Winter 2023 - CSE/ECE 344/544 Computer Vision
End-Sem - May. 02, 2023

Maximum score: 65(UG)/75(PG) **Time: 120 minutes**

1. (20 points) State whether the following statements are **true** or **false** with appropriate justification.
   Marking scheme:- 2 points for correct answer and correct justification, 1 point for correct answer and incorrect justification, 1.5 points for correct answer and partially correct justification, 0 point for incorrect answer despite the justification written

   (a) (2 points) Harris corner detection is invariant to scale.
   False. As the scale changes, a corner may appear as a smooth edge and make the second moment matrix **H** have a low value of its smaller eigenvalue.

   (b) (2 points) For a stereo pair in general position (arbitrary rotation and translation) observing a general scene, a 3×3 homography matrix can be estimated that maps any point in the first image to the corresponding point in the second image.
   False. A 3×3 homography matrix is a planar homography and can only map points from one image to a point on another image if both images are that of the same plane in the 3D scene.

   (c) (2 points) The rank of the essential matrix is 1, when the relative translation between the stereo pair is aligned with either the x-axis or with the y-axis.
   False. The rank of an Essential matrix is *always* 2.

   (d) (2 points) In the image formation pipeline, the world to camera coordinate transformation is strictly an affine transformation.
   False. Since the World to Camera transformation always implements a rotation and translation, it is a Euclidean transformation and not affine.

   (e) (2 points) Projective transformations guarantee that parallel lines will remain parallel.
   False. Parallelism is only invariant under Euclidean, Similarity and Affine Transformations. The projective transformations map vanishing points and vanishing lines to finite points.

   (f) (2 points) The effect of radial distortion reduces as you radially move away from the principal point.
   False. The effect of radial distortion *increases* as you radially move away from the principal point. The model equation has the distortion effect directly proportional to the distance from the principal point.

   (g) (2 points) A line in the 2D projective plane $\mathcal{P}^2$ is represented by the 3-dimensional vector normal to the plane passing through the origin and the line.
   True. Such a plane uniquely describes the line and therefore the normal vector is an unambiguous representation of a line in $\mathcal{P}^2$

   (h) (2 points) Identifying the line at infinity is sufficient to remove projective distortion from the image of a 3D plane.

True. Using the line at infinity, we can find the homography **H**, the transformation which maps the line at infinity to its canonical position.

(i) (2 points) In a stereo system, both epipoles lie on the epipolar plane.
True. Epipoles are the points of intersection of the line joining the camera centres (the baseline) with the image plane. An epipolar plane is a plane containing the baseline.

(j) (2 points) In the Harris corner detection approach, an edge is detected if the second moment matrix **H** is rank-deficient and has large $||\mathbf{H}||_2$
True. Since second moment of matrix H is rank deficient, which means one of the eigenvalue is zero and a large $||\mathbf{H}||_2$ indicates that the other eigenvalue is large. Since we have one eigenvalue large and other eigenvalue zero, the gradient is dominant only in one direction (direction can be arbitrary).

2. (15 points) An important parameter of the imaging system is the *field of view* (FOV). The field of view is the twice the angle between the optical axis (z-axis) and the end of the retinal plane (CCD array). Imagine having a camera system with focal length 24 mm and retinal plane (CCD array) (16 mm $\times$ 12 mm) and that your digitizer samples your imaging surface at 500 $\times$ 500 pixels in the horizontal and vertical directions.

   (a) (5 points) Compute the FOV.

   (b) (5 points) Describe how the size of the FOV is related to the focal length and how it affects the resolution in the image.

   (c) (5 points) Write down the expression that relates the image coordinate and a point in 3D-space expressed in the camera coordinate frame of reference.

   Solution:

   (a) (5 points) vertical (portrait) $FOV_y = 2 \times \tan^{-1}(0.5 \times 12/f)$; (2.5 points) horizontal (landscape) $FOV_x = 2 \times \tan^{-1}(0.5 \times 16/f)$ (2.5 points)

   (b) (5 points) the larger the focal length, the smaller is the FOV (2.5 points). As the FOV increases and the number of image pixels is fixed , the resolution of the image is decreased, (2.5 points) i.e. the visual angle subtended by a single pixel is larger

   (c) (5 points) Let $P_c = \{X_c, Y_c, Z_c\}$ be the point in camera frame, then the expression is: $x_{im} = fX_c/Z_c$ (2.5 points) and $y_{im} = fY_c/Z_c$ (2.5 points)

3. (15 points) Edge detection and model fitting.

   (a) (5 points) Describe how non-maxima suppression works in Canny edge detector. Can you apply a similar strategy to eliminate any detection bounding boxes obtained from an object detector that significantly overlap? Briefly describe your strategy. {*Hint*: Instead of using gradient magnitude, you may use classification score as a measure.}

   (b) (10 points) Say you are trying to detect a circle of radius r in an edge image and wish to use RANSAC to find the circle. What is the minimum subset of points you would need to generate a model hypothesis? If there are 500 points, of which 250 are outliers, how would you decide the number of iterations of RANSAC to find a

reasonable hypothesis? Write the algorithm along with the residual function you will use. Now suppose there are multiple circles in the edge image, how would you identify all the circles? (You are not permitted to use RANSAC repeatedly.) {*Hint:* Use the equation of the circle $(x-a)^2+(y-b)^2=r^2$.}

Solution:

(a) Non-maximum suppression or edge thinning can help to suppress all the gradient values (by setting them to 0) except the local maxima, which indicate locations with the sharpest change of intensity value. The algorithm for each pixel in the gradient image is:

Compare the edge strength of the current pixel with the edge strength of the pixel in the positive and negative gradient directions within a neighborhood. If the edge strength of the current pixel is the largest compared to the other pixels in the mask with the same direction (i.e., the pixel that is pointing in the y-direction, it will be compared to the pixel above and below it in the vertical axis), the value will be preserved. Otherwise, the value will be suppressed. (3 points)

A similar approach can be taken for non-maxima suppression in case of detection. For detection bounding boxes that have a significant overlap (let's say intersection over union (IoU) of bounding boxes is > 0.7), then you retain only the box that has the highest classifier score and eliminate the others. (2 points)

(b) RANSAC for detecting a circle of radius r.

1. The minimum subset of points required for solving for a model of a circle of radius r is 2. Given two points ($\mathbf{p}=[x,y]^\top, \mathbf{p}'=[x',y']^\top$), we can draw a circle of radius r around both these points. These two circles will lead to zero (if $d(\mathbf{p},\mathbf{p}')>r$), one (if $d(\mathbf{p},\mathbf{p}')=r$) or two ($d(\mathbf{p},\mathbf{p}')<r$) intersections. The points of intersection will be the possible center of the circle. The intersection points of these circles give us the model parameters (i.e., the center (a,b)) by solving the quadratic equation $(x-a)^2+(y-b)^2=(x'-a)^2+(y'-b)^2$ for samples where $||\mathbf{p}-\mathbf{p}'||_2 \leq r$. (2 points)
2. If there are 250 outliers out of 500, we have an inlier fraction of 0.5.
   - (Approx.[1]) Probability of picking an *all inlier* minimal subset: 0.5^2
   - Probability that we *have not* picked an all inlier minimal subset: $1-0.5^2$
   - Probability that we have not picked an all inlier minimal subset over N independent trials: $(1-0.5^2)^N$
   - Assuming that this probability of failure is 0.01, we get $N\log(1-0.5^2)=\log(0.01)$.
   - Therefore $N \geq \frac{\log(0.01)}{\log(1-0.5^2)}$ would have a 99% chance of getting an all inlier minimal subset.

   (3 points)
3. The algorithm is given below:
   (a) Fetch the data and the inlier error threshold τ from the user.

[1] Ignoring the error due to sampling without replacement.

(b) Pick minimal subset of size 2, $\{\mathbf{p}, \mathbf{p}'\}$.

(c) Compute the model parameters (a, b) using the two points if $||\mathbf{p} - \mathbf{p}'||_2 \leq r$.

(d) For each model parameters, identify the inliers by plugging in each center's coordinate into the error (residual) function that satisfy this inequality: $(x-a)^2 + (y-b)^2 - r^2 \leq \tau$.

(e) Keep track of the inlier set for each model and start from Step 3b.

(f) Pick the model that has the largest number of inlier.

(3 points)

4. RANSAC does not help in discovering multiple structures from the data. We would then need to use Hough Transforms. The approach is given in Lec-10 deck, slide nos. 37-38. (2 points)

4. (15 points) Local feature detector and descriptors.

(a) (10 points) Write the error function for Harris corner detector in the quadratic form. Now devise a generalization of the Harris corner detector for 3D images, i.e, there is an intensity value for each voxel (x, y, z). Write the algorithm for performing corner detection on a 3D image and the test you would use for selecting corners.

(b) (5 points) How are SIFT descriptors computed? Describe in detail.

Solution:

(a) The error function for Harris corner detector is the following

$$\begin{aligned} E(u,v) &= \sum_{x,y \in W} \left(I(x+u, y+v) - I(x,y)\right)^2 \\ &\simeq \sum_{x,y \in W} \left(I_x(x,y)u + I_y(x,y)v\right)^2 \\ &= \sum_{x,y \in W} I_x^2(x,y)u^2 + 2I_x(x,y)I_y(x,y)uv + I_y^2(x,y)v^2 \\ &= \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \text{ (3 points)} \end{aligned}$$

For a 3D image, we will have

$$\begin{aligned} E(u,v,w) &= \sum_{x,y,z \in W} \left(I(x+u, y+v, z+w) - I(x,y,z)\right)^2 \\ &\simeq \sum_{x,y,z \in W} \left(I_x(x,y,z)u + I_y(x,y,z)v + I_z(x,y,z)w\right)^2 \\ &= \begin{bmatrix} u & v & w \end{bmatrix} \begin{bmatrix} A & B & C \\ B & D & E \\ C & E & F \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \text{ (4 points)} \end{aligned}$$

Algo. for Harris corner detection in 3D images.

- Compute the second moment matrix using gradient images in x, y and z direcions.

- Select corners based on eigendecomposition of the second moment matrix. The smallest eigenvalue λ_{min} of the matrix at a point should be larger than a certain threshold for the point to qualify as a corner feature.

(3 points)

(b) Steps to compute SIFT descriptors:

- Take a 16x16 window around the detected feature.
- Divide the 16x16 window into a 4x4 grid of cells.
- Compute edge orientation (angle of the gradient - 90°) for each pixel. Throw out the weak edges by thresholding the gradient magnitude.
- Create a histogram of the remaining edge orientations for each 4x4 cell with 8 bins in each histogram.
- Stack the 16 histograms with 8 bins in each to get 16 cells * 8 orientations = 128 dimensional feature descriptor.

(5 points)

5. (10 points) [*Extra Credit for UG; required for PG*]

(a) (5 points) Suppose that there are three camera frames C_0, C_1, C_2 and the coordinate transformation from the frame C_0 to frame C_1 is (R_1, T_1) and from C_0 to C_2 is (R_2, T_2). What is the relative transformation from C_1 to C_2 ? What about C_2 to C_1?. Express these transformations in terms of R_1, T_1 and R_2, T_2 only.

(b) (5 points) Given a 3×4 camera matrix $\mathbf{P}$, such that a world 3D point $\widetilde{\mathbf{X}}$ (homogeneous coordinates) gets mapped to the 2D point $\widetilde{\mathbf{x}}$ (again in homogeneous coordinates) via the relation $\widetilde{\mathbf{x}} = \mathbf{P}\widetilde{\mathbf{X}}$, show that the camera centre (center of projection or the optical centre of the camera; also the origin on the camera frame of reference) in the world coordinate frame is the null-space of $\mathbf{P}$. Is this the only vector in the null space of $\mathbf{P}$? Explain why or why not.

Solution:

(a) We can write the transformation from C_i to C_j as ${}^jT_i$. Then we have the following transformations:

$$ {}^1P_0 = \begin{bmatrix} R_1 & T_1 \\ 0 & 1 \end{bmatrix} $$

$$ {}^0P_1 = ({}^1P_0)^{-1}(0.5marks) = \begin{bmatrix} R_1^\top & -R_1^\top T_1 \\ 0 & 1 \end{bmatrix} (1marks) $$

(Inverses 1.5 mark each, if only written inv then 0.5, calc 1 mark each, if inverted 0.5 each) Similarly,

$$ {}^2P_0 = \begin{bmatrix} R_2 & T_2 \\ 0 & 1 \end{bmatrix} $$

$$ {}^0P_2 = ({}^2P_0)^{-1}(0.5marks) = \begin{bmatrix} R_2^\top & -R_2^\top T_2 \\ 0 & 1 \end{bmatrix} (1marks) $$

The transformation between Camera C_1 and C_2 will be:

$$^2P_1 = {}^2P_0\ {}^0P_1 (0.5marks) = \begin{bmatrix} R_2 & T_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_1^\top & -R_1^\top T_1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_2 R_1^\top & -R_2 R_1^\top T_1 + T_2 \\ 0 & 1 \end{bmatrix} (0.5marks)$$

Similarly, the transformation from Camera C_2 to C_1 will be: :

$$^1P_2 = {}^1P_0\ {}^0P_2 (0.5marks) = \begin{bmatrix} R_1 & T_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_2^\top & -R_2^\top T_2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_1 R_2^\top & -R_1 R_2^\top T_2 + T_1 \\ 0 & 1 \end{bmatrix} (0.5marks)$$

(b) See Quiz-3, Q3 solution (3 marks, 1 mark for writing that it is the only point in null space, 1 mark for explanation)

6. (20 points) [*Extra Credit (for UG&PG)*] Refer Fig. 1. Consider a point p on a 2D plane P in 3D space. Suppose that image points (in pixel coordinates) of p in two camera frames are given to be x_1^{pix} and x_2^{pix}. Let the coordinate transformation between the two frames be:

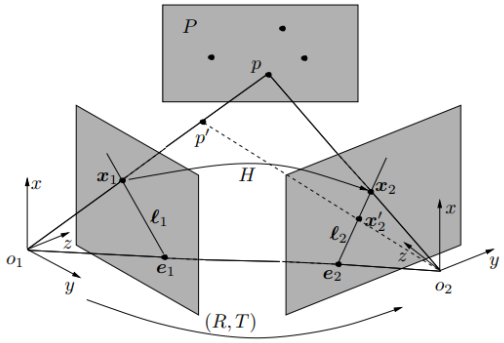

Figure 1: Homography induced by plane

$$X_2 = RX_1 + T \tag{1}$$

where $X_1 \in \mathbb{R}^3$, and $X_2 \in \mathbb{R}^3$ are the coordinates of p relative to camera frames 1 and 2, respectively. As we know the two image points x_1, x_2 of p, in the normalized camera coordinates, satisfy the epipolar constraint:

$$x_2^T E x_1 = 0 \tag{2}$$

Since x_1 and x_2 are represented in homogeneous coordinates, we will also have $X_2^\top E X_1 = 0$. Furthermore, since p lies on a plane P, the corresponding image points will have an additional constraint based on a planar homography of the form $x_2 \sim H x_1$. There exist λ_1 and λ_2 such that $\lambda_1 x_1 = X_1$ and $\lambda_2 x_2 = X_2$, therefore we also have $X_2 = H X_1$. Assume that N is the unit normal vector of the plane P with respect to first camera frame and $d > 0$ is the offset of the plane along the unit vector N.

(a) (10 points) Find H in terms of R, T, N and d.

Solution: The two images x_1, x_2 of p satisfy the epipolar constraint

$$x_2^T E x_1 = 0 \tag{3}$$

However, for points on the same plane P, their images will share an extra constraint that makes the epipolar constraint alone no longer sufficient. Given N is the unit normal vector of the plane P with respect to the first camera frame and d is the distance from the plane P to the optical center of the first camera. Let $N = [n_1, n_2, n_3]^T$. Using plane equation we have

$$N^T X_1 = n_1 X + n_2 Y + n_3 Z = d \tag{4}$$

$$n_1 X + n_2 Y + n_3 Z = d \iff \frac{1}{d} N^T X_1 = 1, \quad \forall X_1 \in P \tag{5}$$

Substituting plane equation 5 into coordinate transformation equation $X_2 = RX_1 + T$, we get

$$X_2 = RX_1 + T \tag{6}$$

$$X_2 = RX_1 + T\frac{1}{d} N^T X_1 \tag{7}$$

$$X_2 = \left(R + T\frac{1}{d} N^T\right) X_1 \tag{8}$$

$$X_2 = HX_1 \tag{9}$$

Where,

$$H \doteq R + T\frac{1}{d} N^T \in \mathbb{R}^{3\times 3} \tag{10}$$

The H matrix depends on the motion parameters R, T as well as the parameters N, d of the plane P.

(b) (10 points) Given a homography H (induced by plane P in 3D) between two images, for any pair of image points (x_1, x_2) corresponding to a 3D point p, *not necessarily* on P. Prove that the associated epipolar lines are:

$$l_2 \sim [x_2]_\times H x_1 \tag{11}$$

$$l_1 \sim H^T l_2 \tag{12}$$

Solution:
Proof of equation ($l_2 \sim [x_2]_\times H x_1$):
Suppose p is on the plane P, let x_1 be its image in the left camera image and its correspponding image in second camera image is x_2. Both x_1 and x_2 satisfy the relation $x_2 \sim Hx_1$, since for any point say x_2' on the same epipolar line $\ell_2 \sim Ex_1 \in \mathbb{R}^3$, the ray $o_2 x_2'$ will intersect the ray $o_1 x_1$ at the point p' out of the plane.

Refering figure 1, if x_1 is the image of some point p', not on the plane P, then $x_2 \sim Hx_1$ is only a point that is on the same epipolar line $\ell_2 \sim Ex_1$ as its actual corresponding image $x_2^{'}$. i.e $\ell_2^T x_2 = \ell_2^T x_2^{'} = 0$.

$$\ell_2 \sim [x_2]_\times x_2^{'} \tag{13}$$

$$x_2^{'} \sim Hx_1 \tag{14}$$

Substituting equation 14 into equation 13 we get the desired result $\ell_2 \sim [x_2]_\times H x_1$.

Proof of equation $\left(\ell_1 = H^T \ell_2\right)$:

Refering figure 1, we have

$$\ell_1^T x_1 = \ell_2^T x_2 = 0 \tag{15}$$

Substituting $x_2 \sim Hx_1$ into equation 15, we get

$$\ell_1^T x_1 \sim \ell_2^T H x_1 \tag{16}$$

The equation 16 holds true iff $\ell_1^T \sim \ell_2^T H \implies \ell_1 \sim H^T \ell_2$. Hence proved.