Version RC/6

EGT3
ENGINEERING TRIPOS PART IIB

Wednesday 6 May 2026 14.00 to 15.40

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 greyscale image, $I(x, y)$, is first low-pass filtered (smoothed) by convolving with a 2-D Gaussian filter, $G_\sigma(x, y)$, before gradients are computed as part of the feature detection process.

(a) (i) Show how smoothing can be performed by two 1-D convolutions. Give an expression for computing the intensity of a smoothed pixel, $S_\sigma(x, y)$, using two discrete 1-D convolutions. Your answer should identify the size and elements of the discrete filter kernels. [15%]

(ii) Find an expression for the Fourier Transform of the 1-D Gaussian, $g_\sigma(x)$. Hence show why the Gaussian is a suitable low-pass filter. [10%]

(iii) Show that repeated convolutions with a series of 1-D Gaussians, each with a particular standard deviation σ_i, is equivalent to a single convolution with a Gaussian of variance $\sum_i \sigma_i^2$. How is this used to compute efficiently low-pass filtering with large values of scale, σ? [15%]

(b) A *blob-like* image feature is usually localised in position and scale by filtering the image with the scale-normalised Laplacian of a Gaussian filter, $\sigma^2 \nabla^2 G_\sigma(x, y)$, over different scales.

(i) Derive the second derivative of a 1-D Gaussian function. [10%]

(ii) Sketch the output of the convolution of a 1-D bar of width $2w$ with the second derivative of the 1-D Gaussian. [10%]

(iii) Hence find the filter scale σ that is needed to localise the centre of a bar with apparent size $2b$ and the centre of a blob with apparent radius r. [10%]

(iv) Show that the scale-normalised Laplacian of a Gaussian filter acts a band-pass filter. [10%]

(c) The low-pass filtered images and their spatial derivatives can also be used for image *edge* and *corner* detection at different scales.

(i) What is an image *edge*? How are edges detected and localised in images? [10%]

(ii) What is meant by a *corner* feature and how are corners detected and localised in the image? [10%]

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

(a) Under what assumptions is this relationship valid? [10%]

(b) For a *calibrated* camera the elements of the projection matrix are known.

(i) What is the perspective projection of the 3-D point with co-ordinates (X, Y, Z) and give an expression for the distance to the point along the optical axis. [10%]

(ii) Derive the equation of the *horizon* of the $X - Y$ plane under perspective projection. [10%]

(c) For an *uncalibrated* camera the elements of the projection matrix, p_{jk}, are unknown.

(i) Find the minimum number of measurements that are needed to be able to recover the projection matrix. What are the degrees of freedom of a real camera? [10%]

(ii) Show how to recover the position and orientation of a known 3-D object. Include details of the optimisation techniques used when the measurements are noisy. [20%]

(d) *Weak perspective* projection consists of an *orthographic* projection onto a plane which is parallel to the image plane followed by a perspective projection onto the image plane.

(i) Derive the projection matrix between a 3-D point and its image under weak perspective projection. [10%]

(ii) Under what viewing conditions is weak perspective a good camera model? What are its advantages? [20%]

(iii) What happens to the $X - Y$ plane's horizon under weak perspective? [10%]

3 (a) A filter bank consists of N engineered 3×3 weight matrices, forming an $N \times 3 \times 3$ tensor W_0. The filter bank is applied to a $D \times D$ greyscale input image I_0, resulting in a $N \times D \times D$ tensor I_1. In this notation $I_l^{(n,r,c)}$ indicates the pixel value which is the output of layer l at channel n, row r, and column c.

(i) Write an expression for the pixel at $I_1^{(1,1,1)}$ (using zero-based indexing as shown in Fig. 1) in terms of W_0 and I_0. [10%]

(ii) I_1 has the same spatial dimensions (*i.e.*, $D \times D$) as I_0. How is this achieved? Be specific about the mechanism used, and how it works. [10%]

(iii) Based on the image in Fig. 1 showing the filter at W_0^0 (*i.e.*, all 3×3 weights at channel 0), what features from the original image will be highlighted in I_1^0? [10%]

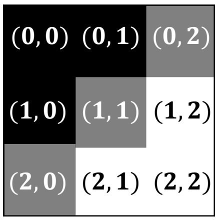

Fig. 1

(b) The application of the filter bank W_0 forms a tensor I_1 which then acts as the input to a convolutional neural net (CNN) shown in Fig. 2.

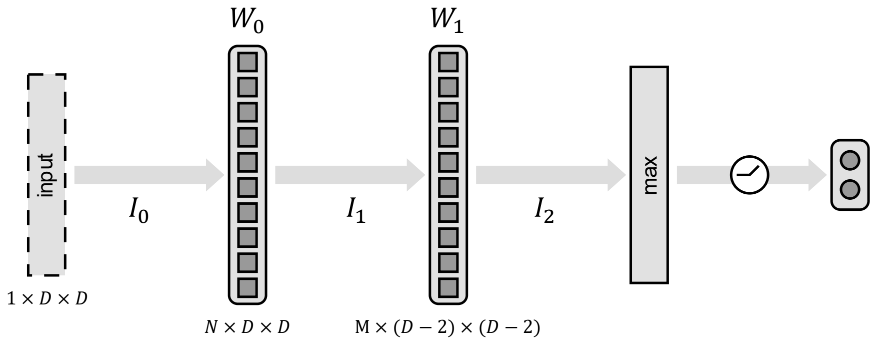

Fig. 2

(i) Weigh the pros and cons of using engineered features like a filter bank as the input to a CNN. When might this be a good idea? [10%]

(ii) The first layer of the CNN has a tensor of filter weights W_1, resulting in an output tensor I_2 of dimension $M \times (D-2) \times (D-2)$. What is a possible shape of W_1, and why are the dimensions of I_2 smaller than I_1 given that shape? [10%]

 (cont.

(iii) The next step in the CNN is a pooling layer. What might a convolutional layer followed by a pooling layer enable the network to represent in terms of the semantics of the image? [10%]

(iv) What is fine-tuning in the context of a CNN, and how could it be used to improve the values in W_0? What are common pitfalls of fine-tuning a network, and how can they be avoided? [10%]

(c) The W_0 filter bank has been replaced by a module that computes the pixel values as:

$$I_1^{(n,r,c)} = \begin{cases} I_0^{(0,r,c)} + \cos rn\pi & n \bmod 4 = 0 \\ I_0^{(0,r,c)} + \cos cn\pi & n \bmod 4 = 1 \\ I_0^{(0,r,c)} + \sin rn\pi & n \bmod 4 = 2 \\ I_0^{(0,r,c)} + \sin cn\pi & n \bmod 4 = 3 \end{cases}$$

What does this do? Why is it redundant for a CNN, but required for a Vision Transformer? [10%]

(d) Transformers tokenise images by dividing them into a regular grid of patches, and then converting those image patches into token vectors. Describe how the patch dictionary tokenisation technique works, and compare and contrast it with learned embeddings. [20%]

4 An object is to be reconstructed in 3-D by photographing it from two viewpoints with a mobile phone camera. The correspondences in the left and right images, (u, v) and (u', v'), are first found by matching feature points extracted in each view.

(a) What is the geometric constraint that these correspondences must satisfy? Express this constraint algebraically using the *fundamental* matrix. Your answer should include a factorization of the fundamental matrix into matrices representing camera translation and rotation. [15%]

(b) The fundamental matrix is to be estimated from pairs of point correspondences.

(i) Describe an algorithm for finding the feature points and their possible matches. [15%]

(ii) What is the minimum number of point correspondences needed to estimate the fundamental matrix? Describe an algorithm for finding consistent matches in the presence of incorrect or outlier measurements. [10%]

(iii) How is the transformation estimated when a large number of consistent matches is available? What special constraints must be enforced? [20%]

(c) The fundamental matrix is to be factorized to recover the projection matrices for each viewpoint.

(i) First describe an algorithm to recover the camera motion (rotation and translation) between the two views if the internal camera parameters (matrix K) are known. Include details of how any ambiguities are resolved. [20%]

(ii) Show how to recover the 3-D positions of each feature point and give the equations that need to be solved. [10%]

(d) What will happen to the 3-D reconstruction if the camera motion is a pure rotation or is very small? [10%]

END OF PAPER