Lecture 2 – Compression Basics

PORTLAND STATE UNIVERSITY

Administrative

- HW1 – due now
- HW2 – now on web site
  - Due next Wednesday
- Programming Assignment 1
  - Will probably go out Monday or Wednesday

PORTLAND STATE UNIVERSITY

How big is HDTV?

- 2 hour movie in HD uncompressed:

$1920 \times 1080 \times 3 \times 30 \times 60 \times 120 = 1,343,600,000,000$

(RGB, fps, min, 2hrs)

25,000,000,000

↓ Blu-ray disc ≈ 25GB

27 discs ?

PORTLAND STATE UNIVERSITY

Preconceived notions…

WinZip on JPEG doesn't really do anything

on text files about $\frac{1}{2}$ size

video gets virtually nothing

PORTLAND STATE UNIVERSITY

Compression

- Ideal compression
  - Get the original data back * or a very good estimate of it
  - Want it as small as possible
- Time vs. space

With compression you can have time or space pick only one!

Use small amount of (computation) time → larger file

Use large amount of time → smaller file

More redundancy in data results in smaller file.

PORTLAND STATE UNIVERSITY

Types of Compression

- Lossless

Reversible compression

Compression/decompression results in the exact data being retrieved

Great for word processing files

Examples: WinZip Application (Lempel Ziv, Welch)

Huffman, Run length encoding, LZ, LZW

Arithmetic coding

$\frac{1}{2}$ to $\frac{1}{4}$ the original size 2-4:1 compression

PORTLAND STATE UNIVERSITY

Types of Compression

- Lossy

Throw away data to achieve compression

Want to minimize perceived loss of data

Images + video use this

audio - cd-audio - lossless

mp3 - lossy $\frac{1}{10}$ the size

Typical: 10:1 for audio
24:1 for images
25-50:1 for video
} dependent upon how much loss

PORTLAND STATE UNIVERSITY

Compression Performance

Compression performance measured as ratio of :

$$\text{Compression ratio} = \frac{\text{uncompressed size}}{\text{compressed size}}$$

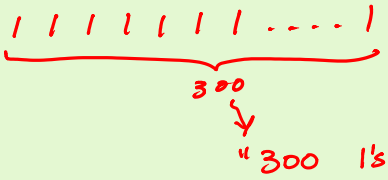
It is stated as N:1 compression
↑ compression ratio

Text file: 10000 bytes
↓ WinZip
↓ 2000 bytes compressed
} 5:1 compression

PORTLAND STATE UNIVERSITY

Compression Toolbox

- Encoding



- Transformation - massage data into form that can be readily encoded

1 2 3 4 300

difference

1 1 1 1 1

= diff

start w/ 1 + 299 1's difference

PORTLAND STATE UNIVERSITY

Compression Terminology

- Symbols - units of data in uncompressed domain

The basketball . . .

one symbol

- Codewords - the compressed domain entry for the symbol

- Codebook - collection of symbol → codeword mappings

01110110 → 101

"T" in ASCII

symbol

codeword

PORTLAND STATE UNIVERSITY

Textual Redundancy

spaces?

The small...

- Doesn't really occur in text files
  - ↳ some words appear more often
  - ↳ some letters appear more often
- graphical images (ASCII ART)
  - ↑ text
- databases, excel exported to text

PORTLAND STATE UNIVERSITY

Huffman Encoding

Takes symbols & creates codewords where smaller codewords are assigned to higher occurring symbols

Optimal for # of bits using a flat codebook
↑ one symbol → one codeword

PORTLAND STATE UNIVERSITY

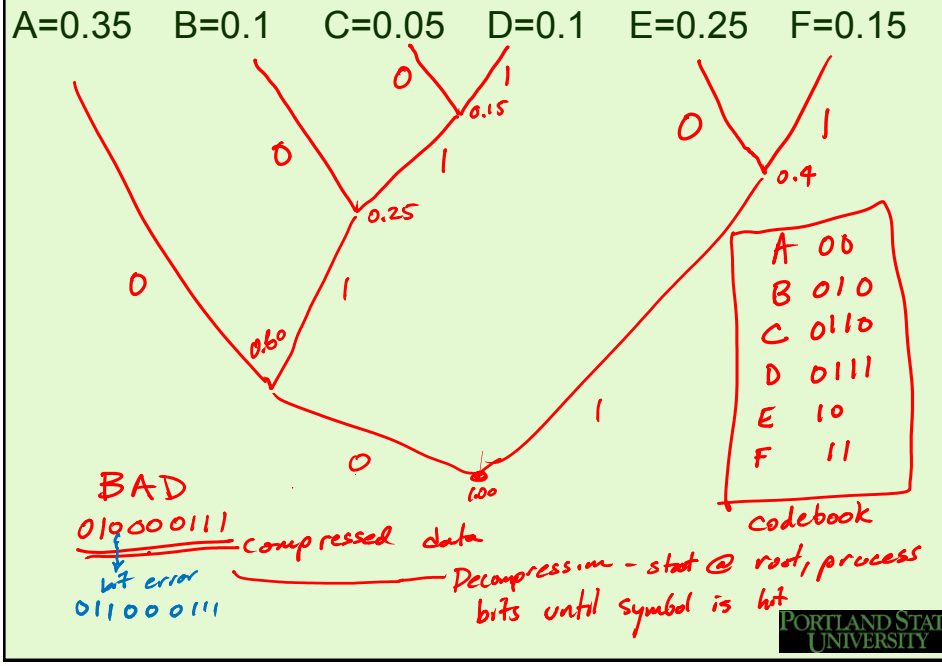
Huffman Algorithm

Given: Characters with either distribution or number of occurrences

1) Find smallest two values *in terms of probability*
2) Combine into a node
   - Mark new node with combined dist.
   - Assign 0 and 1 to branches
3) If all combined into a single tree, goto 4, otherwise go to step 1

4) Starting from the root, labels on the branches to a particular symbol make up the codeword

PORTLAND STATE UNIVERSITY



PORTLAND STATE UNIVERSITY

What's missing?

Need a codebook

PORTLAND STATE UNIVERSITY

Compression Fundamentals

- CF1 – Need to have an agreed upon format
  - Compressor + decompressor need to be in sync
  - Magic cookie "ZIP"
  - JPEG 11011101
  - standards specify these
- CF2 – Pick either
  - A highly compressed single instance ⇐ an optimized codebook
  - Generalized agreed upon codebook
    - ↳ distribution of English language as specified in standard

EXAMPLE - JPEG allows either

PORTLAND STATE UNIVERSITY

Huffman implementation issues

- Multiple same probabilities?

  Just pick one... it doesn't matter

- Uniqueness?

  Huffman tables (codewords) are not unique

  codeword lengths for a particular distribution will be the same

PORTLAND STATE UNIVERSITY

Huffman Compression

- Codebook required

  Implicit or added to file

- Susceptible to error (and propagation)

  Process one bit at a time. If wrong you end up on wrong branch of the tree

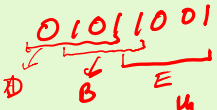
PORTLAND STATE UNIVERSITY

Huffman Compression

- Decoding is complex

We have to do this one bit at a time

A 101
B 1000
C 11
D 0
E 1001



- How do you speed it up?

Expand lookup table

Array index			Length				Length
0	0000	D	1	8	1000	B	4
1	0001	D	1	9	1001	E	4
2	0010	D	1	10	1010	A	3
3	0011	D	1	11	1011	A	3
4	0800	D	1	12	1100	C	2
5	0101	D	1	13	1101	C	2
6	0110	D	1	14	1110	C	2
7	0111	D	1	15	1111	C	2

PORTLAND STATE UNIVERSITY