CS223b Midterm Exam, Computer Vision

Monday February 25th, Winter 2008, Prof. Jana Kosecka

Your name _____

email _____

- This exam is 8 pages long including cover page. Make sure your exam is not missing any pages. The exam has maximum score of 100 points. You have 75 minutes to take the exam.
- The exam is open book, open notes, but no electronic devices that can communicate with the outside world.
- Write your answers in the space provided. If you need extra space, use the back of the preceding sheet.
- Write clearly and be concise.
- SCPD students: If you are taking this exam off campus, you have to fax it (650)-725-1449 exactly 75 minutes after receipt. Alternatively you can e-mail the results to kosecka@ai.stanford.edu.

Questions	Points
1 (15)	
2 (15)	
3 (15)	
4 (20)	
5 (10)	
6 (25)	
total	

1. (15) An important parameter of the imaging system is the *field of view* (FOV). Field of view is twice the angle between the optical axis (z-axis) and the end of the retinal plane (CCD array). Imagine that you have a camera system with focal length 16mm, and retinal plane (CCD array) is $(16mm \times 12mm)$ and that imaging surface is sampled 640×480 pixels in each dimension.

   a) Compute the FOV (horizontal and vertical)

   b) Write down the relationship between the image coordinate and a point in 3D world expressed in the camera coordinate system.

   c) Describe how is the size of FOV related to the focal length and how it affects the resolution in the image.

   d) Given the horizontal FOV you computed, how many images do you need to create 360 degree panorama, assuming that you will need 50% overlap between neighboring views.

   Solution

   - vertical (portrait) FOV $= 2 \times \tan^{-1}(0.5 \times 12/f)$
   - horizontal (landscape) HFOV $= 2 \times \tan^{-1}(0.5 \times 16/f)$
   - the larger is the focal length, the smaller is the FOV. As the FOV increases and the number of image pixels is fixed , the resolution of the image is decreased, i.e. the visual angle subtended by a single pixel is larger.
   - number of images to take $= 360/(0.5 \times HFOV)$

2. (15) Show the resulting image obtained after convolution of the original image with the following approximation of the derivative filter $[-1, 0, 1]$ in the horizontal direction.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	1^*	1	1	1	0	0	0
0	0	0	1	1	1	1	0	0	0
0	0	0	1^*	1	1	1^*	0	0	0
0	0	0	1	1	1	1	0	0	0
0	0	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	−1	−1	0	0
0	0	1	1	0	0	−1	−1	0	0
0	0	1	1	0	0	−1	−1	0	0
0	0	1	1	0	0	−1	−1	0	0
0	0	1	1	0	0	−1	−1	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

(a) Compute gradient magnitude at pixels (3,4), (5,4) and (5,7) (marked with * in the image).

(b) Compute gradient direction at those same points.

(c) Describe in words what does the non-maximum suppression step on the gradient magnitude in edge detection process accomplish ?

Solution:

(a) Gradient magnitude is $M = \sqrt{I_x^2 + I_y^2}$ where $I_x = \frac{\partial I_x}{\partial x}$ and $I_y = \frac{\partial I_y}{\partial y}$ at pixels (3,4), (5,4) and (5,7) (marked with * in the image). First we need to compute at those pixels I_y which is the result of convolution with the derivative filter in vertical direction. At (3,4) $[I_x, I_y] = [1, 1]$ and $M = \sqrt{2}$ and $\theta = \tan^{-1}(1) = 45, \circ$, at (5,4) $[I_x, I_y] = [1, 0]$ and $M = 1$ and $\theta = \tan^{-1}(0) = 0^\circ$ at (5,7) $[I_x, I_y] = [-1, 0]$ and $M = 1$ and $\theta = \tan^{-1}(0) = 0^\circ$.

(b) Non-maximum supression of the gradient magnitude creates a single pixel wide edges, by supressing the non-maximum magnitudes which are in the gradient direction neighborhood of each pixel.

Note: If one follows directly the formula for convolution $g[x] = \sum_{k=-\infty}^{k=\infty} f[k]h[x-k]$ the final result would be correspond to a shift and point-wise multiplication with a flipped (mirror) version of the filter.

3. (15) **Motion recovery.** Consider a set of corresponding points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in retinal coordinates in two views, which are related by pure translation T.

   (a) Write down a simplified version of the epipolar constraint in case the motion is pure translation.

   (b) Describe a linear least squares algorithm for estimation of translation T. What is the minimal number of corresponding points needed to solve for T ?

   (c) i) Suppose now that the camera is not calibrated and you can measure only pixel coordinates points $\boldsymbol{x}'_1$ and $\boldsymbol{x}'_2$. Can you still recover the translation between the two views ?

   ii) Suppose that you know all intrinsic camera parameters expect the focal length. Can you now recover the translation between the two views ?

Solution General epipolar constraint has the following form

$$\boldsymbol{x}_2^T \widehat{T} R \boldsymbol{x}_1 = \boldsymbol{x}_2^T E \boldsymbol{x}_1 = 0, \tag{1}$$

where $E = \widehat{T}R$ is the essential matrix. In case of translational motion only, the rotation is $R = I$ and the epipolar constraint becomes $\boldsymbol{x}_2^T \widehat{T} \boldsymbol{x}_1 = 0$

$$E = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{2}$$

Single correspondence gives rise to the following constraint on translation vector

$$[y_2 + y_1, x_2 - x_1, -x_2 y_1 + y_2 x_1]^T \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} = 0$$

Single correspondence gives us one constraint. We have three unknowns, but since the system is homogeneous and the translation can be recovered only up to scale, only two correspondences are needed. Given at least two point correspondences, the elements of the essential matrix $[t_x, t_y, t_z]^T$ can be obtained as a least squares solution of a system of homogeneous equations. Once the essential matrix T has been recovered, we have two possible solution $T = \pm[t_x, t_y, t_z]$. The physically correct solution is then obtained using the positive depth constraint.

In uncalibrated case we can only recover $T' = KT$, so true translation cannot be recovered. Even when all unknows are known except the focal length the translation still cannot be recovered.

4. (20) A camera in a low flying plane captures an image of a 100x100m planar wheat field. The four corners of the field have the following image coordinates (in pixels) (-30,-50), (120,-30), (100,50), (-30,60), where the origin of the image coordinate system is assumed to be in the center of the image. You can assume that field is perfectly planar and that you know the focal length of the camera. How would you compute the relative orientation of the camera and the field at the instance the image was captured ? What can you say about the distance between the camera and the plane ? Write down the geometry of the problem and describe the individual steps of the algorithm (you don't have to compute the actual values).

Solution The relationship between the planar square field and it's image is captured by a planar homography between the world and the image plane, which in calibrated setting has the following form

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = H\boldsymbol{X} = \begin{bmatrix} r_{11} & r_{12} & t_x \\ r_{21} & r_{22} & t_y \\ r_{31} & r_{32} & t_z \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

Given at least 4 points we can recover H up to scale, by solving system of linear equations. Since we know 3D coordinates of 4 points in the world and also their projections, we have all the information. Each point gives following two constraints on the entries of the homography:

$$x = \frac{h_{11}X + h_{12}Y + h_{13}}{h_{31}X + h_{32}Y + h_{33}} \quad y = \frac{h_{21}X + h_{22}Y + h_{23}}{h_{31}X + h_{32}Y + h_{33}}$$

In case the focal lenght of the camera is uknown the homography has the following form:

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = \begin{bmatrix} fr_{11} & fr_{12} & ft_x \\ fr_{21} & fr_{22} & ft_y \\ r_{31} & r_{32} & t_z \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

The focal length can be estimated by using the constraints on columns of the rotation matrix $r_1^T r_2 = 0$.

$$f = \sqrt{\frac{h_{11}h_{12} + h_{21}h_{22}}{-h_{31}h_{32}}}$$

Once f is known then we can estimate the correct rotation and translation. Using the constraints that $\|r_1\| = \|r_2\| = 1$ we can recover the scale of the homography. Hence the rotation and translation can be recovered even in the absence of the camera focal length.

5. (10) Suppose you observe 3 vanishing points in the image and assume that you know all the intrinsic parameters except the focal length. How would you determine the rotation of the camera with respect to the origin of the world coordinate frame ? Write down the geometry of the problem and describe the individual steps of the algorithm.

   Solution: Assuming that the parallel line come from 34 orthogonal directions, the coordinates of the vanishing points corresponding to the three sets of parallel lines are respectively

   $$v_1 = KRe_1 \;\; v_2 = KRe_2 \quad v_3 = KRe_3$$

   Using the orthogonality constraints between the lines we can get constraints on the entries of the matrix K. Namely

   $$e_i^T e_j = v_i^T K^{-T} K^{-1} v_j = v_i^T S v_j = 0$$

   The unknown matrix S in this case has the form

   $$S = \begin{bmatrix} 1/f^2 & 0 & 0 \\ 0 & 1/f^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

   The focal length can then be recovered from a single constraint.

6. (25) Provide short answers to the following questions ?

   - (3) What is the aperture problem ? If one considers local window of a small aperture around a point and that window has gradient variation only in single direction, then one cannot recover the full motion. Only the component perpendicular to the edge direction can be recovered.

   - (3) What does it mean for the 2D convolution kernel (filter) to be separable ? 2D convolution kernel is separable, if it can be obtained as an outer product of two 2D kernels.

   - (3) When is a purely translational model suitable for image stitching ? The purely translational model is suitable if the camera is viewing a fronto-parallel plane or if the 3D structure is too far away with little depth variation compared to the distance to the camera (as in aerial photography)

   - (3) Given two lines in the image denoted by their projective coordinates l_1 and l_2, how would you compute an intersection point of these two lines ? $\boldsymbol{x} = \boldsymbol{l}_1 \times \boldsymbol{l}_2$

   - (3) The rank of the multiview measurement (point correspondences) matrix M of dimensions (2m x n) under orthographic projection is 3. Explain why ? The measurement matrix M is a product of the motion and shape matrix $M = RS$, where

- (3) How is the rotational invariance enforced in SIFT features ? By finding dominant orientation and rotating the window to that canonical orientation.

- (3) What is an epipole ? Epipole is a point where the line connecting two centers of projection (i.e. the translation vector) intersects the image.

- (1) Does cylindrical projection preserve straight lines NO

- (1) Image thresholding is not a linear operation YES

- (1) Fundamental matrix maps points to points in respective views NO

- (1) Essential matrix is not invertible YES