CST0
COMPUTER SCIENCE TRIPOS Part IA

Friday 9 June 2023 09:00 to 12:00

COMPUTER SCIENCE Paper 3

*Answer **one** question from each of Sections A, B and C, and **two** questions from Section D.*

*Submit the answers in five **separate** bundles, each with its own cover sheet. On each cover sheet, write the numbers of **all** attempted questions, and circle the number of the question attached.*

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

STATIONERY REQUIREMENTS
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

SECTION A

1 Databases

(*a*) A manufacturer makes three models of vehicle that vary in their engine type and number of seats. One model is available in two paint colours. Engines are either petrol or electric. Engines come from different suppliers, dependent on their fuel and horsepower.

(*i*) Draw a suitable E/R diagram. [4 marks]

(*ii*) By writing out a suitable number of short tables, give a small relational database example holding vehicles and engines. Make sure every field has a name. Underline table keys. [3 marks]

(*iii*) Which fields in your example are foreign keys? Explain whether your example satisfies referential integrity. [2 marks]

(*b*) The SQL language contains a `GROUP BY` construct.

(*i*) Explain why the data returned by `GROUP BY` has to be passed through a reduction operator before it can be returned in a table? [2 marks]

(*ii*) What mathematical properties should such a reduction operator have? Explain why. [2 marks]

(*c*) A distributed database may use *eventual consistency*. Explain what this is. Give two advantages. [3 marks]

(*d*) State the principle disadvantage of *eventual consistency*? Provide a simple example. [2 marks]

(*e*) Explain why a graph database typically holds just one graph? [2 marks]

2 Databases

(*a*) Is the relational database join operator associative? Describe an example where different associations might result in significant execution time differences. [4 marks]

(*b*) Give an example relation where there is more than one potential key. Ensure your example has at least two reasons why some of these candidate keys are not suitable for use as the primary key. Explain the reasons. [4 marks]

(*c*) In this question, an XML document is a tree with named nodes, called elements, whose leaves are character strings. In addition, an element has an unordered list of string pairs where the strings are an attribute name and its value.

```
xml_t = | LEAF of string
        | ELEMENT of string * (string * string) ulist * xml_t list
```

(*i*) A document may be stored in XML in various ways, varying from rigidly structured to loosely structured. Describe when this can be useful. How can variations in structure be tolerated or reported? [5 marks]

(*ii*) How might all the data held in a relational DBMS sensibly be exported into a single XML document? [4 marks]

(*d*) What is the purpose of the project operator in the relational algebra? Would there be an equivalent operator in a document or graph database? [3 marks]

SECTION B

3 Introduction to Graphics

(a) The Phong reflection model is expressed as

$$I = I_{\mathrm{a}}\, k_{\mathrm{a}} + \sum_i I_i\, k_{\mathrm{d}}(L_i \cdot N) + \sum_i I_i\, k_{\mathrm{s}}(R_i \cdot V)^n\,. \tag{1}$$

(i) In which situation do the terms $L_i \cdot N$ and $R_i \cdot V$ become negative? How should this case be handled? [2 marks]

(ii) The equation above does not account for cast shadows. How can we simulate shadows in ray tracing? [1 mark]

(iii) How can we simulate soft shadows with blurry boundary in ray tracing? [3 marks]

(iv) You work on an accelerated rendering pipeline in which the results of shading computation can be stored in object's texture and then reused in future frames. We assume that only the Phong reflection model is used. Determine and enumerate which terms of the reflection model (ambient, diffuse and specular) can be reused in the future frames when the camera, objects or lights move. [4 marks]

(b) You came up with an idea to render shadows in rasterization. You first render the depth map of the scene from the point of view of the light source and store the result in a texture, which you call a shadow map. Then, when rendering the scene, you use the shadow map as a replacement for shadow rays.

(i) What condition needs to be met to determine that a fragment is in shadow? [2 marks]

(ii) How can you find the (u, v) coordinates in the shadow map for a fragment with the world coordinates p? You have view, V, and projection, P, matrices of the camera that rendered the shadow map. The model matrix is an identity matrix. [4 marks]

(iii) Discuss the artifacts that you are likely to see when rendering shadows with your method. [4 marks]

4 Introduction to Graphics

You are asked to implement a Java class for storing images with arbitrary pixel order and an arbitrary number of colour channels. The skeleton of such a class is provided below. If `row_major` is set to `true` in the constructor, the class stores pixels in the row-major order and in the column-major order otherwise. If `interleaved` is set to `true` in the constructor, the class stores colour values in the interleaved order and in the planar order otherwise.

```
public class ExImage {
    final protected byte[] data;
    final protected int width, height, colour_channels, sx, sy, sc, first_pixel;

    public ExImage(int width, int height, int colour_channels,
                   boolean row_major, boolean interleaved)
    {
        data = new byte[width*height*colour_channels];
        this.colour_channels = colour_channels;
        this.width = width;  this.height = height; this.first_pixel = 0;
        ...
    }
    protected int get_index( int x, int y, int cc )
    ...
    public void set_pixel( int x, int y, byte[] value )
    ...
    public byte[] get_pixel( int x, int y )
    ...
}
```

(*a*) Write the missing piece of code in the constructor for setting the strides `sx`, `sy` and `sc` of the `ExImage` object. [4 marks]

(*b*) Implement `get_index`, `get_pixel`, and `set_pixel` methods. [3 marks]

(*c*) You want to add a region-of-interest functionality to the class. Write the code for a constructor with the signature

```
public ExImage( ExImage src_img, int ox, int oy, int width, int height)
```

that creates an object that operates on the region `(ox, oy, ox+width, oy+height)` of the image `src_img` without creating a copy of the data. [3 marks]

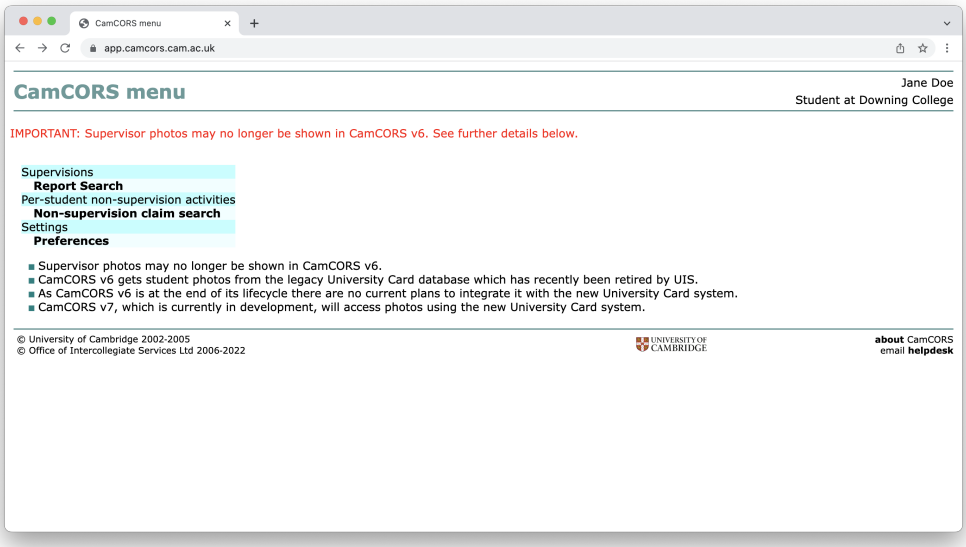
(*d*) Would you recommend storing linear or display-encoded pixel values in this class? Justify. [3 marks]

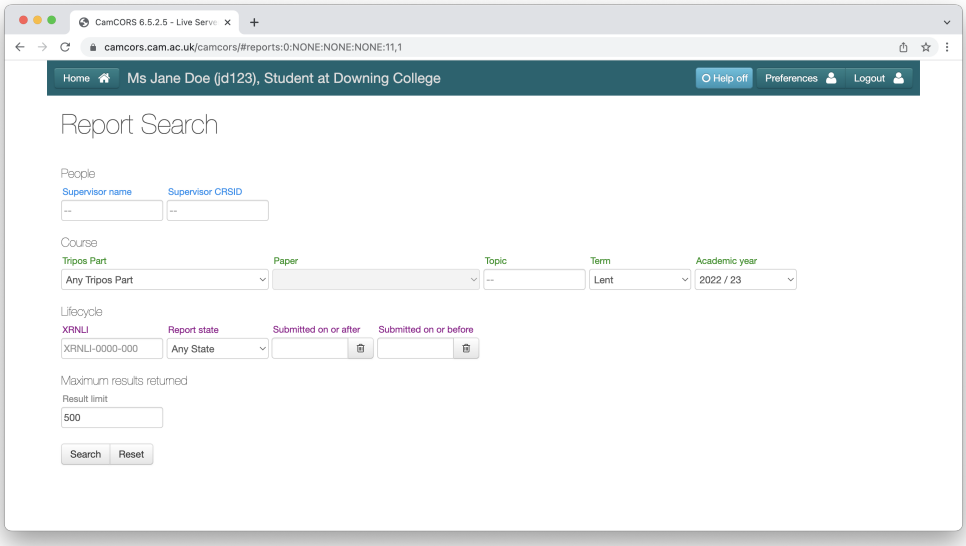
(*e*) The object of the `ExImage` class stores RGB values that are shown on a display with non-standard primaries $r(\lambda)$, $g(\lambda)$, $b(\lambda)$, where λ is the wavelength. Derive a formula for converting those RGB values to the display-encoded BT.709 RGB colour space. You are given CIE 1931 colour matching functions $x(\lambda)$, $y(\lambda)$, $z(\lambda)$ and a matrix $M_{XYZ\rightarrow 709}$ for converting from CIE 1931 XYZ to BT.709. Both the display and the target colour use a gamma of 2.2. Write equations rather than code. [7 marks]

SECTION C

5 Interaction Design

Considering CamCORS, the online supervision reporting system used by Cambridge colleges, answer the following questions.





(*a*) Describe one aspect about human memory as it applies to interaction design, and then identify and explain a way in which the design as illustrated above made use of this aspect. [3 marks]

(*b*) Describe one aspect about human attention as it applies to interaction design, and then identify and explain one way in which the design as illustrated above made use of this aspect. [3 marks]

(*c*) Describe what information architecture is, and then create a diagram outlining the information architecture of the system illustrated above. [3 marks]

(*d*) Identify one aspect of the information architecture of CamCORS that might be confusing for users and explain why this could be a usability issue. [3 marks]

(*e*) Describe how you would go about evaluating the current design to find out if the identified aspect does indeed pose usability problems or not. Your description should include which research method you expect to use and why, a brief summary of how you would go about using the method in this specific case, what data you expect to collect, and how you plan to analyse it. [8 marks]

6 Interaction Design

Anxiety and depression are among the leading mental health issues in the world. One of the ways to reduce stress and anxiety in everyday life is through practising mindfulness. Mindfulness involves paying attention (on purpose) to what is going on inside and outside ourselves, moment by moment, without judgement. Mindfulness is a technique that can be learned/taught and practised over time.

In your role as a designer at a design agency, you have been tasked with creating a mindfulness app that university students can use.

(*a*) Choose two user research methods that would be appropriate for gathering data for this project, and motivate your choice. Explain which user research methods discussed in the course would be difficult to use for gathering requirements, and describe why that is the case. [5 marks]

(*b*) Describe what stakeholder analysis is. Provide any assumptions you may need to make, and then identify and describe the stakeholders of the app to be developed. [4 marks]

(*c*) Describe what requirements analysis is. Identify four key requirements that the application must meet. [4 marks]

(*d*) Propose one product-specific design principle that this mindfulness app design should follow and motivate your proposition. [3 marks]

(*e*) Describe what prototyping is and explain two prototyping methods you would use and why you would use them while exploring the design space for this mindfulness app. [4 marks]

SECTION D

7 Machine Learning and Real-world Data

In an annotation task with 4 classes (I, II, III and IV), three annotators (A, B, C) are making decisions, as in Figure 1.

	A	B	C
Item1	III	III	I
Item2	IV	I	III
Item3	II	II	I
Item4	I	IV	IV
Item5	II	IV	II
Item6	I	I	I
Item7	IV	IV	III
Item8	II	I	II

(*a*) Raw agreement amongst $k > 2$ annotators can be calculated based on pairwise agreement. Explain how this can be done, and calculate the value in the above case, showing your workings. [4 marks]

(*b*) We now want to use a chance-corrected agreement metric and choose Kappa.

(*i*) Explain why chance-corrected agreement metrics are useful. [2 marks]

(*ii*) How is chance agreement in Kappa calculated? Give the formula and calculate the value in the case above. [2 marks]

(*iii*) Give the formula for Kappa and calculate its value in our situation. [2 marks]

(*c*) New annotated data is discovered, which stems from two other annotators. Annotatator D only participated in annotation from item3 onwards, whereas Annotator E stopped annotating after item8 due to sickness. We want to use their partial annotation data, together with that from annotators A-C.

(*i*) One possible treatment is to pretend that annotators D and E were a single person, by randomly discarding one judgement for the doubly annotated items. Give at least two reasons why this is problematic. [4 marks]

(*ii*) Adapt the Kappa metric given above so that it can deal with partial annotation data. Give the motivation behind your idea as well as a formula for the final metric. [4 marks]

(*iii*) The annotation is now parcelled out into small sections (2 items each) and moved to a crowd-sourcing platform. Describe at least one potential problem with your agreement metric from (*c*)(*ii*) in this setting. [2 marks]

8 Machine Learning and Real-world Data

You are the program chair of a large academic conference. When papers are submitted to the conference, your job is to decide which paper should go into which topical area. For instance, a paper on computer architecture should go into the "Hardware" area. The areas are fixed before submission takes place. For each area, you have recruited an expert that will organise reviews for each paper in their area. These experts are called "area chairs". You decide to use statistical classification to route the incoming papers into the various areas. You have at your disposal several decades of papers, labelled with the area they were manually assigned to.

(*a*) Explain how you can set up a Naive Bayesian classifier for this task and derive the required parameter estimates. Give all necessary formulae. [3 marks]

(*b*) You now want to quantify how well your classifier is doing. Given that you can ask your area chairs for instant feedback, which two different evaluation methods can you realise in your setting, and how would you do this? Your answer should give details about data split and metrics. [2 marks]

(*c*) Up to now, we have assumed that areas are stable across years. You now find out that for your upcoming conference, for the first time in the history of your field, some areas have been changed. For each of the cases below, explain what would happen if you simply ran your classifier from (*a*) unchanged in the new situation, and propose the best course of action in the light of your existing classifier, giving your reasons.

(*i*) An area has become unpopular and is no longer treated in this year's conference. [2 marks]

(*ii*) An entirely new area has been proposed, treating material never before covered in your conference. [2 marks]

(*iii*) An existing area has split into two new areas. [2 marks]

(*iv*) Two existing areas have been merged into one. [2 marks]

(*d*) Rank the four situations listed in (*c*) with respect to how damaging they are to your classification strategy, giving your criteria for ranking. [2 marks]

(*e*) You want to know which areas are most similar to each other, and you want to use data to answer this question. In addition to the textual data described above, you also have access to your papers' citation network. This means you know which papers cite which existing papers, although you may not have full text available for cited papers, only identifiers. How would you determine the similarity between areas, and how would you visualise the results? [5 marks]

9 Machine Learning and Real-world Data

You are a member of a thinktank advising the government on financial policy issues. You are asked to develop a model to predict the effect of inflation on interest rates. The government has various mechanisms of controlling inflation, and the banking sector responds to the changes in inflation rates.

The historical data collected has three options for changes in inflation, increase (`inc`), decrease (`dec`), hold (`hold`), and categorizes interest rates as high (`high`) medium (`med`), low (`low`). The data over 7 time periods is as follows:

timestep	1	2	3	4	5	6	7
inflation changes	hold	dec	dec	hold	inc	hold	dec
interest rates	med	med	low	low	med	high	med

You decide to use a first-order hidden Markov model (HMM), modelling the changes in inflation as the hidden states and the interest rate as the observations.

(a) Define and estimate the components of an appropriate HMM for this application, without smoothing. Assume that all hidden states are equally likely to start the sequence. Ignore the end state. [4 marks]

(b) Assume the interests rates are currently high. How do you reduce them in consecutive time periods from high, to medium and then to low? In other words, which is the most probable sequence of inflation changes that results in the sequence of observations: `high medium low`? [6 marks]

(c) Using the HMM parameters you have estimated, answer the following questions, explaining your answers.

(i) Is it possible for the interest rates to change from low to high in consecutive time periods?

(ii) Are low interests rates more likely to remain low or to increase to medium ones? [4 marks]

(d) Given that your goal is to build a realistic model of the relation between inflation changes and interest rates, describe three shortcomings of the HMM developed above. [6 marks]

END OF PAPER