by the spatial dimensions, so (Channels, Height, Width) for 2D convolutions and (Channels, Depth, Height, Width) for 3D convolutions.

For both sequences, calculate the total amount of memory required to store the activations and trainable parameters. For this you can fill in the blank fields in the table below.

Layer	Input Shape	Output Shape	# Trainable Parameters	Memory
Conv2D(32, 64, 3)	(32, 128, 128)	(64, 128, 128)	9x32x64+64	
Conv2D(64, 128, 3)	(64, 128, 128)	(128, 128, 128)	9x64x128+128	
Conv3D(1, 64, 3)	(1, 32, 128, 128)	(64, 32, 128, 128)	9x1x64+64	
Conv3D(64, 128, 3)	(64, 32, 128, 128)	(128, 32, 128, 128)	9x64x128+128	

Memory1 = (64x128x128+128³ + 9x32x64+64 + 9x64x128+128) x4 / 1000² = 12.95 MB

Memory2 = (64x32x128x128 + 128³x32 + 9x1x64+64 + 9x64x128+128) x4 / 1000² = 402.95 MB

b) You are working with a GC-Net-style architecture to solve a disparity estimation problem. For two particular pixels p_1 and p_2 in the cost volume, the network estimates the following matching costs:

- p_1: $c_\theta(d) = [1.0, 3.0, 10.0, 3.0, 1.0]$
- p_2: $c_\theta(d) = [10.0, 2.0, 1.0, 2.0, 10.0]$

where $d \in \{0, 1, 2, 3, 4\}$. For both pixels, calculate the expectation of the disparity and discuss the result. GC-Net uses the discrepancy between the expected and the ground-truth disparity as a loss function during training. What kind of behaviour does this encourage?

Hint: *Compare the distributions over disparities implied by the cost vectors with the final result.*

Softmax is define as $\sigma(z_i) = \frac{e^{z_i}}{\sum_{i=1}^{n} e^{z_i}}$

For p_1 :

$\sigma(-c_\theta) = [0.44, 0.06, 0.00, 0.06, 0.44]$.

$E[d] = 0.44 \times 0 + 0.06 \times 1 + 0.00 \times 2 + 0.06 \times 3 + 0.44 \times 4 = 2$.

For p_2 :

$\sigma(-c_\theta) = [0.00, 0.21, 0.58, 0.21, 0.00]$

$E[d] = 2$.