Version RC/4

EGT3
ENGINEERING TRIPOS PART IIB

Monday 4th May 2015 9.30 to 11

Module 4F12

COMPUTER VISION AND ROBOTICS

*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
Engineering Data Book
CUED approved calculator allowed

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) Why are the raw pixel intensity values of an image, $I(x,y)$, rarely used in computer vision algorithms? Describe two simple ways of making the images more invariant or insensitive to changes in lighting brightness and contrast. [10%]

(b) Images are often smoothed with a low-pass filter before image gradients are computed.

(i) What smoothing filter is used in practice? Give an expression for computing the intensity of a smoothed pixel, $S(x,y)$, with two discrete 1D convolutions. [20%]

(ii) Show how smoothing at different scales can be computed efficiently in an *image pyramid*. Your answer should include details of the implementation of smoothing within an octave and subsampling of the image between octaves. [20%]

(c) Consider an algorithm to detect and match interest points (features of interest) in a 2D image.

(i) Show how image features such as *blob-like* shapes can be localized in both position and scale. [15%]

(ii) Show how the neighbourhood of each image feature can be normalised to a 16×16 patch of pixels. [15%]

(iii) The SIFT (Scale Invariant Feature Transform) descriptor is often used to describe interest points in order to match them in different images and over different viewpoints. Describe the main steps in computing this descriptor and how it achieves its invariance to lighting, image and viewpoint changes. What are its limitations? [20%]

2 The relationship between a 3D 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 as follows

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

(a) Under what assumptions is this relationship valid? How are the image co-ordinates, (u,v), and world co-ordinates, (X,Y,Z), computed from their 3D and 4D homogeneous representations, **x** and **X**, respectively? [10%]

(b) (i) Give expressions for the image co-ordinates of a point, u_i and v_i, in terms of the 3D world co-ordinates, (X_i,Y_i,Z_i) and the elements of the 3×4 projection matrix above. [10%]

(ii) For a *calibrated* camera (i.e. the elements of the projection matrix are known) show how the visual ray from the optical centre and through an image point (u_i,v_i) can be defined by the intersection of two planes. Give algebraic expressions for these two planes. [10%]

(iii) For an *uncalibrated* camera, show how these equations can be used to recover the elements of an unknown projection matrix from the projections of known 3D points. How many correspondences are needed? [20%]

(c) (i) Show how the projection matrix can be modified to represent the projective transformation obtained when viewing the ground plane ($Z = 0$). [10%]

(ii) How many degrees of freedom does this new projective transformation have? Describe, using sketches, how a square on the ground may appear after the transformation. Be sure to account for each degree of freedom. [20%]

(iii) By considering the projection of points on the plane at infinity, or otherwise, derive the equation of the *vanishing* line (horizon) of the ground plane. [20%]

3 Consider stereo vision where a 3D point has image projections (u,v) and (u',v') in the left and right cameras. The 3D point has coordinates $\mathbf{X}$ and $\mathbf{X}' = \mathbf{R}\mathbf{X} + \mathbf{T}$ in the left and right camera coordinate systems, respectively. The internal calibration parameter matrices of the left and right cameras are represented by $\mathbf{K}$ and $\mathbf{K}'$ respectively.

(a) (i) What is meant by the *epipolar constraint* in stereo vision and how is it used to find image correspondences? [20%]

(ii) The epipolar constraint can be expressed algebraically with the *fundamental* matrix

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

Derive this constraint and give an expression for the fundamental matrix in terms of the translation, $\mathbf{T}$, and rotation, $\mathbf{R}$, between cameras. [30%]

(b) A single camera moves in a static scene. Correspondences between the first two views are to be used to recover the relative camera positions and orientations.

(i) Show how the fundamental matrix can be estimated from these stereo pair correspondences. What additional constraints must the fundamental matrix satisfy and how are these met in practice? [15%]

(ii) Show how to recover the projection matrices for the left and right camera viewpoints from the fundamental matrix and knowledge of the internal calibration parameters of the camera. Include details of any ambiguity. [15%]

(c) How are the 3D positions of points in the scene recovered? What are the advantages of using a single camera that can move in a static scene? Give details of methods to improve the accuracy of the reconstruction. [20%]

4 A *convolutional neural network* is to be trained for pedestrian detection. A labelled dataset has been collected that contains N greyscale images $\{Z^{(n)}\}_{n=1}^{N}$ and binary labels $\{\mathrm{t}^{(n)}\}_{n=1}^{N}$ which indicate whether pedestrians are present in each image.

The network contains three stages. The first stage carries out a 2D 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 following objective function,

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

(a) Provide a probabilistic interpretation for the network's output and use this to justify the form of the objective function. [20%]

(b) Describe how to train the network's convolutional weights W using gradient descent. Compute the derivative required to implement gradient descent. Simplify your expression and interpret the terms. [40%]

(c) Describe enhancements to the architecture of the network that might improve its ability to perform pedestrian detection. [40%]

END OF PAPER

THIS PAGE IS BLANK