Version RC/6

EGT3
ENGINEERING TRIPOS PART IIB

Wednesday 3 May 2023   2 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 grey scale 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) Give an expression for computing the intensity of a smoothed pixel, $S_\sigma(x, y)$, using two discrete 1-D convolutions. [10%]

(ii) Why is smoothing necessary? By considering the Fourier transform of the Gaussian, or otherwise, show that the Gaussian kernel is a suitable low-pass filter and identify the relationship between the scale parameter, σ, and the cut-off frequency of the low-pass filter. [10%]

(iii) How is low-pass filtering at multiple scales implemented efficiently using an *image pyramid* to sample *scale-space*? Illustrate your answer by considering an image pyramid of low-pass filtered images, $S_{\sigma_0} \ldots S_{64\sigma_0}$, with $s = 3$ distinct images in each octave. How many different Gaussian filters are needed to build the pyramid and what is the size of the images in each octave? [20%]

(b) Differentiation of the smoothed image, $S_\sigma(x, y)$, can be implemented with discrete convolutions.

(i) By first considering the Taylor series expansion of $S_\sigma(x, y)$ show that approximations for the second-order derivatives, $\partial^2 S_\sigma / \partial x^2$ and $\partial^2 S_\sigma / \partial y^2$, can be computed by convolving $S_\sigma(x, y)$ with discrete 1-D convolutions. Identify the filter coefficients needed for each derivative. [15%]

(ii) Hence derive the 2-D filter needed to compute the Laplacian of the smoothed image, $\nabla^2 S_\sigma(x, y)$. [5%]

(iii) Show that convolution with the Laplacian of the Gaussian, $\nabla^2 G_\sigma(x, y)$, can be considered as *band-pass* filtering. [10%]

(c) The band-pass filtered images can be used for image *edge* and *blob* detection at different scales.

(i) What is an image edge? How are edges detected in an image which has been convolved with the Laplacian of a Gaussian, $\nabla^2 S_\sigma(x, y)$? [10%]

(ii) What is an image blob and how are they detected? Show how to recover the approximate scale and size of the blob. [10%]

(iii) How are edges and blobs used in object recognition? How do they achieve geometric or photometric invariance? [10%]

2 A camera is used to view a three dimensional object under perspective projection. The image co-ordinates, (u_i, v_i), of the 3-D world point, (X_i, Y_i, Z_i), are given by:

$$u_i = \frac{p_{11}X_i + p_{12}Y_i + p_{13}Z_i + p_{14}}{p_{31}X_i + p_{32}Y_i + p_{33}Z_i + p_{34}}$$

$$v_i = \frac{p_{21}X_i + p_{22}Y_i + p_{23}Z_i + p_{24}}{p_{31}X_i + p_{32}Y_i + p_{33}Z_i + p_{34}}$$

(a) (i) By introducing *homogeneous* co-ordinates show how the mapping from world to pixel co-ordinates can be expressed as a 3×4 projection matrix. [10%]

(ii) What is meant by a *vanishing point*? Find the *vanishing* point of lines which are parallel to the Z-axis. [10%]

(iii) What is meant by the *horizon*? Derive the equation of the horizon of the $X - Y$ ground plane. [10%]

(b) A set of N known reference 3-D points, $\{(X_i, Y_i, Z_i)\}_{i=1}^{N}$, and their corresponding projections, $\{(u_i, v_i)\}_{i=1}^{N}$, are to be used to calibrate the camera.

(i) What is meant by camera calibration? List the parameters that need to be estimated. [10%]

(ii) How many reference points, N, are needed in practice and what are the properties of a good calibration object? [10%]

(iii) Show how the unknown parameters can be estimated from the image measurements of the projections of the known 3-D points. Include details of the optimisation techniques used when the measurements are noisy. [20%]

(c) After calibration the known projection parameters can be used to recover 3-D scene co-ordinates.

(i) Show how the visual ray from the optical centre that passes 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%]

(ii) Show that there is an ambiguity in the recovery of the 3-D co-ordinates of a visible point. How can this ambiguity be removed? Your answer should give details of the algebraic equations that need to be solved and their geometrical interpretation. [20%]

3 A five layer convolutional neural network (CNN) is used to extract features for a face recognition application from a 32 × 32 resolution grey scale image.

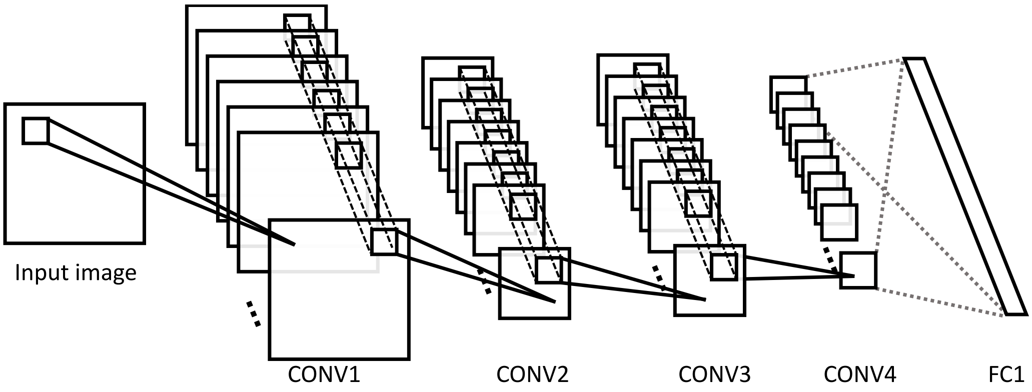

Fig. 1

Its architecture is illustrated in Fig. 1 and details of each layer are provided below:

(CONV1, K = 5 × 5, S = 2, C = 16, P = 2, A = ReLU) → (CONV2, K = 3 × 3, S = 1, C = 32, P = 1, A = ReLU) → (CONV3, K = 5 × 5, S = 2, C = 64, P = 2, A = ReLU) → (CONV4, K = 3 × 3, S = 1, C = 128, P = 1, A = ReLU) → (FC1, C = 128, A = Linear)

Here K is the kernel size, S is the kernel stride, P is padding, 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) and the convolutional layers (CONV1-4). The output of the convolutional layer CONV4 is flattened into a vector before it is passed to the fully connected layer FC1.

(a) Explain the role of each layer in this network and provide a detailed calculation of the total number of parameters used.

[20%]

(b) (i) A *Siamese network* setup with the *contrastive loss* is used to train this network. Explain this setup in detail, including the precise mathematical definition of the objective function used. You can ignore weight regularisation. [10%]

(ii) Obtain the expression of the derivative required to implement the gradient descent update for the parameters, $w^c_{i,j,c'}$, of the final convolutional layer, CONV4. Here i, j correspond to the spatial output coordinates, c' is the input channel index and c is the output channel index respectively. Let $y_{i,j,c'}$ be the output of the previous layer, CONV3, and hence the input to CONV4. Simplify your answer. [25%]

(iii) When training the network described in Fig.1, poor performance was observed when: images of people who were not represented in the training data and when darker or lighter images than typical images in the training data were used. What two simple changes could be made to the proposed architecture in order to counter the poor performance? Provide the details of these architectural changes. [15%]

(c) In order to perform face recognition on larger resolution images (e.g. 224 × 224) on a modern phone, more powerful deep neural networks are needed such as ResNet or Vision Transformer (ViT). Explain three advantages and/or disadvantages of using transformer-based networks over convolutional neural networks. [10%]

(d) You are asked to re-purpose the network shown in Fig.1 to build an efficient solution for joint segmentation and recognition of faces in group photos (resolution larger than 224 × 224) of people. The following constraints are placed on your solution:

- all faces in the group photos are of the same size and can be well recognised within 32 × 32 square of pixels;
- along with the original dataset of 32 × 32 images and corresponding face identity labels (some identities can have more than one image available) used to train the network in Fig.1, you are given an additional very small dataset of pairs of group photos and corresponding per-pixel labels of the identities of faces.

Explain what changes to the architecture, training and testing procedure, data processing and objective function you would make in order to achieve the desired objective. [20%]

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

(a) The correspondences in the left and right images, (u, v) and (u', v'), satisfy the *epipolar constraint* :

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

(i) Derive this constraint and identify the dependence of the matrix parameters on the translation, **T**, and rotation, **R**, of the camera's movement between images and the camera internal parameters, **K**. [20%]

(ii) Hence derive an algebraic expression for the epipolar line corresponding to a point in the left image with pixel coordinates (u, v). [10%]

(b) Under which viewing conditions (camera motion or scene geometry) will the transformation between point correspondences in successive images be described by the 2-D projective transformation below? [20%]

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

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

(i) First a large number of *keypoints* are detected in each image and potential matches from one image to the other are found by comparing their descriptors. How are matches found in the presence of scale, orientation and lighting changes? [20%]

(ii) How are consistent matches obtained in the presence of incorrect or outlier measurements? [10%]

(iii) Show how the transformations can be estimated and then used to recover the camera motion if the camera internal parameters are known. [20%]

END OF PAPER