Version IB/4

EGT3
ENGINEERING TRIPOS PART IIB

Monday 26 April 2021 1.30 to 3.10

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 and at the top of each answer sheet.*

STATIONERY REQUIREMENTS
Write on single-sided paper.

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed.
You are allowed access to the electronic version of the Engineering Data Books.

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

The time taken for scanning/uploading answers is 15 minutes.

Your script is to be uploaded as a single consolidated pdf containing all answers.

1 (a) A grey scale image, $I(x, y)$, is *low-pass* and *band-pass* filtered at multiple scales as part of the feature detection and matching process.

(i) The smoothed image, $S(x, y)$, is computed by filtering with the 2-D Gaussian, $G_\sigma(x, y)$. Give an expression for computing the intensity of a smoothed pixel efficiently using two discrete convolutions. Include expressions for the filter coefficients and the size of the filter kernel. [15%]

(ii) By considering the Fourier transform of the Gaussian, or otherwise, show that filtering with the Gaussian is low-pass filtering. Identify the relationship between the scale parameter, σ, and the cut-off frequency of the low-pass filter. [10%]

(iii) Why is low-pass filtering performed by convolution and not by using the Fourier transform in this case? [10%]

(iv) Show that convolution with the Laplacian of the Gaussian, $\nabla^2 G_\sigma(x, y)$, can be considered as band-pass filtering. Derive a discrete 3×3 filter which approximates the 2-D Laplacian. [15%]

(v) How is the Laplacian of a Gaussian implemented efficiently without the need for differentiation? How are low-pass and band-pass filtering at different scales implemented efficiently using an *image pyramid*? [10%]

(b) Consider an algorithm to detect and match image features in a 2-D image.

(i) Show how to compute the image position and scale of each feature from the image pyramid. [10%]

(ii) A 16×16 patch of pixels around each feature is sampled at the correct scale and orientation from the image pyramid. Why is this necessary and how is this achieved in practice? [10%]

(iii) The SIFT (Scale Invariant Feature Transform) descriptor is often used to describe the image feature and used for matching in different images and over different viewpoints. What properties of the image feature does the SIFT descriptor encode and how does it achieve its invariance to lighting and viewpoint changes? What are its limitations? [10%]

(iv) How are the best matches (correspondences) found in different images? Give details of an algorithm that can be used. [10%]

2 An object is viewed from two viewpoints with a mobile phone camera.

(a) (i) Under which viewing conditions will the correspondences in the two views satisfy the *epipolar constraint* shown below?

$$\tilde{\mathbf{x}}'^{\top}\mathbf{F}\tilde{\mathbf{x}} = 0, \text{ where } \tilde{\mathbf{x}}' = \begin{bmatrix} u' & v' & 1 \end{bmatrix}^{\top} \text{ and } \tilde{\mathbf{x}} = \begin{bmatrix} u & v & 1 \end{bmatrix}^{\top}.$$

Identify the dependence of the matrix $\mathbf{F}$ parameters on the translation, $\mathbf{T}$, and rotation, $\mathbf{R}$, of the camera's movement between images and the camera internal parameters, $\mathbf{K}$. [10%]

(ii) Show that epipolar lines are parallel when the camera motion is a pure translation parallel to the x-axis with no rotation and no change in the internal parameters. Why is this type of camera motion important in 3-D reconstruction? It may be useful to know that for any vector $\mathbf{t}$ and non-singular matrix $\mathbf{M}$ the following equality holds up to scale: $[\mathbf{t}]_{\times}\mathbf{M} = \mathbf{M}^{-\top}\left[\mathbf{M}^{-1}\mathbf{t}\right]_{\times}$. [15%]

(b) State the camera motions and scene geometries under which the transformation between point correspondences in successive images can be expressed as a 2-D projective transformation: [15%]

$$\tilde{\mathbf{x}}' = \mathbf{H}\tilde{\mathbf{x}}, \text{ where } \tilde{\mathbf{x}}' = \begin{bmatrix} u' & v' & 1 \end{bmatrix}^{\top} \text{ and } \tilde{\mathbf{x}} = \begin{bmatrix} u & v & 1 \end{bmatrix}^{\top}.$$

(c) The transformations in (a) and (b) above are estimated from point correspondences.

(i) How many point correspondences are required to estimate the transformation in each case? Explain your answer. [10%]

(ii) How are consistent matches obtained in the presence of incorrect or outlier measurements? Give details of the RANSAC (Random Sample Consensus) algorithm. [20%]

(d) *Conic sections* are planar curves and include the circle, ellipse, parabola and hyperbola. Consider the image of a conic section in the first view with equation $au^2 + buv + cv^2 + du + ev + f = 0$.

(i) Show that the conic section can be expressed using a symmetric matrix, $\mathbf{C}$, in homogeneous coordinates as: [10%]

$$\tilde{\mathbf{x}}^{\top}\mathbf{C}\tilde{\mathbf{x}} = 0, \text{ where } \tilde{\mathbf{x}} = \begin{bmatrix} u & v & 1 \end{bmatrix}^{\top}.$$

(ii) Hence, derive the equation of the conic section in the second viewpoint and comment on how a circle in the first view is transformed by the movement of the camera. [20%]

3 (a) A camera is to be calibrated from a single perspective image of a known 3-D object as shown in Fig.1. The pixel coordinates (u, v) corresponding to each known calibration point on the object, $\mathbf{X}$, with world coordinates (X, Y, Z) are measured.

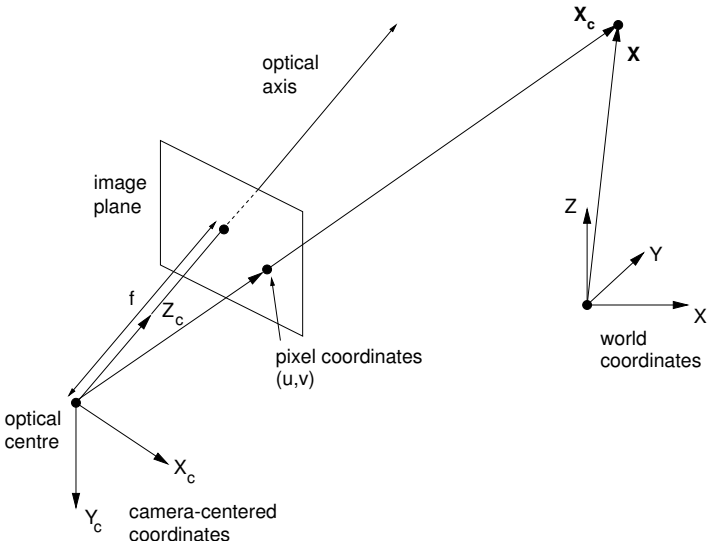

Fig. 1

(i) What properties of the 3-D object are desirable for calibration? [10%]

(ii) The relationship between the 3-D object coordinates and their perspective projection can be expressed by a 3×4 projection matrix. Derive this matrix and state any assumptions made. [15%]

(iii) Show that each image measurement gives two linear constraints on the unknown projection matrix parameters. What is the minimum number of calibration points needed? How is the set of linear equations solved from N reference points? Give details of the algorithm. [20%]

(iv) Why is the linear solution not optimal and how can a better solution be obtained? How are the position, orientation and focal length of the camera recovered? [10%]

(v) Explain how an augmented reality application running on a mobile phone can obtain the projection matrix from a single image of a known planar surface instead of a known 3-D object. [15%]

(b) A sawmill uses an automated inspection system to measure the volume of wood cut by measuring thickness, Z, of wood chips passing along a conveyor belt. A point mark is projected onto the wood chip surface using a laser light source and observed using a fixed camera. See Fig. 2.

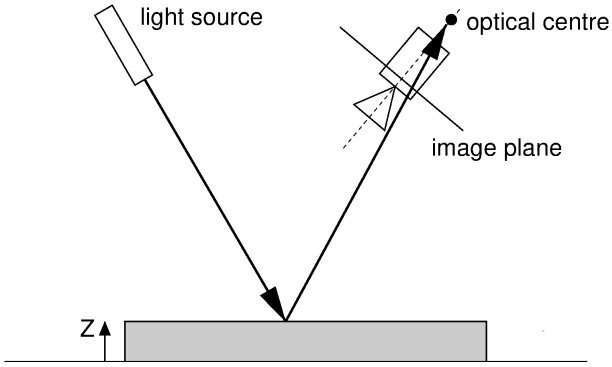

Fig. 2

(i) Show that the image of the projected point, l, moves along a line in the image as the thickness of the wood chip, Z, changes and that the transformation can be described in homogeneous co-ordinates by a 1-D projective transformation:

$$\begin{bmatrix} l \\ 1 \end{bmatrix} = \begin{bmatrix} p & q \\ r & s \end{bmatrix} \begin{bmatrix} Z \\ 1 \end{bmatrix}$$

[10%]

(ii) Describe a simple method for calibrating the inspection system and show how the thickness of the wood chip can be recovered continuously during inspection? [20%]

4 A six layer convolutional neural network (CNN) is used to classify a 28×28 resolution grey scale image into one of ten classes.

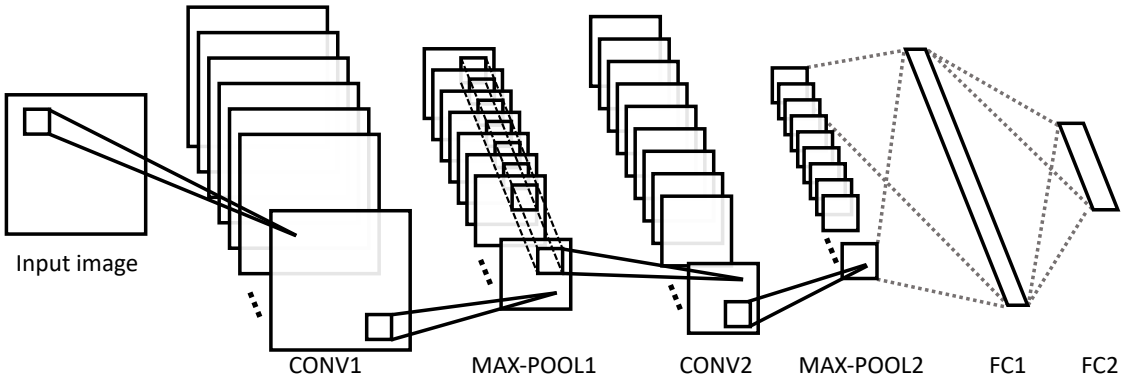

Fig. 3

Its architecture is illustrated in Fig. 3 and details of each layer are provided below:

(CONV1, K = 5×5, S = 1, C = 32, A = ReLU) $\rightarrow$ (MAX-POOL1, K = 2×2, S = 2) $\rightarrow$ (CONV2, K = 5×5, S = 1, C = 48, A = ReLU) $\rightarrow$ (MAX-POOL2, K = 2×2, S = 2) $\rightarrow$ (FC1, C = 256, A = ReLU) $\rightarrow$ (FC2, C = 10, A = Softmax).

Here K is the kernel size, S is the kernel stride, A is the activation function and C is the number of channels of a convolutional layer or the number of output units of a fully connected layer. A bias term is used in both the fully connected (FC1, FC2) and the convolutional layers (CONV1, CONV2). Two pixel padding is applied in both of the convolutional layers, CONV1 and CONV2. No padding is applied to the max-pooling layers, MAX-POOL1 and MAX-POOL2.

(a) Describe the role of each layer. Explain why non-linear activation functions such as Rectified Linear Units (ReLU) are used in CNNs. What is the role of the Softmax activation function in the final layer FC2? [20%]

(b) Provide a detailed calculation of the total number of parameters used in this network. You can assume that the output of the max-pooling layer MAX-POOL2 is flattened into a vector before it is passed to the fully connected layer FC1. [15%]

(c) Why is it important to consider the number of parameters when designing convolutional neural networks? List two architecture design choices which contribute to the reduction of the total number of parameters in the commonly used VGG-16 architecture. [20%]

(d) (i) Define an objective function for the training of the aforementioned six layer network. You can ignore weight regularisation. [10%]

(ii) Let the output $\mathbf{y} = (y_0, \ldots, y_c, \ldots, y_9)^\top$ of the FC2 layer be defined as:

$$y_c = \frac{\exp(\sum_{i=0}^{255} x_i w_{c,i} + b_c)}{\sum_{k=0}^{9} \exp(\sum_{i=0}^{255} x_i w_{k,i} + b_k)}$$

Here $\mathbf{x} = (x_0, \ldots, x_i, \ldots, x_{255})^\top$ is the input to the FC2 layer and $\mathbf{w}_c = \left(w_{c,0}, \ldots, w_{c,i}, \ldots, w_{c,255}\right)^\top$ and b_c correspondingly are weights and biases of the c-th neuron of the fully connected layer FC2. Obtain the expression of the derivative required to implement gradient descent for the parameters $w_{c,i}$. Simplify your answer. [25%]

(e) During training, you noticed that your network takes significantly longer to converge than a similar network implemented by one of your classmates. What layer would you insert after the convolutional layers CONV1 and CONV2 in order to match the performance, if you knew that both architectures and training procedures were otherwise identical. Explain your choice. [10%]

END OF PAPER

THIS PAGE IS BLANK