DATA.ML.300 Computer Vision

This exam contains two parts: 1) questions related to each week exam, and 2) traditional final exam. Please return the answers only to ONE of these two options. For option 1) see further instructions below.

By answering the questions related to a particular week exam, you can improve the corresponding week exam points (you can answer even if you have not participated to the particular week exam before). Note that the points for each week exam are summed and the sum is compared to your current points from the corresponding week exam. The maximum of these will be used as your points for that week.

You can improve as many week exams as you like (and have time for). Please write your answers to a separate sheet and note clearly which week exam and task they correspond to. Calculator is allowed.

Week exam 1

1. General questions

(a) What is a Gaussian filter and where it can be applied?
(b) What is the benefit of using homogenous coordinates in the case of pinhole camera model?
(c) How Fourier transform could be used in calculating the linear filtering result?
(d) What is a Gaussian image pyramid?

2. Transformations

(a) A perspective camera has the following camera matrix:

$$P = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Determine the image points corresponding to 3D point X = (6,2,2). Report your answer in non-homogenous coordinates.

(b) Write the matrix equations for 3D similarity, affine, and perspective transformations. Use homogenous coordinates. How many degrees of freedom each transform has and how many point correspondences are needed to estimate them?

3. Homogenous coordinates

(a) Convert the following (in-homogenous) points into homogenous coordinates (1, 5), (100, 500), and (4, 4, 1). Similarly, convert the following homogenous points into corresponding in-homogenous form (i.e. to normal coordinates) (1,5,1), (7,1,3), (24,12,6) and (8,6,1,2). What does a homogenous point (1, 1, 0) correspond to?

(b) A line ax + by + c = 0 can be presented in a vector form as $l=(a,b,c)^T$ and, using homogenous coordinates, a point x is on the line l if $x^Tl = 0$. The intersection of two lines l and l' is given by the vector cross product between l and l'. Similarly, the line l passing though points x and x' is given by the vector cross product between x and x'. Use homogenous coordinates and above formulas to determine the intersection of lines l1 and l2. The l1 runs through points (2,4) and (8,8), and l2 runs through points (14,10) and (18,6).

Hint: The 3D vector cross product is calculated as:

$$\begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \times \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} a_y b_z - b_y a_z \\ a_z b_x - b_z a_x \\ a_x b_y - b_x a_y \end{pmatrix}$$

Week exam 2

1. General questions

(a) What is the main goal in image retrieval task?
(b) What do hyperparemeters mean in image classification (give one example)?
(c) What is k-nearest neighbour classifier? What are its pros and cons?
(d) Give one example how pertained classification network can be used in image retrieval?

2. 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).

(b) In Figure 1 below you see a very small neural network, which has one input unit, one hidden unit (logistic), and one output unit (linear). The nonlinear function σ in the logistic unit is defined by the formula $\sigma(z) = 1/(1 + e^{-z})$. Let's consider one training case. For that training case, the input value is 1 (as shown in the figure) and the target output value t is 2. We are using the standard squared loss function: $E = (t - y)^2/2$, where y is the output of the network. The values of the weights and biases are shown in the figure and they have been constructed in such a way that you don't need a calculator. Hint: the derivative of logistic function is defined as $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$. Answer the following questions:

i. What is the output of the hidden unit and the output unit, for this training case?
ii. What is the loss, for this training case?
iii. What is the derivative of the loss with respect to w2, for this training case?

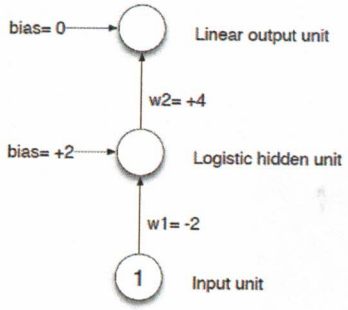


Figure 1: A simple neural network

3. Image retrieval

(a) Describe the bag-of-visual-words image representation technique. How it can be utilised in image retrieval?

(b) Figure 3 (below) illustrates a database of four images and corresponding visual words for each image (W1, W2, ...). Construct an inverted index for this example dataset.

(c) We have a database of 10 images. Our retrieval algorithm has ranked them in the following order with respect to a given query (see query and ranked database in Figure 2 below). Based on the manual annotations, we know that the images with green box are relevant to the current query. Draw a precision-recall curve for the retrieval result (use the axis given in Fig 2).

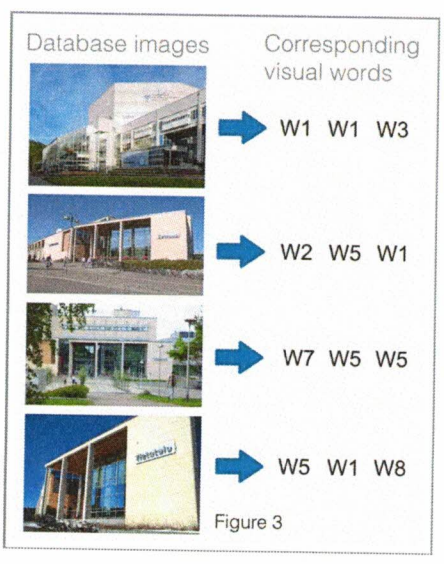


Figure 3

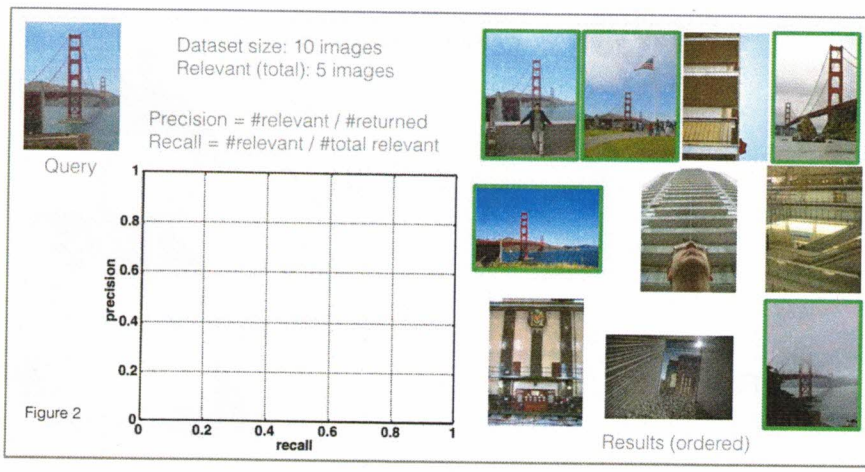


Figure 2

Week exam 3

1. General questions

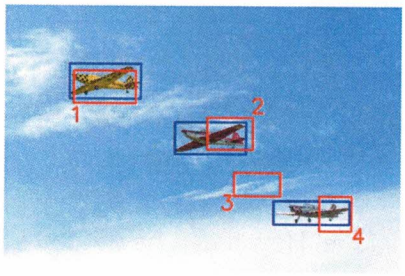
(a) What is the goal in object category detection and how it differs from image classification and object segmentation?
(b) Name the main components in the sliding window based object detector.
(c) What is bootstrapping and how it can be used in training detectors?
(d) What is the difference between one and two stage CNN object detectors?

2. Classical object detectors

(a) Describe different phases in extracting Histogram of Oriented Gradients (HoG). Use picture.

(b) The following image (Figure 1) depicts an example detection result. The blue boxes are the know ground truth locations of the objects and the red boxes are the obtained detection. The number next to each detection denotes the corresponding ranking (i.e. detection 1 has the highest classification score, detection 2 next highest, and so on.). The corresponding intersection over union values are: 1) 0.9, 2) 0.57, 3) 0, and 4) 0.49 (i.e. the IoU measure for each detection with respect to the highest overlapping ground truth). Draw the corresponding precision-recall curve using 0.5 IoU value as a detection threshold.

Hint:
Precision = #returned correct detections / #returned detections
Recall = #detected objects / #total number of objects



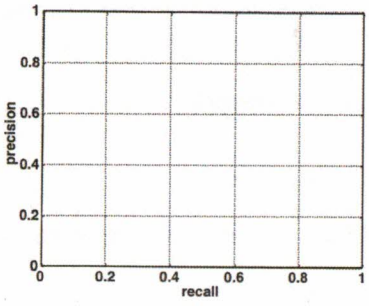


Figure 1

3. CNN based detectors

(a) Explain the main phases in the "CornerNet" object detection approach.

(b) The following image (Figure 2) depicts the Faster-RCNN object detector. Shortly describe the objective of each component in the system (i.e. what it takes in and what it aims to produce as an output).

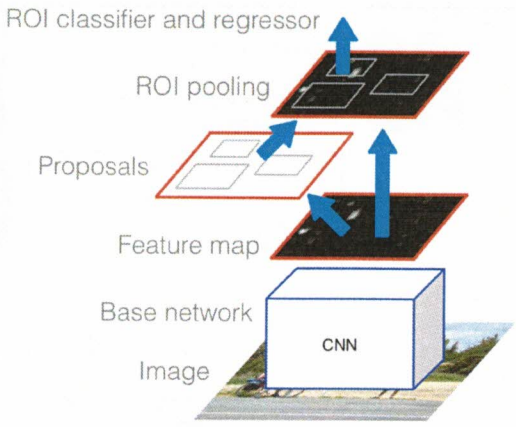


Figure 2

Week exam 4

1. General questions

(a) What are the main stages in Canny edge detector?
(b) Outline the cost function that is minimised when fitting a line with least squares method. (no need to solve it)
(c) Why it is usually beneficial to sample *minimal subset* of data points in RANSAC instead of using more data points?
(d) What is the main motivation in using "robust cost functions" in model fitting instead of normal quadratic function used in vanilla Least Square fitting?

2. Local features

(a) Figure 1 illustrates three different kinds of local image areas (the box). For each case, explain if it makes a good local keypoint or not. Justify your answer. (local key point = an image point that can be accurately and reliably detected from multiple images from the same scene).

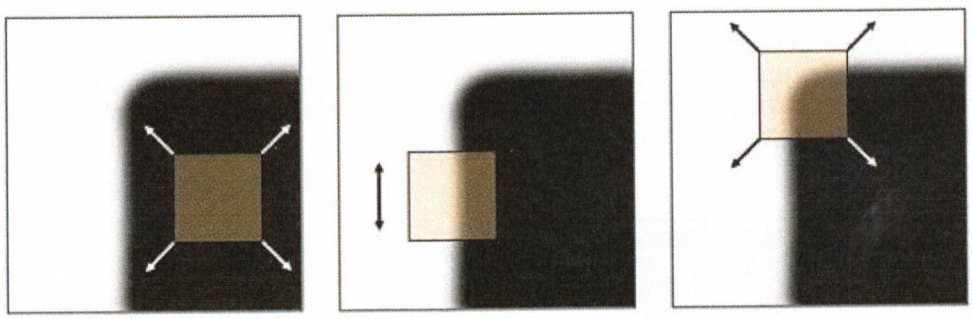


Figure 1

(b) Describe how scale normalised Laplacian of Gaussian function (see figure 2) can be used in scale covariant blob detection.

$$\nabla^2_{\text{norm}} g = \sigma^2 \left(\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} \right)$$

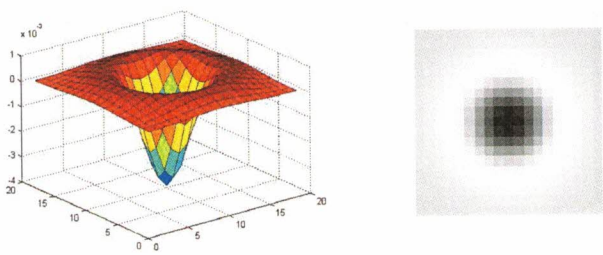


Figure 2. Scale normalised Laplacian of Gaussian

3. Robust model fitting

(a) Describe the main stages of the RANSAC algorithm in the general case.

(b) What is the idea in Hough transform and how it can be used in model fitting? Give one example.

Week exam 5

1. General questions

(a) What is the brightness constraint in optical flow estimation?
(b) What is so called aperture problem?
(c) What are motion field and optical flow? What is the main difference?
(d) What kind of features are good for tracking and why?

2. 2D transformations

(a) Figure 1 depicts two images of database objects and a scene where they need to be detected. Describe the main steps how this kind of object instance recognition task can be solved using local features and image alignment. For each step, explain the main goal and name at least one method to implement this.

Figure 1

(b) Figure 2 depicts two images taken from the same scene. Describe the main steps how these images can be aligned to form a panorama image shown in Figure 3. For each step, explain the main goal and name at least one method to implement this.

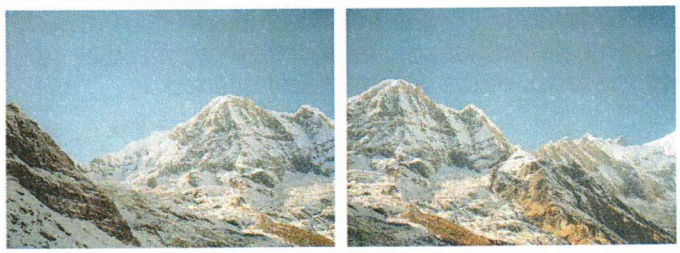


Figure 1: image pair from the same scene

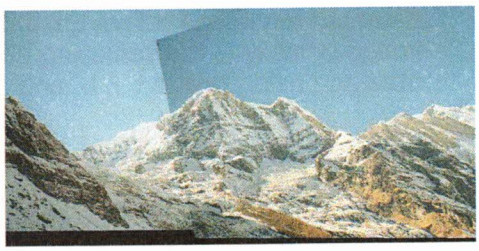


Figure 2: panorama image

3. Optical flow and tracking

(a) Assume we have two frames obtained at time instants (t-1) and t as shown in Figure 3. In optical flow, our target is to estimate the motion (u,v) of a pixel at position (x,y). Starting from the brightness constraint, derive the optical flow equation:

$$\nabla I \cdot (u,v) + I_t = 0$$

How many unknown this equation has per pixel? Hint:

$$I(x+u(x,y), y+v(x,y), t) \approx I(x,y,t) + I_x\, u(x,y) + I_y\, v(x,y)$$

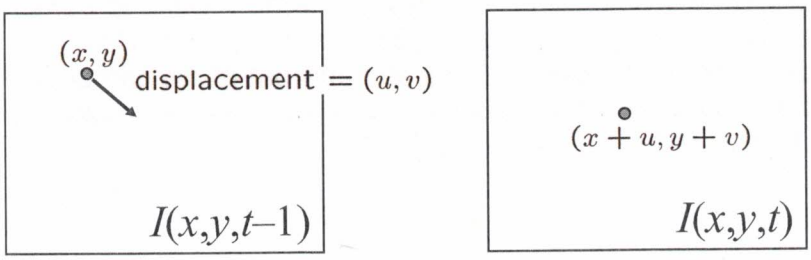


Figure 3. $I(x,y,t)$ denotes the brightness of a pixel at position (x,y) at time instant t.

(b) Explain the multi-resolution approach for optical flow estimation. What are the main advantages of the approach?

Week exam 6

1. General questions

(a) Why the recovery of the scene structure from a single image is an ill-posed problem?
(b) What does auto calibration mean in the context of camera calibration?
(c) What is the relation between depth and disparity in stereo vision?
(d) What is the main difference between essential and fundamental matrices?

2. Camera calibration and single view metrology

(a) Briefly explain the "linear method" for camera calibration? What are the pros and cons of this approach?

(b) Figure 2 illustrates a scenario, where we are trying to estimate the height H (distance between top T and bottom B) from a single image using the known reference height R. We have detected the image points t, r, and b that correspond to 3D points T, R, B, respectively. The image point v_Z is the vanishing point in the vertical direction. Show how the height H can be obtained using points t, r, b, and . Hint: use the cross ration of four points defined as

$$\frac{\|\mathbf{P}_3 - \mathbf{P}_1\| \, \|\mathbf{P}_4 - \mathbf{P}_2\|}{\|\mathbf{P}_3 - \mathbf{P}_2\| \, \|\mathbf{P}_4 - \mathbf{P}_1\|}$$

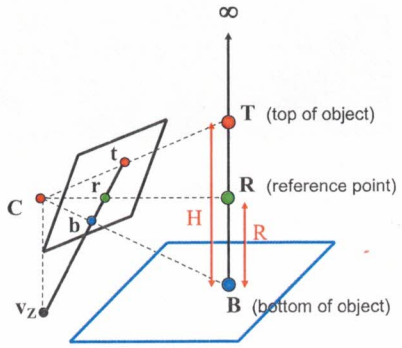


Figure 1: Estimating the height H from a single image using reference height R.

3. Epipolar geometry and stereo

(a) What is epipolar line and how it relates observations in two cameras? How Essential and Fundamental matrices are related to these?

(b) Figure 2 presents a stereo system with two parallel pinhole cameras separated by a baseline b so that the centers of the cameras are c_l = (0,0,0) and 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 = 4 cm, b = 12 cm, and f = 2 cm. Compute Z_P.

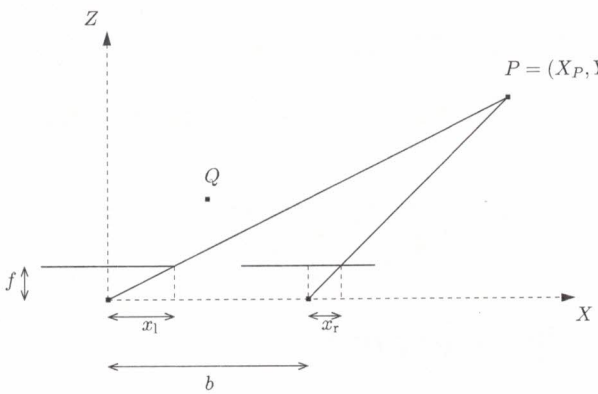


Figure 2: Top view of a stereo pair where two pinhole cameras are placed side by side.

Week exam 7

1. General questions

(a) What is the projective ambiguity in the context of Structure from Motion.
(b) What are the main differences between multi-view stereo and Structure from Motion?
(c) What is bundle adjustment and why it is important in Structure from Motion?
(d) What is inverse depth and why it is used in some multi-view stereo applications?

2. Structure from Motion

(a) We know that all images used in Structure from Motion (SfM) are captured by a single moving camera. How this information can be used to "upgrade" projective SfM solution? Give rough idea how can be done (no need to solve).

(b) You are given m images of n fixed 3D points

I.e. you have m cameras, n 3D points, and each point is detected in every camera (see illustration in Figure 1). The task is to estimate m projection matrices P_i and n 3D points X_j from mn correspondences x_{ij} (up to projective transformation). Explain the main steps in solving this problem in sequential manner.

$$\lambda_{ij}\mathbf{x}_{ij} = \mathbf{P}_i\mathbf{X}_j, \quad i = 1,\dots,m, \quad j = 1,\dots,n$$

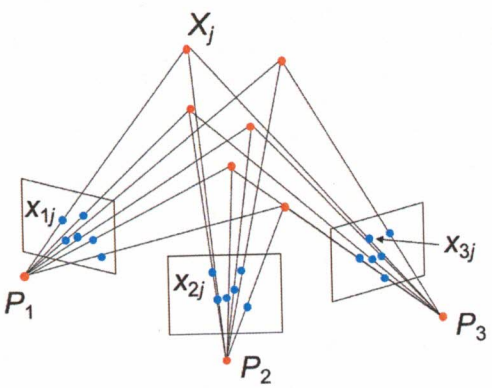


Figure 1: Illustration of the setup in Structure from Motion problem in the case of 3 cameras.

3. Multi-view geometry

(a) Explain the main principle in epipolar geometry based multi-view stereo reconstruction approach.

(b) What is space carving method and how it works in multi-view stereo reconstruction?

DATA.ML.300 Computer Vision

Final exam

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 (1 p)

   (b) Scale Invariant Feature Transform (SIFT) (1 p)

   (c) Camera calibration (1 p)

   (d) Structure from motion (1 p)

   (e) Essential matrix (1 p)

   (f) Hough transform (1 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. Feature tracking

   (a) Describe the main elements of the Shi-Tomasi feature tracker (i.e. how the features are selected and tracked, and tracks terminated or added). (2 p)

   (b) Describe the Lucas-Kanade method for estimating the displacement of an image patch. What kind of equations need to be solved and what is the brightness constancy constraint in this context? (2 p)

   (c) What are the benefits of Lucas-Kanade method when compared to simple template matching? (2 p)

4. Neural networks

   (a) What is a Perceptron? Explain the construction and give a rough idea how it can be trained (hint: use picture). (2 p)

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

   (c) Explain how neural networks are typically used in image classification? What kind of neural networks are popular in this context and why? (1 p)

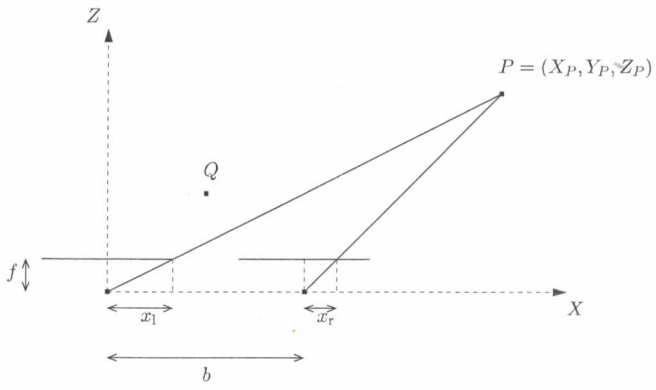


Figure 1: Top view of a stereo pair where two pinhole cameras are placed side by side.

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

5. Triangularisation 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) 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:

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

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)

(c) 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)

(d) How does the 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)