CS-E4850 Computer Vision

Exam 4th of April 2017, 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 will need pen and paper, and also calculator is allowed but is not necessary.

1. Explain briefly the following terms and concepts:

   (a) Camera projection matrix (2 p)

   (b) RANSAC algorithm (2 p)

   (c) Hough transform (2 p)

   (d) Optical flow (2 p)

   (e) Epipolar line (2 p)

   (f) Object detection by sliding windows (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 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) Describe what is the role of *spatial verification* in object instance recognition and how it is usually performed? (2 p)

4. Convolutional neural networks (CNNs)

   (a) What are the main differences between convolutional neural networks and conventional fully connected networks? (2 p)

   (b) Describe at least two computer vision problems in which CNNs have been shown to perform well and explain why CNNs are particularly suitable for such problems. (2 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)

5. Triangulation

Two cameras are looking at the same scene. The projection matrices of the two cameras are $\mathbf{P}_1$ and $\mathbf{P}_2$. They see the same 3D point $\mathbf{X} = (X, Y, Z)^\top$. The observed coordinates for the projections of point $\mathbf{X}$ are $\mathbf{x}_1$ and $\mathbf{x}_2$ in the two images, respectively. The numerical values are as follows:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} \frac{3}{4} \\ 0 \end{bmatrix}.$$

(a) Compute the 3D coordinates of the point $\mathbf{X}$. (Hint: Perhaps the simplest way in this case is to write the projection equations in homogeneous coordinates by explicitly writing out the unknown scale factors, and to solve X, Y, Z and the scale factors directly from those equations.) (1 p)

(b) Present a derivation for the linear triangulation method and explain how $\mathbf{X}$ can be solved using that approach in the general case (i.e. no need to compute with numbers in this subtask). (2 p)

(c) A third camera $\mathbf{P}_3$ is added to the scene. Describe how the linear triangulation method above can be extended to use the information from all the three cameras. (1 p)

(d) If there is noise (i.e. measurement errors) in the observed image coordinates of point $\mathbf{X}$, the linear triangulation method above is not the optimal choice but a nonlinear approach can be used instead. What error function is typically minimized in the nonlinear approach? (1 p)

(e) How does the nonlinear triangulation approach differ from the bundle adjustment procedure which is commonly used in structure-from-motion problems (i.e. how is the bundle adjustment problem different)? (1 p)

6. Geometric transformations

A rectangle with corners $A = (0, 2), B = (3, 2), C = (3, 0), D = (0, 0)$ is transformed by a transformation T so that the new corners are $A' = (-9, 0)$, $B' = (-9, 6)$, $C' = (-1, 6)$, $D' = (-1, 0)$.

(a) Assume T is a similarity transformation (i.e. rotation, scaling, translation). Solve the transformation using the corresponding corners (A, A') and (B, B'). (Hint: Drawing a picture may be helpful for checking the result.) (2 p)

(b) Describe the least-squares method for solving an affine transformation from n pairs of corresponding points. (2 p)

(c) Assume that T is an affine transformation. Solve the transformation using all four corners of the rectangles. (Hint: You may also use the picture to infer or check the result.) (1 p)

(d) How many point correspondences are sufficient for determining an affine transformation in the general case? Justify your answer. (1 p)