SGN-45006 Fundamentals of Robot Vision

Lecturer Esa Rahtu

The maximum number of points for each task is shown in parenthesis. The use of calculator is allowed, but it is not necessary.

1. Explain briefly the following terms and concepts:

   (a) Camera projection matrix (2 p)
   (b) Scale Invariant Feature Transform (SIFT) (2 p)
   (c) Camera calibration (2 p)
   (d) Structure from motion (2 p)
   (e) Convolutional neural network (2 p)
   (f) Boosting sceme in classification (2 p)

2. Model fitting using RANSAC algorithm

   (a) Describe the main stages of the RANSAC algorithm in the general case. (2 p)
   (b) In this context, why it is usually beneficial to sample *minimal subsets* of data points instead of using more data points? (Minimal subsets have the minimal number of data points required for fitting.) (1 p)
   (c) Mention at least two examples of models that can be fitted using RANSAC. Describe how the models are used in computer vision and what is the size of the minimal subset of data points required for fitting in each case. (1 p)
   (d) Describe how RANSAC can be used for panoramic image stitching. Why is RANSAC needed and what is the model fitted in this case? (2 p)

3. Geometric 2D transformations

   (a) Using homogeneous coordinates, write the matrix form of the following 2D transformations: translation, similarity (rotation+scaling+translation), affine and homography. How many degrees of freedom does each transformation have and why? How many point correspondences are needed to estimate each? (3 p)
   (b) A rectangle with corners $A = (-1, 1)$, $B = (1, 1)$, $C = (1, -1)$, $D = (-1, -1)$ is transformed by a transformation so that the new corners are $A' = (1, 3)$, $B' = (3, 3)$, $C' = (-2, 1)$, $D' = (-6, 1)$, respectively. An affine transformation does not explain the observations perfectly, but there is reason to believe that the transformation is affine and there is noise in the observations. Write down the equations to solve the transformation using the least squares method. **Note:** You don't actually have to solve the transformation. (3 p)

4. Neural networks

   (a) What is a Perceptron? Explain the construction (hint: use picture) and how it can be trained to perform classification task (assume you have training samples with input feature vector x and class label 1 or -1). (2 p)

(b) What are the main differences between convolutional neural networks (CNNs) and conventional fully connected networks? (1 p)

(c) Give a rough example of typically used structures in CNN based models. Illustrate your example with a picture. (1 p)

(d) Explain what are *feature maps* in the context of CNNs. (1 p)

(e) Explain the main differences between one and two stage object detection methods? Give example of both types. (1 p)

5. Epipolar geometry and stereo

(a) Figure 1 presents a stereo system with two parallel pinhole cameras separated by a baseline b so that the centers of the cameras are $\mathbf{c}_l = (0,0,0)$ and $\mathbf{c}_r = (b,0,0)$. Both cameras have the same focal length f. The point P is located in front of the cameras and its disparity d is the distance between corresponding image points, i.e., $d = |x_l - x_r|$. Assume that $d = 1$ cm, $b = 6$ cm and $f = 1$ cm. Compute Z_P. (2 p)

(b) Let's denote the camera projection matrices of two cameras by $\mathbf{P}_1 = [\mathbf{I} \quad \mathbf{0}]$ and $\mathbf{P}_1' = [\mathbf{R} \quad \mathbf{t}]$, where $\mathbf{R}$ is a rotation matrix and $\mathbf{t} = (t_1, t_2, t_3)^T$ describes the translation between the cameras. Show that the epipolar constraint for corresponding image points $\mathbf{x}$ and $\mathbf{x}'$ can be written in the form $\mathbf{x}'\mathbf{E}^T\mathbf{x} = 0$, where matrix $\mathbf{E}$ is the essential matrix $\mathbf{E} = [\mathbf{t}]_X\mathbf{R}$. (2 p)

(c) In the configuration illustrated in Figure 1 the camera matrices are $\mathbf{P}_1 = [\mathbf{I} \quad \mathbf{0}]$ and $\mathbf{P}_1' = [\mathbf{I} \quad \mathbf{t}]$, where $\mathbf{I}$ is the identity matrix and $\mathbf{t} = (-6,0,0)^T$. The point Q has coordinates $(3,0,3)$. Compute the image of Q on the image plane of the camera on the left and the corresponding epipolar line on the image plane of the camera on the right. (Hint: The epipolar line is computed using the essential matrix.) (2 p)

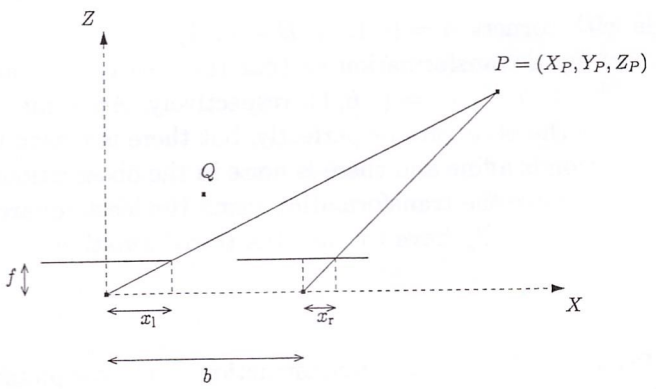

Figure 1: Top view of a stereo pair where two pinhole cameras are placed side by side.