ECE 499/599
Data Compression & Information Theory

Thinh Nguyen
Oregon State University

Adminstrivia

Office Hours

TTh: 2-3 PM Kelley Engineering Center 3115

Class homepage

- http://www.eecs.orst.edu/~thinhq/teaching/ece499/spring06/spring06.html

Adminstrivia

Textbook

Title: Introduction to Data Compression, third edition
Author: Khalid Sayood
Publisher: Morgan Kaufmann

Adminstrivia

Grade Policy

25% Homework
30% Midterm
5% Class participation
40% Final

Syllabus

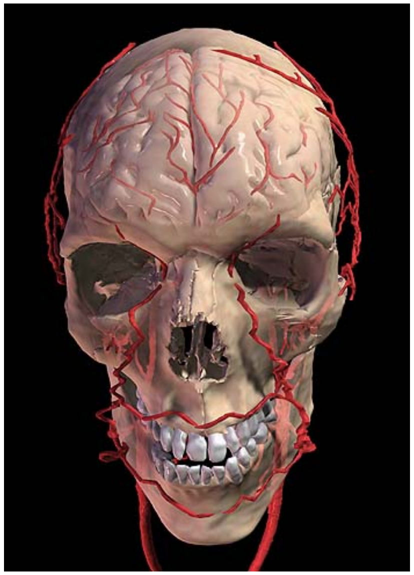
- Basic Information Theory
- Prefix Codes.
- Huffman Codes.
- Tunstall and Golomb Codes.
- Arithmetic Codes .
- Dictionary Codes: LZW, LZ77.
- Predictive coding and Burrows Wheeler.
- Lossy image compression and scalar quantization.
- Vector quantization.
- Nearest-neighbor search for VQ.
- Transform coding (DCT) and JPEG '87.
- Subband coding (wavelets) and SPIHT
- EBCOT and JPEG 2000.
- Intro to Video Coding and H.261/MPEG-1.
- Mpeg2 and Mpeg4.
- Audio and MP3's.

Why Compression?

- Multimedia applications generates a lot of data
  - Need to compress data for efficient storage
  - Need to compress data for efficient transmission.

Why Compression?

- Examples of applications that use compression.
  - Video: DVD, video conferencing
  - Image: JPEG
  - Audio: MP3
  - Text: Winzip
  - Visualization: 3D medical volume visualization



Compression is everywhere!

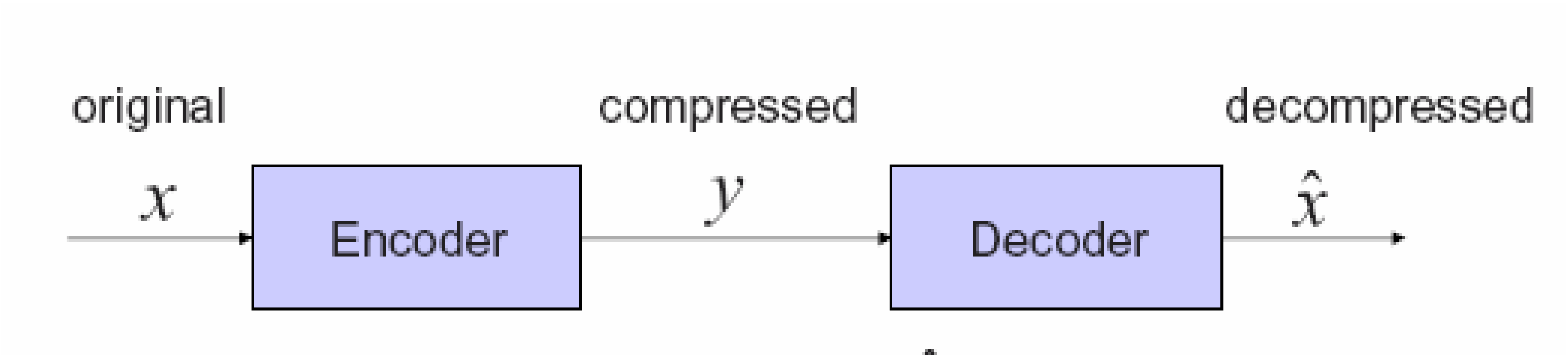
Why compression?

Speech	8000 samples/s	8 Kbytes/s
CD audio	44,100 samples/s, 2 bytes/sample, stereo	176 Kbytes/s
NTSC	30 fps, 640x480 pixels, 3 bytes/pixel	30 megabytes/s
Volume visualization voxels	30 fps, 1000x1000x1000 voxels, 3 bytes/voxels	90 gigabytes/s

Lecture 1:
Basic Compression Concepts

Thinh Nguyen
Oregon State University

Compression



- Lossless compression
  - Also called entropy coding, reversible coding.
- Lossy compression
  - Also called irreversible coding.
- Compression ratio = |x|/|y|
  - |x| is the number of bits in *x*.

Compression: Beware!

- Compression ratio = $|x|/|y|$
- Two ways to make the ratio larger:
  - Decrease the size of the compressed version.
  - Increase the size of the uncompressed version!

Compression Classification

Compression

Lossless

- Decoded data = original data
- Comp. ratio < lossy comp. ratio
- Eliminate redundancy
- Used where errors are not allowed, e.g, computer programs.
- LZ, JBIG

Lossy

- Decoded data ~ original data
- Comp. ratio > lossless comp. ratio
- Keeping important information
- Used where small errors are allowed, e.g, images, videos.
- JPEG, MPEG

Lossless Compression

- Data is not lost - the original is really needed.
  - text compression.
  - compression of computer binaries to fit on a floppy.
  - Compression ratio typically no better than 4:1
- Statistical Techniques:
  - Huffman coding.
  - Arithmetic coding.
  - Golomb coding.
- Dictionary techniques:
  - LZW, LZ77.
  - Burrows-Wheeler Method.
- Standards
  - Zip, bzip, GIF, PNG, JBIG, Lossless JPEG.

Lossy Compression

- Data is lost, but not too much:
  - Audio.
  - Video.
  - Still images, medical images, photographs.
  - Compression ratios of 10:1.

- Major techniques include:
  - Vector Quantization.
  - Wavelets.
  - Block transforms.

- Standards:
  - JPEG, JPEG 2000, MPEG (1, 2, 4, 7).

Why data compression possible?

- Redundancy exists in many places
  - Texts
    - Redundancy(German) > Redundancy(English)
  - Video and images
    - Redundancy (videos) > redundancy(images)
  - Audio
    - Redundancy(music) ? Redundancy(speech)
- Eliminate redundancy – keep essential information
  - Assume 8 bits per character
  - Uncompressed: aaaaaaaaab: 10x8 = 80 bits
  - Compressed: 9ab = 3x8 = 24 bits
- Reduce the amount of bits to store the data
  - Small storage, small network bandwidth, low storage devices.
    - Ex: 620x560 pixels/frame
    - 24 bits/pixel 1 MB
    - 30 fps 30 MB/s (CD-ROM 2x 300KB/s)
    - 30 minutes 50 GB

Why data compression possible?

- Always possible to compress?
  - Consider a two-bit sequence.
  - Can you always compress it to one bit?

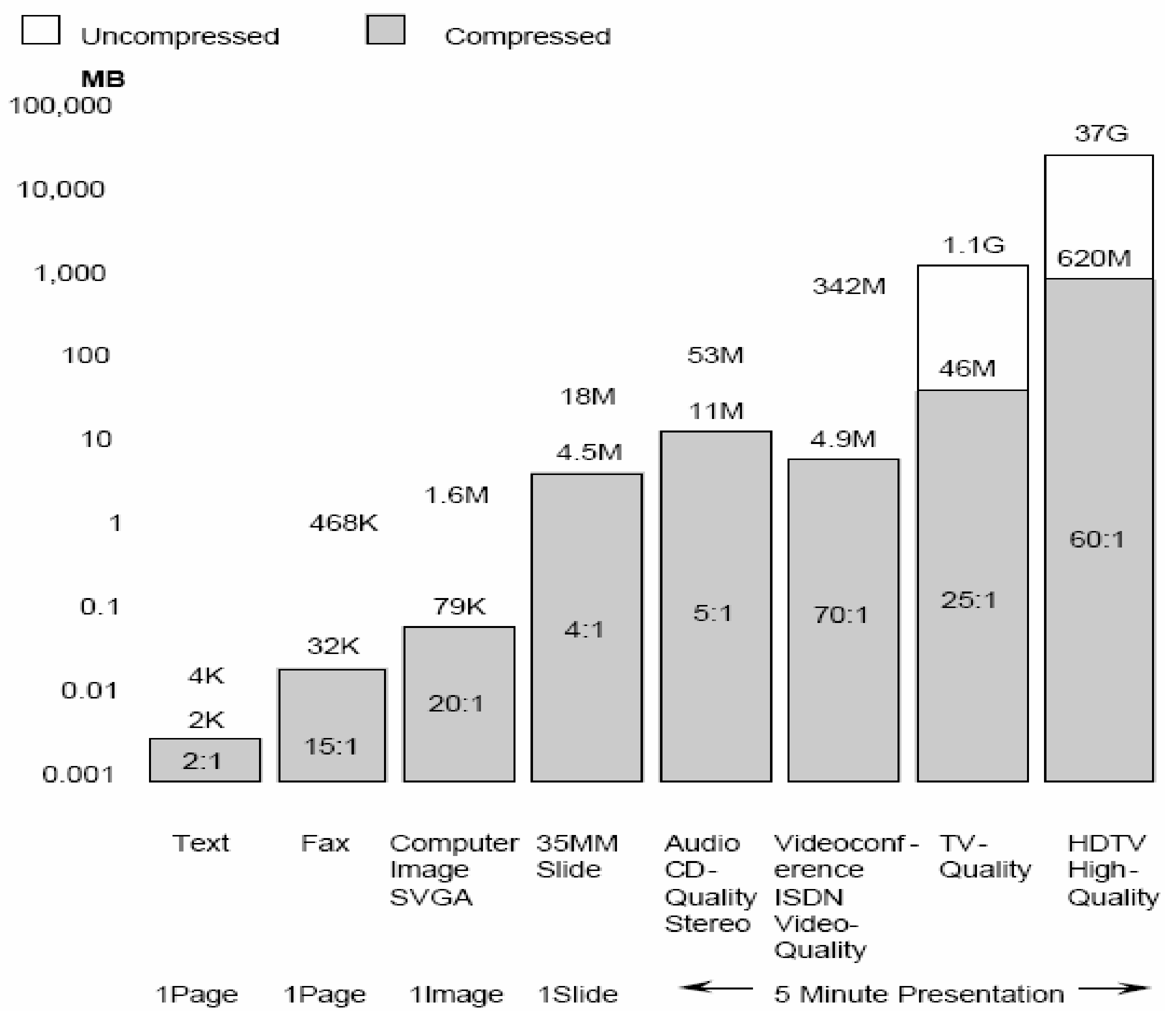
- Information theory is needed to understand the limits of compression and give clues on how to compress well. We will study information theory shortly!

Compression Techniques

- JPEG (DCT), JPEG-2000 (Wavelet)
  - Images
- JBIG
  - Fax
- LZ (gzip)
  - Text
- MPEG
  - Video

16:1 compression ratio

Typical Compression Ratios



Digital Representation of Data

- Digitization
  - Analog
  - Discrete Time
  - Digital
- Why digitize?
  - Universality of representation
  - Robustness to error, aging, distortion, noise

Digital Representation

Analog Signal

↓ Sample in time

Discrete Time Signal

↓ Quantize amplitude

Digital Signal

Advantages of Digital Representation

- Storage of different information types on the same devices -> easy integration of different media.
- Transmission of various information types over a single digital network.
- Processing and manipulation of various information by computer programs for editing, quality improvement, or recognition of meaningful information.

Disadvantages of Digital Representation

- Quantization distortion
- Sampling distortion (aliasing)
- Need large amount of digital storage capacity
- → Compression

- We will deal with only digital information in this class

Digital Representation

- Analog data:
  - Also called continuous data.
  - Represented by real numbers.
- Digital data:
  - Finite set of symbols {a1, a2, ..., an}.
  - All data represented as sequences (strings) in the symbol set.
  - Example: {a, b, c, d, r}: abracadabra.
  - Digital data can be an approximation to analog data.

Symbols

- Roman alphabet plus punctuation.
  - ASCII – 256 symbols.
- Binary – {0, 1}: 0 and 1 are called bits.
- All digital information can be represented in binary.
  - {a, b, c, d} fixed length representation:
  - a→00; b→01; c→10; d→11.
  - 2 bits per symbol.

Symbols

- Suppose we have n symbols. How many bits b (as a function of n) are necessary to represent a symbol in binary?
- What if some symbols occur more frequently than others, can we reduce the average number of bits to represent the symbols?