CS-E4850 Computer Vision

Exam 18th of February 2020, Lecturer: Juho Kannala

There are plenty of questions, answer as many as you can in the available time. The number of points awarded from different parts is shown in parenthesis in the end of each question. The maximum score from the whole exam is 42 points.

You need pen and paper, also calculator is allowed but should not be necessary.

1. Explain the following concepts (e.g. what does the concept mean, what are its key properties, and how it is utilised in computer vision):

   (a) Separable filter (2 p)

   (b) Harris corner detector (2 p)

   (c) RANSAC algorithm (2 p)

   (d) Kanade-Lucas-Tomasi (KLT) feature tracker (2 p)

   (e) Camera calibration (2 p)

   (f) Bundle adjustment (2 p)

2. Local feature detection and description using SIFT

   (a) Describe the detector part of the Scale Invariant Feature Transform (SIFT). In particular, explain the motivation and idea of the scale selection. (2 p)

   (b) Describe the descriptor part of SIFT. That is, describe how the pixel neighborhood around a detected keypoint is converted to a 128 dimensional feature vector. (2 p)

   (c) Mention at least two computer vision tasks or applications where SIFT is commonly used. Explain also what is the benefit of using SIFT in the applications (e.g. when compared to earlier methods which are not scale invariant). (2 p)

3. Large-scale object instance recognition

   (a) Describe the bag-of-visual-words image representation technique and its pros and cons for object instance recognition. (2 p)

   (b) Describe what is *inverted index* and how it can be used to improve efficiency of object instance recognition from large image databases? (1 p)

   (c) Explain the concept *term frequency - inverse document frequency* (tf-idf) weighting and its purpose. (1 p)

   (d) Explain what is the precision-recall curve (that is often used for evaluating image retrieval systems). Compute precision and recall in the following case: We search for car images from a database of 10000 images. It is known that there are 600 car images in the database. An automatic image retrieval system retrieves 200 car images and 50 other images from the database. (2 p)

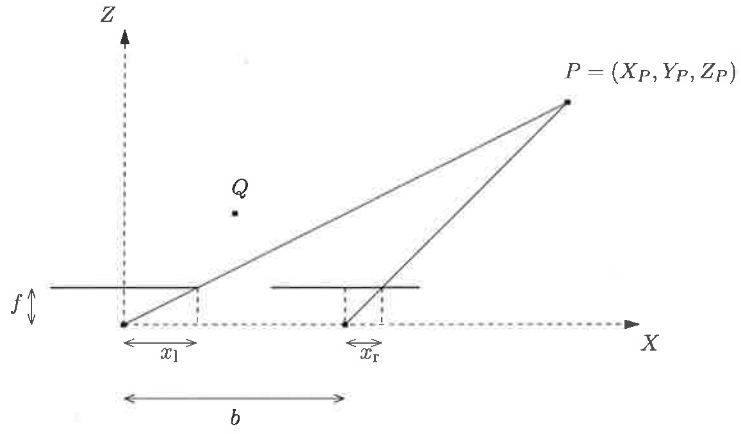

Figure 1: Top view of a stereo pair where two pinhole cameras are placed side by side.

4. 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 $\mathbf{P} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}$ and $\mathbf{P}' = \begin{bmatrix}\mathbf{R} & \mathbf{t}\end{bmatrix}$, where $\mathbf{R}$ is a rotation matrix and $\mathbf{t} = (t_1, t_2, t_3)^\top$ 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 $\mathbf{x}'^\top \mathbf{E}\mathbf{x} = 0$, where matrix $\mathbf{E}$ is the essential matrix $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$. (2 p)

   (c) In the configuration illustrated in Figure 1 the camera matrices are $\mathbf{P} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}$ and $\mathbf{P}' = \begin{bmatrix}\mathbf{I} & \mathbf{t}\end{bmatrix}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{t} = (-6, 0, 0)^\top$. 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)

5. 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? How many point correspondences are needed to estimate each? (3 p)

   (b) A similarity transformation consists of rotation, scaling and translation and is defined using conventional Cartesian coordinates as follows:

$$\mathbf{x}' = s\mathbf{R}\mathbf{x} + \mathbf{t} \quad \Leftrightarrow \quad \begin{pmatrix} x' \\ y' \end{pmatrix} = s \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

   Describe a procedure for solving the parameters s, θ, t_x, t_y of a similarity transformation from two point correspondences $\{\mathbf{x}_1 \rightarrow \mathbf{x}'_1\}$, $\{\mathbf{x}_2 \rightarrow \mathbf{x}'_2\}$. Use the

procedure to compute the transformation from the following point correspondences: $\{(\frac{1}{2},0) \rightarrow (0,0)\}$, $\{(0,\frac{1}{2}) \rightarrow (-1,-1)\}$. (3 p)
(Hint: Drawing the point correspondences on a grid paper may help you to check your answer.)

6. Neural networks

   (a) Explain how neural networks are typically used in image classification? What kind of neural networks are popular in this context and why? (2 p)

   (b) Explain the basic concepts of the backpropagation algorithm. (What it does? How it works? When it can be used? Why it may sometimes fail?) (2 p)

   (c) Explain the object detection problem and the basic principles of the single-shot multibox detector (SSD). In addition, describe the loss function and other key concepts of SSD training. (2 p)