UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

B.Sc. Examination 2015

Computing

IS53010A Data Compression

Duration: 2 hours and 15 minutes

Date and time:

There are five questions in this paper. You should answer no more than three questions. Full marks will be awarded for complete answers to a total of three questions. Each question carries 25 marks. The marks for each part of a question are indicated at the end of the part in [.] *brackets.*

There are 75 marks available on this paper.

Electronic calculators must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Question 1

(a) What is the distinction between *lossy* and *lossless* data compression? What is the aim of lossy compression in general terms? Give an example of real life data that are suitable for lossless compression. [5]

(b) Encode the string `AABACCABBAAACCC` following the LZW algorithm. Assume that the dictionary initially contains single characters A-F and occupies cells at 0–5 only. Demonstrate the content changes of the main variables and the dictionary. [8]

(c) A binary tree (0-1 tree) can be used to represent a code containing a few codewords of variable lengths. Consider each of the four codes for the characters A, B, C, D below and draw the binary tree for each code.

i. (0011, 0001, 110, 111)
ii. (110, 111, 0, 1)
iii. (0000, 001, 1, 0001)
iv. (0000, 0001, 001, 1)

For each tree drawn, comment on whether the code being represented by the binary tree is a prefix code, and justify your conclusion. [12]

Question 2

(a) Keith claims that the binary code (1, 01, 001, 010) is a prefix code since it satisfies the Kraft inequality. Check if the code indeed satisfies the Kraft inequality and explain what is wrong with Keith's claim. [7]

(b) One important step of the Arithmetic decoding algorithm is to update boundary values. Identify an update error in the Arithmetic algorithm below and correct the error. [5]

```
1. L <- 0 and d <- 1
2. If x is within [L,L+d*p1)
3.     then output s1, leave L unchanged, and
4.          set d<-d*p1
5.     else if x is within [L+d*p1, L+d)
6.          then output s2, set L<- L*d+p2 and d<-d*p2
7. If the_number_of_decoded_symbols
8.             < the_required_number_of_symbols
9.     then go to step 2.
```

(c) Following the simplified version of the LZ77 algorithm, demonstrate how to encode, step by step, the string `AABACCABBAAACCC`. Assume that the length of the history buffer $H = 6$ and the length of the lookahead buffer $L = 6$. [13]

Question 3

(a) Explain the meaning of the following terms: [4]

i. rate-distortion problem
ii. transformation

(b) A student has given the following answer as step (5) in her assignment of demonstrating the execution steps of an Adaptive Huffman encoding algorithm. Suppose the sequence of symbols to be encoded is CAAABB initially. Highlight your answer for step (5) in tracing the states (or values) of the input, output, alphabet and the tree structure on each step and give reasons if there is any difference between yours and her answer. Indicate if there is an error in her answer. [6]

```
(5)
read-input:   B
Output    :   h(DAG) ASCII("B")

A={A(3), C(1), B(1), DAG(0)}

Tree:       5
          /     \
       A(3)      2
               /    \
             1      C(1)
           /   \
         B(1)  DAG(0)
```

(c) Explain what is meant by a *minimum-variance Huffman code*. Demonstrate, with an example, what technique can be used to derive a minimum-variance Huffman code. You may focus on one step of the Huffman encoding algorithm. [5]

(d) Consider a binary source file consisting of characters A and B with probability of 0.2 for B. Discuss the cause of inefficiency of applying to this source the static Huffman compression algorithm. Demonstrate how the problem may be solved. [10]

Question 4

(a) Explain what is used to represent the so-called colour depth in a common RGB colour model. What is the value of the colour depth in a representation where three bytes are assigned to every pixel? If 247 distinct colours are required for an application, what is the smallest colour depth value required? Give your reasons. [5]

(b) Consider a source of the alphabet (A, B, C). David claims that the static Huffman coding algorithm can generate an *optimal* prefix code on the source when the probability distribution of the symbols is (1/3, 1/3, 1/3). Do you agree with him? Give your reasons. [5]

(c) Derive the *Reflected Grey Code* for each of the colour codes in decimal below. Explain why Reflected Grey Codes are regarded as a better representation than normal binary codes for coding greyscale images. [5]

9	10
10	11

(d) Decode the following string using the HDC algorithm. Explain the meaning of each control symbol used. What is the compression ratio? What is the entropy of the source? [10]

```
r4n1Ar2n6BB3322r31n30ABr3Cn2BC
```

Question 5

Consider the two 8×8 matrices (a) and (b) below that represent two image sources. The matrix entries represent the pixel colour values of the images. Show all your work in answering questions.

(a)

0	0	0	0	2	4	7	7
0	0	0	0	2	5	6	7
0	0	1	1	2	2	5	7
0	0	1	1	2	2	6	7
0	0	1	1	2	2	7	7
0	0	3	3	2	4	6	7
0	0	4	2	2	3	7	7
0	0	5	1	2	2	7	7

(b)

0	0	0	0	2	4	7	7
0	0	0	0	2	5	6	7
0	1	1	1	2	2	5	7
1	1	1	1	3	4	6	7
1	2	1	1	4	5	7	6
2	2	3	3	4	4	6	6
3	4	4	2	3	4	7	6
5	5	5	6	5	7	6	5

(a) What is the minimum number of bits required for fix-length coding for each of the two sources? [2]

(b) Draw a histogram (or bar-chart) to contrast the probability distribution of the two sources plotting probability values against pixel colour values. [4]

(c) Which source would contain more information on average? Justify your answer. [4]

(d) Compress the two sources applying the canonical minimum-variance Huffman encoding algorithm directly. On which source would Huffman coding achieve a better compression result? Justify your answers. [7]

(e) Demonstrate how a simple differential compression scheme on rows may be applied as a preprocessing before a runlength algorithm. On which source would your preprocessing scheme achieve a better compression result? Justify your answer. [8]