Version IB/4

EGT3
ENGINEERING TRIPOS PART IIB

Wednesday 4 May 2022 2.00 to 3.40

Module 4F12

COMPUTER VISION

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 (a) Consider the task of interpreting a 2-D greyscale image, $I(x, y)$. What is the central motivation for extracting *generic salient features* from a 2-D image, $I(x, y)$, rather than working directly with raw pixel values? What desirable properties do we seek for such features? [10%]

(b) Suppose we wish to extract edges from image, $I(x, y)$, which we assume has a size of 320×240 pixels.

(i) The process of extracting edges directly from raw pixel data may be affected by high-frequency noise, which is amplified by differentiation. Describe a convolution operation to address this issue. Explain how the operation can be implemented efficiently for 2-D images. Identify the total number of multiplications that can be saved over a naive implementation when the filter size is 5×5 and the convolution is "valid" (i.e. no zero padding is applied to image, $I(x, y)$). [15%]

(ii) By using a Taylor series expansion, or otherwise, derive a 3×3 discrete filter which can be used to approximate the 2-D Laplacian. How can this operation be implemented instead with 1-D filters? Assuming that the 1-D filters are to be applied sequentially to the image, which approach (applying a single 2-D filter or two 1-D filters one after the other) is more efficient with respect to multiplications performed and with respect to memory used? [15%]

(iii) Give details of an edge detection algorithm that recovers both the orientation and position of edges in the image, $I(x, y)$. How are the operations of computing gradients and removing high-frequency noise performed efficiently? [15%]

(c) Consider the task of finding correspondences between multiple images that capture different views of an object.

(i) Describe one limitation of using edge detections as *keypoints* for finding correspondences that would be resolved by using *corners*. Describe a second limitation of using edge detections as keypoints that would be resolved by using *blobs*. [10%]

(ii) The Scale-Invariant Feature Transform (SIFT) has been widely adopted for matching tasks. Describe in detail how different components of the SIFT keypoint detection and description stages contribute to achieving desirable invariances for finding correspondences. In which scenarios does SIFT typically fail? [20%]

(iii) Suppose two images capture a building that forms a tourist attraction in summer sunshine and heavy winter snow, respectively. Give four detailed examples of nuisance factors for finding correspondences across seasons that differ from those that arise among two images taken at similar times on the same day. [15%]

2 The relationship between a 3-D world point, (X, Y, Z), and its corresponding pixel at image co-ordinates, (u, v), under perspective projection can be written using *homogeneous* co-ordinates by a 3×4 *projection matrix*, $\mathbf{P}$:

$$\tilde{\mathbf{x}} = \mathbf{P}\tilde{\mathbf{X}}.$$

(a) (i) Under what assumptions is this relationship valid? How are the image co-ordinates, (u, v), and world co-ordinates, (X, Y, Z), computed from their 3-D and 4-D homogeneous co-ordinates, $\tilde{\mathbf{x}} = [x_1, x_2, x_3]^\top$ and $\tilde{\mathbf{X}} = [X_1, X_2, X_3, X_4]^\top$? [10%]

(ii) Give a factorisation of the projection matrix into components which encode the camera position, $\mathbf{T}$, camera orientation, $\mathbf{R}$, and internal camera calibration parameters, $\mathbf{K}$. Clearly identify the elements, total number of parameters and degrees of freedom for each component. [15%]

(iii) For an *uncalibrated* camera the elements of the projection matrix, p_{jk}, are unknown. Show how a set of N known reference 3-D points and their corresponding image points can be used to estimate the projection matrix. Include details of the optimisation techniques used when the measurements are noisy and there is a large number of 3-D reference points. [20%]

(iv) Write down the *projection matrix* for *weak perspective* projection and explain the advantages of using it for calibration. Under what viewing conditions is weak perspective a good camera model? [15%]

(b) A floor tiling company would like to build an automatic system for rectifying images of tiled floors in their catalogue. In particular, they would like to ensure that the edges of each square tile in the image accurately coincide with a pre-defined square grid.

(i) Show that the transformation between 3-D world points on the floor plane and corresponding image co-ordinates can be expressed as a 2-D projective transformation, $\mathbf{H}$. What is the rank of the matrix, $\mathbf{H}$, in the general case? [10%]

(ii) Show that the 2-D homogeneous representation of the line, $\tilde{\mathbf{l}}'$, on the image can be expressed in terms of a corresponding line, $\tilde{\mathbf{l}}$, on the tiled floor and matrix, $\mathbf{H}$. [10%]

(iii) How can elements of the matrix, $\mathbf{H}$, be recovered from line correspondences? What is the minimum number of line correspondences required? Why may line correspondences be preferred to point correspondences in this case? How can the rectified images be obtained? [20%]

3 (a) A typical Convolutional Neural Network (CNN) incorporates convolutional, non-linear and pooling stages. Describe each stage mathematically and explain the rationale behind their design. [20%]

(b) Consider applying a simple convolutional neural network to a 1-D image. Let $x_i^{(n)}$ denote the nth image in the training set. The network consists of two 1-D convolutions, each followed by the point-wise non-linearity, f. The network output vector, $\mathbf{y}^{(n)} = \left[y_1^{(n)}, y_2^{(n)}, \cdots, y_i^{(n)}, \cdots\right]^\top$, is obtained by applying a fully-connected layer with some non-linearity, g (e.g. softmax). The objective function, G, is used to train this network. The intermediate computation performed by this simple network is as follows:

$$
\begin{aligned}
a_i^{1,(n)} &= \textstyle\sum_k w_k x_{i-k}^{(n)}, & y_i^{1,(n)} &= f(a_i^{1,(n)}),\\
a_i^{2,(n)} &= \textstyle\sum_l v_l y_{i-l}^{1,(n)}, & y_i^{2,(n)} &= f(a_i^{2,(n)}),\\
a_i^{(n)} &= \textstyle\sum_d W_{i,d} y_d^{2,(n)}, & \mathbf{y}^{(n)} &= g(\mathbf{a}^{(n)}).
\end{aligned}
$$

Obtain an expression for the derivative required to implement gradient descent for the parameters, w_k. You can assume that $\frac{\partial G}{\partial y_i^{(n)}}$ and $\frac{\partial y_i^{(n)}}{\partial a_t^{(n)}}$ are known. [20%]

(c) A simple Convolutional Neural Network (CNN) is trained to perform a task of road sign classification. A database of RGB images covering 100 different types of road signs (e.g. "stop sign") under different viewing conditions is collected. Training images are obtained by tightly cropping the road signs from the original images and resizing these crops to 32×32 pixels for efficiency. Each image has a ground truth class label of a corresponding type of road sign associated with it. The CNN architecture consists of: (i) one fully-connected layer with a softmax non-linearity, (ii) three convolutional layers (K=5×5, S=1, A=ReLU), (iii) one max-pooling layer (K=2×2, S=2) arranged in some unspecified order. Here K is the kernel size, S is the kernel stride, A is the activation function. A bias term is used in all fully-connected and convolutional layers. Zero padding of two pixels is applied in all three convolutional layers. No padding is applied to the max-pooling layer.

The network is trained using the cross-entropy objective function. At test time, the most likely class label is returned as the predicted label for the test image.

(i) Propose a suitable order of the aforementioned layers, knowing that this network has 1703812 trainable parameters and that the number of output channels in the three convolutional layers are $2^n, 2^{n+1}$ and 2^{n+2}, where n is some positive integer. [20%]

(ii) In order to perform road sign classification on significantly larger resolution images (e.g. 224×224), an engineer has proposed to perform feature extraction by stacking B (e.g. $B > 4$) blocks of multiple layers one after another. Each block consists of three convolutional layers (K=7×7, S=1, A=ReLU) with zero padding of 3 pixels followed by one max-pooling layer. List three potential issues with such a straightforward approach to building a deeper network and explain in detail how design choices in the VGG-16 and ResNet-34 architectures address these issues effectively. [25%]

(iii) A company would like to deploy the road sign classification system described above to another country without needing any additional training steps. Note that the additional training would normally be required in order for the CNN to work on previously unseen classes of road signs. How should the original training and testing procedures be changed in order to accommodate this new scenario? In your answer, include a detailed description of an alternative objective function to be used. [15%]

4 A static scene is observed twice with a single camera. Corresponding points in a pair of images, (u, v) and (u', v'), are found by matching interest points extracted in each view.

(a) A 3-D point has co-ordinates $\mathbf{X}$ and $\mathbf{X}' = \mathbf{R}\mathbf{X} + \mathbf{T}$ in the left and right camera co-ordinate systems, respectively.

(i) Explain what is meant by the *epipolar constraint*? Derive this constraint. [20%]

(ii) Give an expression for the *fundamental matrix*, $\mathbf{F}$, in terms of the rotation matrix, $\mathbf{R}$, translation vector, $\mathbf{T}$, and internal calibration parameter matrix, $\mathbf{K}$. Derive how the *epipolar constraint* can be expressed in terms of $\mathbf{F}$. [15%]

(iii) Obtain algebraic expressions for *epipolar lines* for a point, (u, v), in the left image for the particular case: $\mathbf{R} = \mathbf{I}$, $\mathbf{T} = [-d, 2d, 0]^\top$, and

$$\mathbf{K} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Sketch the families of epipolar lines in the left and right images. [15%]

(b) Assume that the internal calibration parameter matrix, $\mathbf{K}$, of the camera is known and the relative position, $\mathbf{T}$, and orientation, $\mathbf{R}$, of the two images are unknown.

(i) Explain, in detail, how projection matrices of the left and right cameras can be recovered from known point correspondences in the general case. In your answer, discuss the minimum number of correspondences required and give details of any ambiguity. [15%]

(ii) Identify a set of configurations of the relative position, $\mathbf{T}$, and orientation, $\mathbf{R}$, for which the method described in Part (b)(i) would not be applicable. Provide an alternative approach to estimating the aforementioned projection matrices under this particular set of configurations. Can the 3-D co-ordinates of visible points in the scene be recovered in this case? [15%]

(c) Consider that additional images are taken with the same camera. Briefly describe a method to recover camera motion and to estimate 3-D co-ordinates of visible points in the scene. [20%]

END OF PAPER