Introduction to Computer Vision
Assignment 3: Stereo and Structure from Motion

Submission: Individually, here
Due Date: Wednesday, Dec 16, 2020

1. You are given two camera matrices: $P_1 = K\begin{bmatrix} I_3 & 0 \end{bmatrix}$, $P_2 = K\begin{bmatrix} R_2 & t_2 \end{bmatrix}$ where:

$$K = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$t_2 = \begin{bmatrix} 1 & 1 & -1 \end{bmatrix}^T$$

   (a) Find x_1 and x_2, the projections of the point $X = \begin{bmatrix} -1 & 1 & 2 & 1 \end{bmatrix}^T$ to P_1 and P_2.

   (b) You are given the fundamental matrix between the two images:

$$F = \begin{bmatrix} 0 & \sqrt{2} & -\sqrt{2} \\ -1 & -1/\sqrt{2} & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

   Show that F fulfills the requirement $x_2^T F x_1 = 0$.

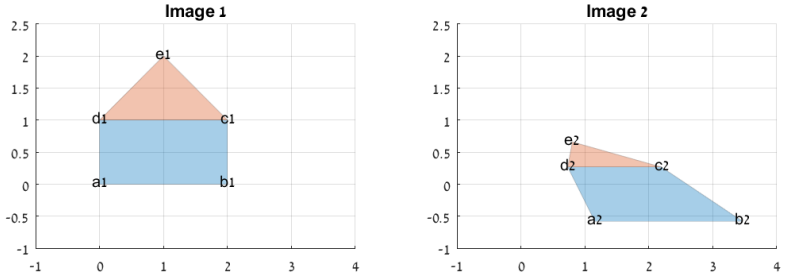
   (c) Let l_1 be the epipolar line that contains the projection of x_2 in the first image, and l_2 the epipolar line that contains the projection of x_1 in the second image. Find l_1, l_2.

   (d) Show that $e_1 = \begin{bmatrix} \frac{\sqrt{2}-1}{\sqrt{2}} & 1 & 1 \end{bmatrix}^T$ and $e_2 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$ are the right and left epipoles and that e_1 is on l_1 and e_2 is on l_2.

2. Two cameras view a house, as seen in the following figure. 5 of the corners of the house are marked A-E. 4 of these 3D points (**A**,**B**,**C**,**D**) are on the plane $Z = 1$ and **E** is on the plane $Z = 2$.

$$P_1 = \begin{bmatrix} I_3 & 0 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} R_2 & t_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & \sqrt{3}/2 & -1/2 & 0 \\ 0 & 1/2 & \sqrt{3}/2 & 0 \end{bmatrix}$$



(a) Find the homography $H \in \mathbb{R}^{3\times3}$ that satisfies:

$$x_2 \propto Hx_1$$

For $x \in \{a, b, c, d\}$

(b) What is the location of the 3D point **E**? (Hint: Use P_1, e_1 and the given depth of **E**).

(c) When applying H on e_1 do we get e_2? Why?

3. A camera is imaging an object at two time instances. Let $P_i = (X_i, Y_i, Z_i)^T$, $i = 1, ..., N$ be a set of 3D points in space at the first time instance, and $p_i = (x_i, y_i, f)^T = \frac{f}{Z_i} P_i$. At the second time instance the points are displaced relative to the camera by $P_i' = RP_i + t$, where R is the rotational component and t is the translational component of

either the camera or object motion. In the second image therefore the points are projected to $p'_i = \frac{f}{Z'_i} P'_i$. Assume point correspondences are given across the two images.

Can depth (Z_i) be recovered if the relation between the two images is:

(a) A camera rotation ($t = 0$)?

(b) An objeted rotation (about its center of mass)?

If depth cannot be recovered, **show explicitly** that it can be eliminated from the equations relating the two images. (*Hint:* show the existence of a mapping from p_i to p'_i that does not include 3D structure (i.e., depth Z_i). Explain why the existence of such a mapping proves the claim.)

4. (a) **Show** that the right epipole (the epipole in the first image) is given by $v = \alpha R^T t$ (for some $\alpha \neq 0$). (*Hint:* to show this show that $Ev = 0$.) Show that v is the intersection of all epipolar lines (and thus must be the epipole).

   (b) **Derive** an expression for the left epipole (in other words, find u such that $u^T E = 0$). (*Hint:* invert the rigid transformation that relates P and Q and use the expression derived for the right epipole.)

5. A calibration pattern is a set of 3D points at known positions. Such a pattern is used to calibrate a stereo rig (determine the transformation between the two cameras).

   Suppose we take two images of a calibration pattern with the cameras at unknown positions and orientations (so a point P from the pattern is expressed as $R_1 P + t_1$ in the coordinate system of image 1; and $P \rightarrow R_2 P + t_2$ in image 2). To calibrate the cameras in this case we need a more general expression for the essential matrix.

   Express the essential matrix between the two images in terms of R_1, t_1, R_2, t_2.

6. (a) **Show** that the intersection of any two parallel lines in the projective plane $\mathbb{P}^2$ lies at infinity.

   (b) Let H be a homography. **Show** that H maps lines to lines.

   (c) Suppose H represents an affine transformation. **Show** where H maps the line at infinity.

7. `RANSAC` is a method for finding a transformation between images from point matches when some (possibly many) of the matches are erroneous. The objective of `RANSAC` is to identify the transformation that is supported by the largest number of matches.

   Given two images, suppose we use a feature detector to identify m points in the first image and n points in the second image, and suppose we do not know in advance which of the m points in the first image matches which of the n points in the second image. **Write** an expression for the number of transformations that may be computed with `RANSAC` in the worst case and the total run-time complexity for the following cases:

   (a) The transformation relating the two images is a homography.

   (b) The two images are related by a general fundamental matrix and we use the 8-point algorithm to compute the transformation.

8. Similar to the previous question, suppose now we match feature descriptors and obtain m *pairs* of points, $\{(p_i, q_i)\}_{i=1}^{m}$, where $p_i \in I_1$ and $q_i \in I_2$. Answer the following questions in case: (i) the transformation relating the two images is a homography; and (ii) the images are related by an essential matrix and we use the 5-point algorithm to compute the transformation.

   (a) **What** is the number of transformations that may be computed with `RANSAC`.

   (b) Suppose of the m pairs of features half are accurate and half are incorrect (outliers). **What** is the probability that a minimal set drawn randomly will yield a correct transformation? **What** is the expected number of sets needed to be drawn to guarantee the success of `RANSAC`?

9. In Tomasi & Kanade's algorithm for multiview structure from motion we assume that the scene is static and that the projection is orthographic.
Assume there are m images and n points and that the 2D points are centered around zero.
Point P_j is projected to camera i using some rotation matrix whose first row is ${r_i}^T$ and second row is ${s_i}^T$. The projected points (x_{ij}, y_{ij}) are given as input, and we are looking for the optimal P_j, r_i, s_i that minimize the objective:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}({r_i}^T P_j - x_{ij})^2 + ({s_i}^T P_j - y_{ij})^2$$

S.t

$$\|r_i\| = \|s_i\| = 1,\ {r_i}^T s_i = 0$$

(a) Explain the objective and why minimizing it solves the SFM problem. Why do we require that $\|r_i\| = \|s_i\| = 1,\ {r_i}^T s_i = 0$?

(b) What should you change in the objective if the projection is scaled orthographic? (Meaning each camera is multiplied by some scalar c_i)

Tips:
In an orthographic projection point P_j is projected to camera i as:

$$(x_{ij}, y_{ij})^T = \Pi(R_i P_j + t_i)$$

In a scaled orthographic projection point P_j is projected to camera i as:

$$(x_{ij}, y_{ij})^T = c_i \Pi(R_i P_j + t_i)$$

where

$$\Pi(X, Y, Z)^T = (X, Y)^T$$

In this question t_i=0.