Version RC/3

EGT3
ENGINEERING TRIPOS PART IIB

Monday 8 May 2017 2 to 3.30

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.

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

1 (a) A grey scale image, $I(x,y)$, is smoothed by convolving it with two 1-D Gaussian filters as part of the feature detection process:

$$S(x,y) = \sum_{u=-n}^{n} \sum_{v=-n}^{n} g_\sigma(u) g_\sigma(v) I(x-u, y-v)$$

(i) Explain why smoothing is necessary and how an appropriate value of σ is selected. [10%]

(ii) Show that the smoothing operation above is equivalent to convolving the image with a 2-D Gaussian. What is the advantage of performing the smoothing as two 1-D convolutions? [20%]

(iii) Determine the size, n, and the coefficients (filter weights) of the discrete approximation of the 1-D Gaussian smoothing filter, $g_\sigma(x)$, for $\sigma = 1$. Give details of any design choices made. [30%]

(b) Consider an algorithm to match image features in 2-D images. The neighbourhood of each image feature is first geometrically normalised to a 16×16 patch of pixels by sampling pixels at an appropriate scale and orientation and then a descriptor is to be computed.

(i) A simple way to compare patches is to calculate the normalised cross-correlation. Describe a descriptor to achieve this. [10%]

(ii) Describe the descriptor used in the scale-invariant feature transform (SIFT) and show how is it computed from the patch of pixels? [20%]

(iii) Comment on the advantages and disadvantages of the descriptors in (b) (i) and (ii) above. [10%]

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

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

(a) (i) Under what assumptions is this relationship valid? Show how $\mathbf{X}$ is related to $(X_1,X_2,X_3,X_4)^T$. [10%]

(ii) What is the algebraic and geometric significance of the variable, s ? [10%]

(iii) How are 3-D points at infinity represented? [10%]

(iv) Give a factorisation of the projection matrix into components which encode the camera position, camera orientation and *internal* camera calibration parameters. [20%]

(b) Consider a point, $\mathbf{X}_i$, on a 3-D line which is parameterised such that $\mathbf{X}_i = \mathbf{a} + \lambda_i \mathbf{b}$.

(i) Find the perspective projection of points, $\mathbf{X}_i$, on the 3-D line using the projection matrix given above. [10%]

(ii) Hence show that under perspective projection, parallel lines in 3D intersect at *vanishing points* in the image. Determine the image co-ordinates of the vanishing point and show that it depends only on the orientation of the camera and the direction of the line $\mathbf{b}$ but not on its position relative to the camera centre, $\mathbf{a}$. [20%]

(iii) Recover the equation of the *horizon* of the ground plane (X-Y world plane). [20%]

3 Consider multiple views of a static scene which have been taken 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) Under which two viewing conditions will correspondences in the two views be described by the 2-D *projective transformation* given below? Identify the relationship of the transformation parameters and the camera motion in each case. [20%]

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

(b) Consider the degrees of freedom of the 2-D projective transformation.

(i) An object appears as a square in the first image. Describe, using sketches, how it might appear in the second image. [20%]

(ii) What would happen under *weak* perspective projection? [10%]

(c) Under which viewing conditions will the correspondences in the two views be described by the *fundamental matrix* shown below? Identify the relationship of the matrix parameters and the camera motion. [20%]

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$

(d) Describe an algorithm to recover the camera motion between two views from a set of point correspondences if the internal camera parameters, $\mathbf{K}$, are known. [30%]

4 (a) Describe the components that make up a standard convolutional neural network that is used for image classification. Briefly explain the rationale for the form of each of the components. [40%]

(b) A large number of satellite images of the earth have been collected and labelled according to whether they are images of urban areas or not. A convolutional neural network is to be trained to automatically classify new satellite images. The training dataset comprises N greyscale images $\{Z^{(n)}\}_{n=1}^{N}$ and binary labels $\{\mathrm{t}^{(n)}\}_{n=1}^{N}$ which indicate whether the image is of an urban area.

The network contains three stages. The first stage carries out a 2-D convolution between the image pixels $Z_{i,j}^{(n)}$ and convolutional weights $W_{i,j}$,

$$a_{i,j}^{(n)} = \sum_{k,l} W_{k,l} Z_{i-k,j-l}^{(n)}$$

The second stage applies a point-wise non-linearity $y_{i,j}^{(n)} = f\left(a_{i,j}^{(n)}\right)$.

The third stage applies a set of output weights $V_{i,j}$ and a logistic non-linearity in order to form the scalar output of the network,

$$x^{(n)} = \frac{1}{1+\exp\left(-\sum_{i,j} V_{i,j} y_{i,j}^{(n)}\right)}$$

The network's weights will be trained using the cross-entropy objective function,

$$G(V,W) = -\sum_{n=1}^{N}\left(\mathrm{t}^{(n)}\log(x^{(n)}) + (1-\mathrm{t}^{(n)})\log(1-x^{(n)})\right) + \frac{\alpha}{2}\sum_{i,j} V_{i,j}^2 + \frac{\beta}{2}\sum_{i,j} W_{i,j}^2$$

(i) Compute the derivative required to implement gradient descent learning of the network's convolutional weights W. Simplify your expression and interpret the terms. [40%]

(ii) The company would like to extend the network to estimate the average population density of the area in each image. Describe how to extend the architecture of the network to perform this additional task. Explain your design. [20%]

END OF PAPER

THIS PAGE IS BLANK