Information Theory and Coding

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Computer Science Tripos Part II

Michaelmas Term 2012 / 13

Professor John Daugman

Exercises and Supplementary Slides

JJ Thomson Avenue
Cambridge CB3 0FD

http://www.cl.cam.ac.uk/

Exercise Problems: Information Theory and Coding

Exercise 1

Prove that the information measure is additive: that the information gained from observing the combination of N independent events, whose probabilities are p_i for $i = 1....N$, is the *sum* of the information gained from observing each one of these events separately and in any order.

Exercise 2

(a) What is the maximum possible entropy H of an alphabet consisting of N different letters? In such a maximum entropy alphabet, what is the probability of its most likely letter? What is the probability of its least likely letter? Why are fixed length codes inefficient for alphabets whose letters are not equiprobable? Discuss this in relation to Morse Code.

(b) Let X and Y represent random variables with associated probability distributions $p(x)$ and $p(y)$, respectively. They are not independent. Their conditional probability distributions are $p(x|y)$ and $p(y|x)$, and their joint probability distribution is $p(x, y)$.

1. What is the marginal entropy $H(X)$ of variable X, and what is the mutual information of X with itself?
2. In terms of the probability distributions, what are the conditional entropies $H(X|Y)$ and $H(Y|X)$?
3. What is the joint entropy $H(X, Y)$, and what would it be if the random variables X and Y were independent?
4. Give an alternative expression for $H(Y) - H(Y|X)$ in terms of the joint entropy and both marginal entropies.
5. What is the mutual information $I(X; Y)$?

(c) Consider two independent integer-valued random variables, X and Y. Variable X takes on only the values of the eight integers {1, 2, ..., 8} and does so with uniform probability. Variable Y may take the value of *any* positive integer k, with probabilities $P\{Y = k\} = 2^{-k},\ k = 1, 2, 3, ...$.

1. Which random variable has greater uncertainty? Calculate both entropies $H(X)$ and $H(Y)$.
2. What is the joint entropy $H(X, Y)$ of these random variables, and what is their mutual information $I(X; Y)$?

Exercise 3

(a) Calculate the entropy in bits for each of the following random variables:

(i) Pixel values in an image whose possible grey values are all the integers from 0 to 255 with uniform probability.

(ii) Humans classified according to whether they are, or are not, mammals.

(iii) Gender in a tri-sexual insect population whose three genders occur with probabilities 1/4, 1/4, and 1/2.

(iv) A population of persons classified by whether they are older, or not older, than the population's median age.

(b) Let $p(x)$ and $q(x)$ be two probability distributions specified over integers x.

(i) What is the *Kullback-Leibler distance* between these distributions?

(ii) If we have devised an optimally compact code for the random variable described by $q(x)$, what does the *Kullback-Leibler distance* tell us about the effectiveness of our code if the probability distribution is $p(x)$ instead of $q(x)$?

(iii) Which axiom of distance metrics is violated by this distance?

(iv) What happens to this metric if there are some forbidden values of x for which $p(x) = 0$, and other values of x for which $q(x) = 0$?

Exercise 4

Consider an alphabet of 8 symbols whose probabilities are as follows:

A	B	C	D	E	F	G	H
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$

1. If someone has selected one of these symbols and you need to discover which symbol it is by asking 'yes/no' questions that will be truthfully answered, what would be the most efficient sequence of such questions that you could ask in order to discover the selected symbol?
2. By what principle can you claim that each of your proposed questions is maximally informative?
3. On average, how many such questions will need to be asked before the selected symbol is discovered?
4. What is the entropy of the above symbol set?
5. Construct a uniquely decodable prefix code for the symbol set, and explain why it is uniquely decodable and why it has the prefix property.
6. Relate the bits in your prefix code to the 'yes/no' questions that you proposed in 1.

Exercise 5

Suppose that X is a random variable whose entropy $H(X)$ is 8 bits. Suppose that $Y(X)$ is a deterministic function that takes on a different value for each value of X.

(i) What then is $H(Y)$, the entropy of Y?

(ii) What is $H(Y|X)$, the conditional entropy of Y given X?

(iii) What is $H(X|Y)$, the conditional entropy of X given Y?

(iv) What is $H(X,Y)$, the joint entropy of X and Y?

(v) Suppose now that the deterministic function $Y(X)$ is not invertible; in other words, different values of X may correspond to the same value of $Y(X)$. In that case, what could you say about $H(Y)$?

(vi) In that case, what could you say about $H(X|Y)$?

Exercise 6

Suppose that the following sequence of Yes/No questions was an optimal strategy for playing the "Game of 7 questions" to learn which of the letters $\{A, B, C, D, E, F, G\}$ someone had chosen, given that their *a priori* probabilities were known:

"Is it A?"	"No."
"Is it a member of the set $\{B, C\}$?"	"No."
"Is it a member of the set $\{D, E\}$?"	"No."
"Is it F?"	"No."

1. Write down a probability distribution for the 7 letters, $p(A), ..., p(G)$, for which this sequence of questions was an optimal strategy.
2. What was the uncertainty, in bits, associated with each question?
3. What is the entropy of this alphabet?
4. Now specify a variable length, uniquely decodable, prefix code for this alphabet that would minimise the average code word length.
5. What is your average coding rate R for letters of this alphabet?
6. How do you know that a more efficient code could not be developed?

Exercise 7

Consider a binary symmetric communication channel, whose input source is the alphabet $X = \{0, 1\}$ with probabilities $\{0.5, 0.5\}$; whose output alphabet is $Y = \{0, 1\}$; and whose channel matrix is

$$\begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

where ϵ is the probability of transmission error.

1. What is the entropy of the source, $H(X)$?

2. What is the probability distribution of the outputs, $p(Y)$, and the entropy of this output distribution, $H(Y)$?

3. What is the joint probability distribution for the source and the output, $p(X, Y)$, and what is the joint entropy, $H(X, Y)$?

4. What is the mutual information of this channel, $I(X; Y)$?

5. How many values are there for ϵ for which the mutual information of this channel is maximal? What are those values, and what then is the capacity of such a channel in bits?

6. For what value of ϵ is the capacity of this channel minimal? What is the channel capacity in that case?

Exercise 8

Consider Shannon's third theorem, the *Channel Capacity Theorem,* for a continuous communication channel having bandwidth W Hertz, perturbed by additive white Gaussian noise of power spectral density N_0, and average transmitted power P.

(a) Is there any limit to the capacity of such a channel if you increase its signal-to-noise ratio $\dfrac{P}{N_0 W}$ without limit? If so, what is that limit?

(b) Is there any limit to the capacity of such a channel if you can increase its bandwidth W in Hertz without limit, but while not changing N_0 or P? If so, what is that limit?

Exercise 9

1. An error-correcting Hamming code uses a 7 bit block size in order to guarantee the detection, and hence the correction, of any single bit error in a 7 bit block. How many bits are used for error correction, and how many bits for useful data? If the probability of a single bit error within a block of 7 bits is $p = 0.001$, what is the probability of an error correction failure, and what event would cause this?

2. What class of continuous signals has the greatest possible entropy for a given variance (or power level)? What probability density function describes the excursions taken by such signals from their mean value?

3. What does the Fourier power spectrum of this class of signals look like? How would you describe the entropy of this distribution of spectral energy?

4. Suppose that a continuous communication channel of bandwidth W Hertz and a high signal-to-noise ratio, which is perturbed by additive white Gaussian noise of constant power spectral density, has a channel capacity of C bits per second. Approximately how much would C be degraded if suddenly the added noise power became 8 times greater?

Exercise 10

1. Consider a noiseless analog communication channel whose bandwidth is 10,000 Hertz. A signal of duration 1 second is received over such a channel. We wish to represent this continuous signal exactly, at all points in its one-second duration, using just a finite list of real numbers obtained by sampling the values of the signal at discrete, periodic points in time. What is the length of the shortest list of such discrete samples required in order to guarantee that we capture all of the information in the signal and can recover it exactly from this list of samples?

2. Name, define algebraically, and sketch a plot of the function you would need to use in order to recover completely the continuous signal transmitted, using just such a finite list of discrete periodic samples of it.

3. Explain why smoothing a signal, by low-pass filtering it *before* sampling it, can prevent aliasing. Explain aliasing by a picture in the Fourier domain, and also show in the picture how smoothing solves the problem. What would be the most effective low-pass filter to use for this purpose? Draw its spectral sensitivity.

4. Consider a noisy analog communication channel of bandwidth Ω, which is perturbed by additive white Gaussian noise whose power spectral density is N_0. Continuous signals are transmitted across such a channel, with average transmitted power P (defined by their expected variance). What is the channel capacity, in bits per second, of such a channel?

5. If a continuous signal $f(t)$ is *modulated* by multiplying it with a complex exponential wave $\exp(i\omega t)$ whose frequency is ω, what happens to the Fourier spectrum of the signal?

   Name a very important practical application of this principle, and explain why modulation is a useful operation.

   How can the original Fourier spectrum later be recovered?

6. Which part of the 2D Fourier Transform of an image, the amplitude spectrum or the phase spectrum, is indispensable in order for the image to be intelligible?

   Describe a demonstration that proves this.

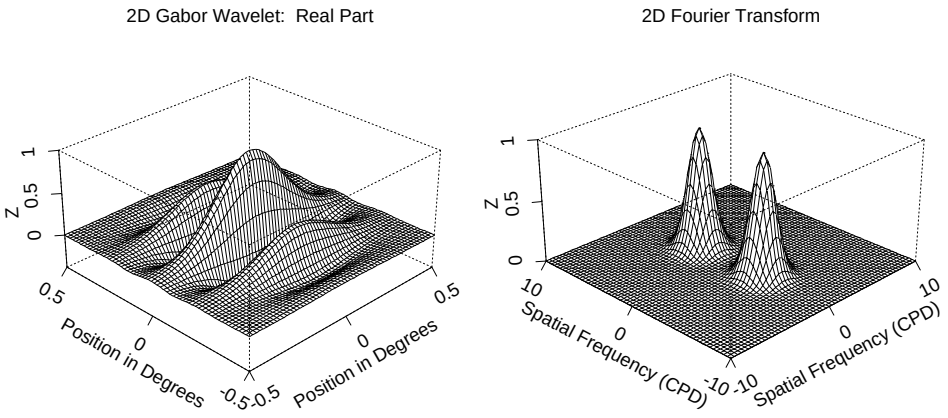
Exercise 11

The signal-to-noise ratio SNR of a continuous communication channel might be different in different parts of its frequency range. For example, the noise might be predominantly high frequency hiss, or low frequency rumble. Explain how the information capacity C of a noisy continuous communication channel, whose available bandwidth spans from frequency ω_1 to ω_2, may be defined in terms of its signal-to-noise ratio as a function of frequency, $\mathrm{SNR}(\omega)$. Define the bit rate for such a channel's information capacity, C, in bits/second, in terms of the $\mathrm{SNR}(\omega)$ function of frequency.

(Note: This question asks you to generalise beyond the material lectured.)

Exercise 12

(a) Explain why the real-part of a 2D Gabor wavelet has a 2D Fourier transform with two peaks, not just one, as shown in the right panel of the Figure below.



(b) Show that the set of all Gabor wavelets is closed under convolution, *i.e.* that the convolution of any two Gabor wavelets is just another Gabor wavelet. [HINT: This property relates to the fact that these wavelets are also closed under multiplication, and that they are also self-Fourier. You may address this question for just 1D wavelets if you wish.]

(c) Show that the family of sinc functions used in the Nyquist Sampling Theorem,

$$\mathrm{sinc}(x) = \frac{\sin(\lambda x)}{\lambda x}$$

is closed under convolution. Show further that when two different sinc functions are convolved, the result is simply whichever one of them had the lower frequency, *i.e.* the smaller λ.

(d) For each of the four classes of signals in the left table below, identify its characteristic spectrum from the right table. ("Continuous" here means supported on the reals, *i.e.* at least piecewise continuous but not necessarily everywhere differentiable. "Periodic" means that under multiples of some finite shift the function remains unchanged.) Give your answer just in the form 1-A, 2-B, etc. Note that you have 24 different possibilities.

Class	*Signal Type*
1.	continuous, aperiodic
2.	continuous, periodic
3.	discrete, aperiodic
4.	discrete, periodic

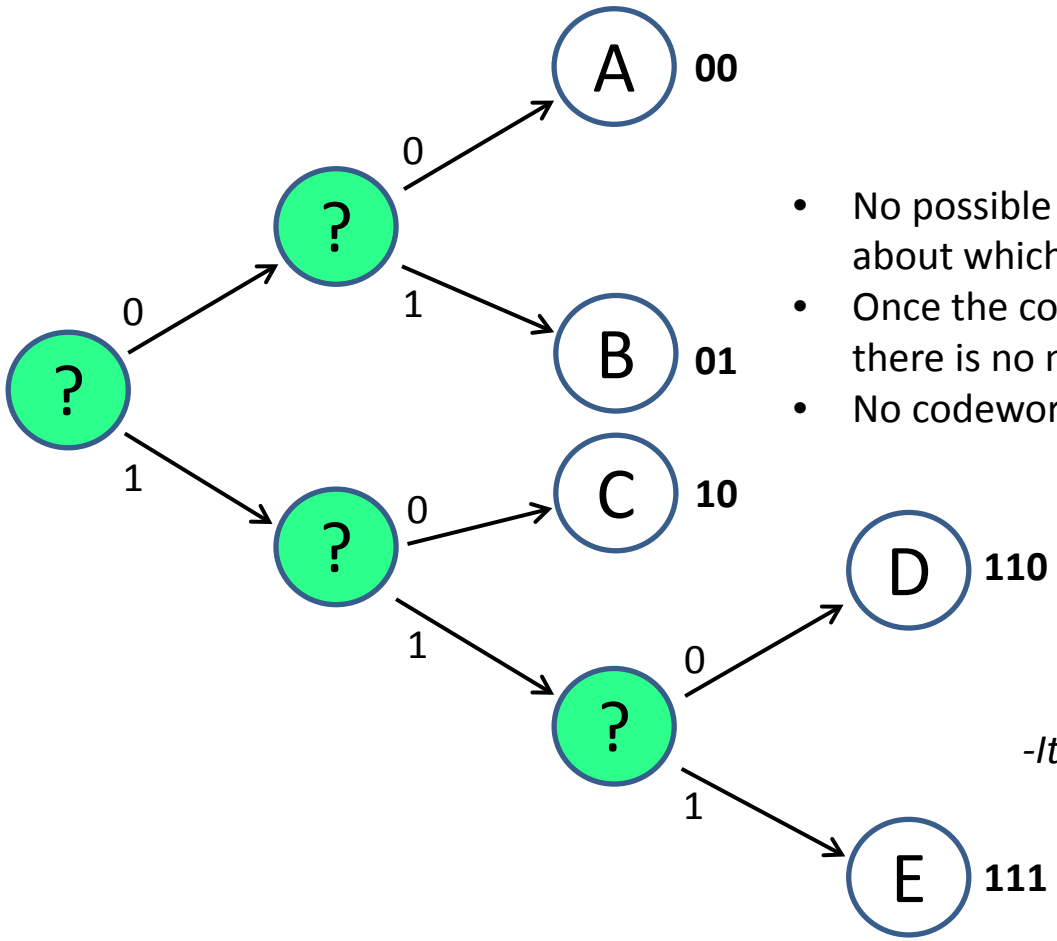
Class	*Spectral Characteristic*
A.	continuous, aperiodic
B.	continuous, periodic
C.	discrete, aperiodic
D.	discrete, periodic

(e) Define the Kolmogorov algorithmic complexity K of a string of data. What relationship is to be expected between the Kolmogorov complexity K and the Shannon entropy H for a given set of data? Give a reasonable estimate of the Kolmogorov complexity K of a fractal, and explain why it is reasonable.

Constructing variable-length symbol codes with desirable properties

Example of a uniquely decodable, instantaneous, prefix code over 5 letters {A,B,C,D,E}

p(A)=1/4
p(B)=1/4
p(C)=1/4
p(D)=1/8
p(E)=1/8



- No possible received string of bits is ambiguous about which symbols were encoded.
- Once the codeword for any symbol is received, there is no need to wait for more bits to resolve it.
- No codeword is a prefix of another codeword.

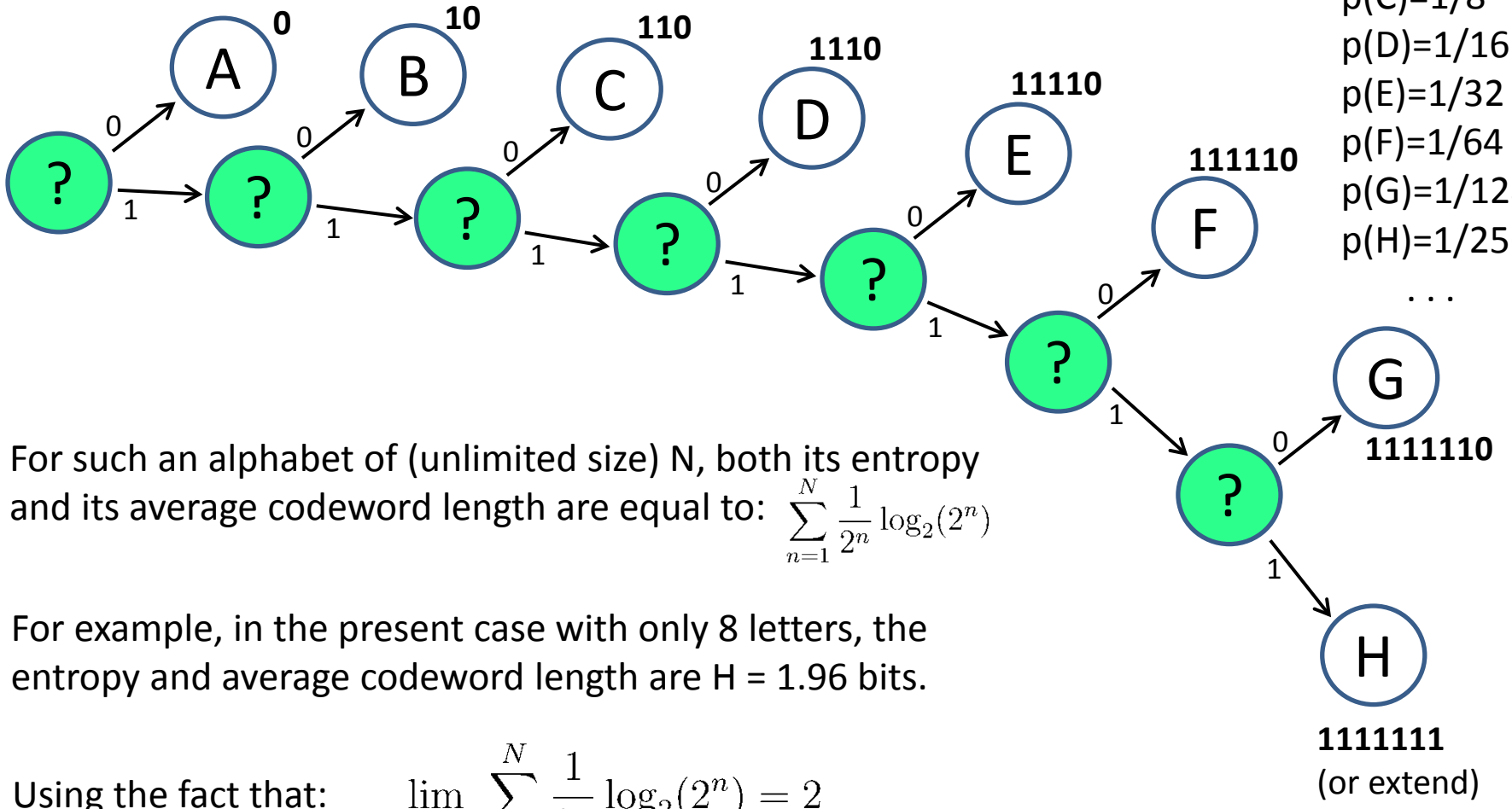
How efficient is this code?
-It achieves optimal Shannon efficiency:

Note the entropy of this alphabet is:
$3 \cdot (1/4) \cdot 2 + 2 \cdot (1/8) \cdot 3 =$ **2.25 bits**

Note the average codeword length is also:
$3 \cdot (1/4) \cdot 2 + 2 \cdot (1/8) \cdot 3 =$ **2.25 bits/codeword**

Example of an alphabet of unlimited size, with a special probability distribution, that can be uniquely encoded with average codeword length **< <u>2 bits/codeword</u>** !

p(A)=1/2
p(B)=1/4
p(C)=1/8
p(D)=1/16
p(E)=1/32
p(F)=1/64
p(G)=1/128
p(H)=1/256
. . .



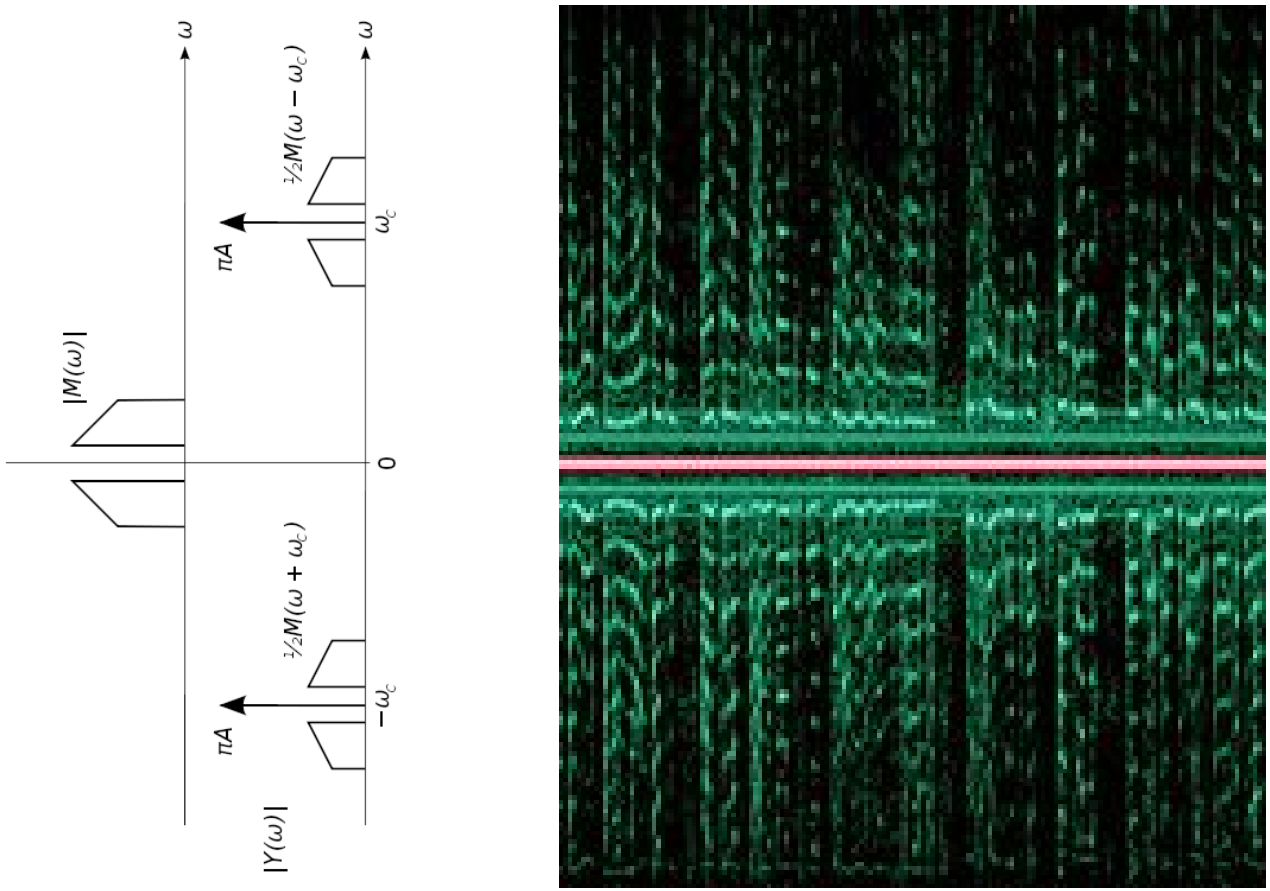
For such an alphabet of (unlimited size) N, both its entropy and its average codeword length are equal to: $\sum_{n=1}^{N} \frac{1}{2^n} \log_2(2^n)$

For example, in the present case with only 8 letters, the entropy and average codeword length are H = 1.96 bits.

Using the fact that: $\lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{2^n} \log_2(2^n) = 2$

we see that even if the size of this alphabet grows indefinitely, it can still be uniquely encoded with an average codeword length just below 2 bits/codeword.

Example of double-sideband modulation in AM broadcasting



Left: Double-sided spectra of baseband and (modulated) AM signals.
Right: Spectrogram (frequency spectrum versus time) of an AM broadcast shows its two sidebands (green), on either side of the central carrier (red).

A major application of the modulation property

The last two theorems are the basis for broadcast telecommunications that encode and transmit using **a**mplitude **m**odulation of a carrier (e.g. "AM radio"), for receivers that decode the AM signal using a tuner.

Radio waves propagate well through the atmosphere in a frequency range (or "spectrum") measured in the gigaHertz, with specific bands allocated by government for commercial broadcasting, mobile phone operators, etc. A band around 1 megaHertz (0.3 to 3.0 MHz) is allocated for AM radio, and a band around 1 gigaHertz (0.3 to 3.0 GHz) for mobile phones, etc.

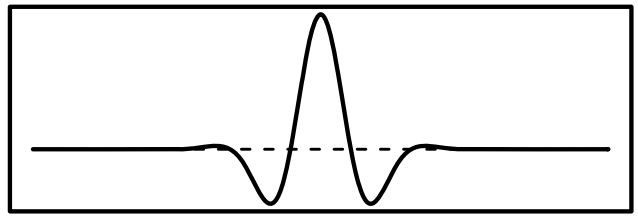
A human audio signal $f(t)$ occupies less than 10 kHz, but its spectrum $F(\omega)$ is shifted up into the MHz or GHz range by multiplying the sound waveform $f(t)$ with a carrier wave e^{ict} of frequency c, yielding $F(\omega - c)$. Its band**width** remains 10 kHz, so many many different channels can be allocated by choices of c. The AM signal received is then multiplied by e^{-ict} in the tuner, shifting its spectrum back down by c, restoring $f(t)$.

This ("single sideband" or SSB) approach requires a complex carrier wave e^{ict}. Devices can be simplified by using a purely real carrier wave $\cos(ct)$, at the cost of shifting in both directions $F(\omega - c)$ and $F(\omega + c)$ as noted, doubling the bandwidth and power requirements.

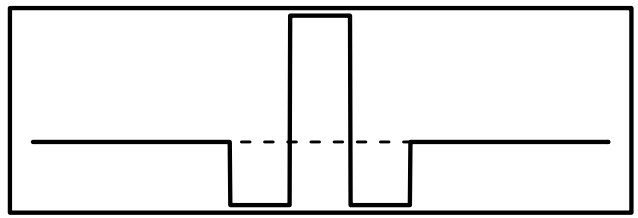
Gabor real and imaginary parts resemble Newton kernels in the calculus

Gabor Wavelets as 1st- and 2nd-order Differential Operators

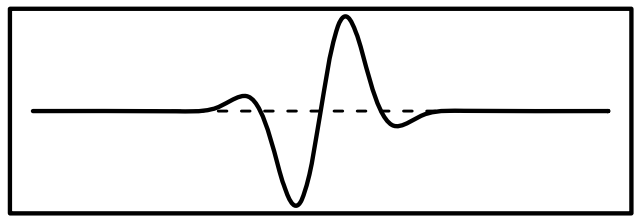
$$\mathbf{Re}\{e^{-x^2}e^{i3x}\} = e^{-x^2}\cos(3x)$$



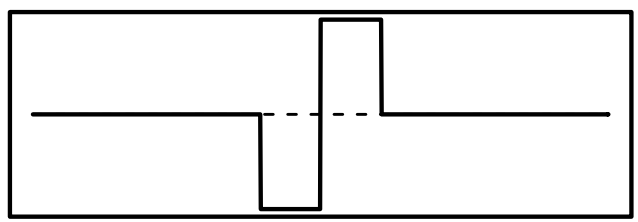
2nd finite difference kernel: $-f''(x_i)$
$\approx -f(x_{i-1}) + 2f(x_i) - f(x_{i+1})$



$$\mathbf{Im}\{e^{-x^2}e^{i3x}\} = e^{-x^2}\sin(3x)$$



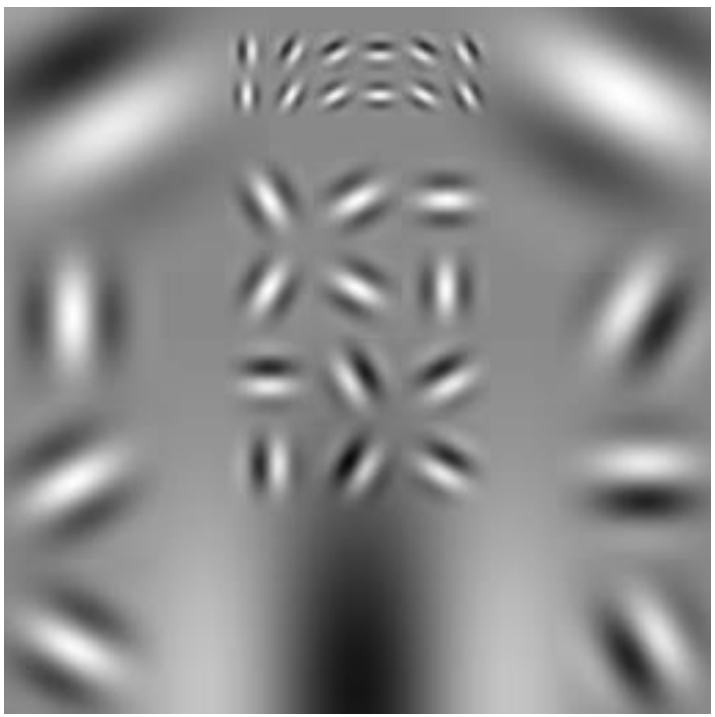
1st finite difference kernel: $f'(x_i)$
$\approx -f(x_i) + f(x_{i+1})$



Wavelets in computer vision and pattern recognition

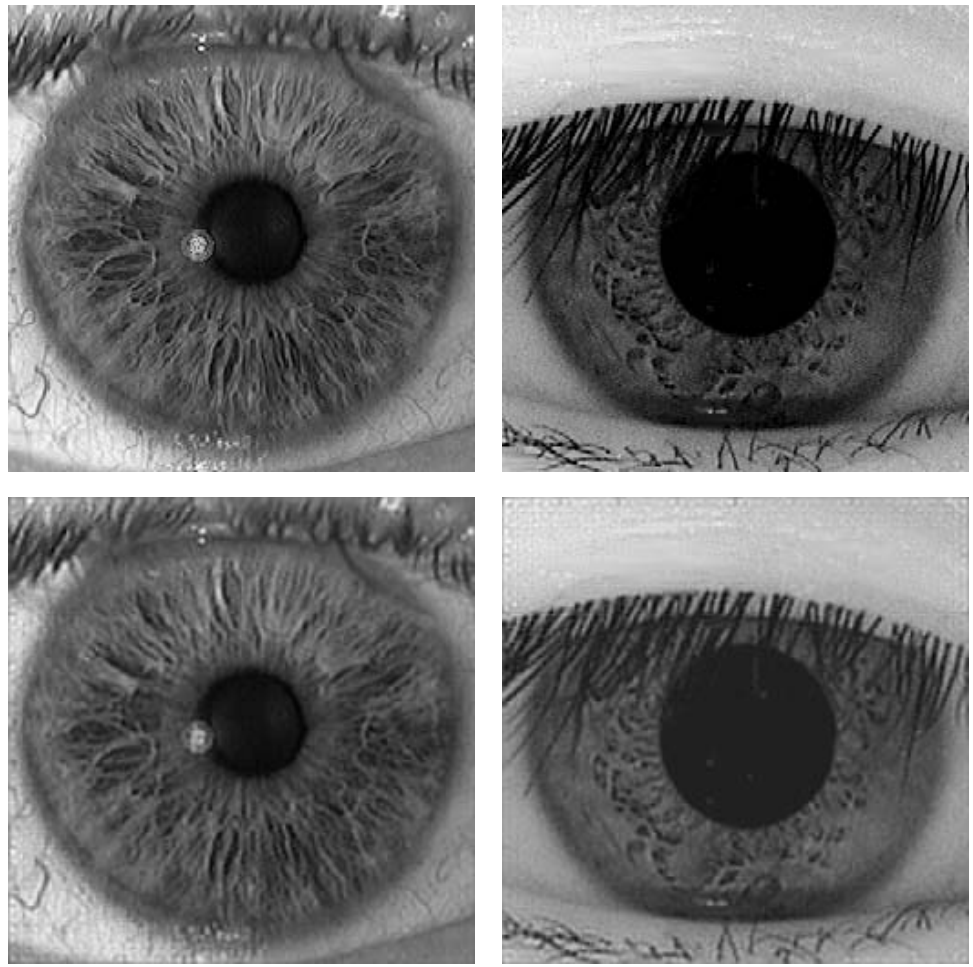
2D Gabor wavelets (defined as a complex exponential plane-wave times a Gaussian windowing function) are extensively used in computer vision.

As multi-scale image encoders, and as pattern detectors, they form a complete basis which can extract image structure with a vocabulary of: location, scale, spatial frequency, orientation, and phase (or symmetry). This collage shows a 4-octave ensemble of such wavelets, differing in size (or spatial frequency) by factors of two, having five sizes, six orientations, and two quadrature phases (even/odd), over a lattice of spatial positions.



Complex natural patterns are very well represented in such terms.

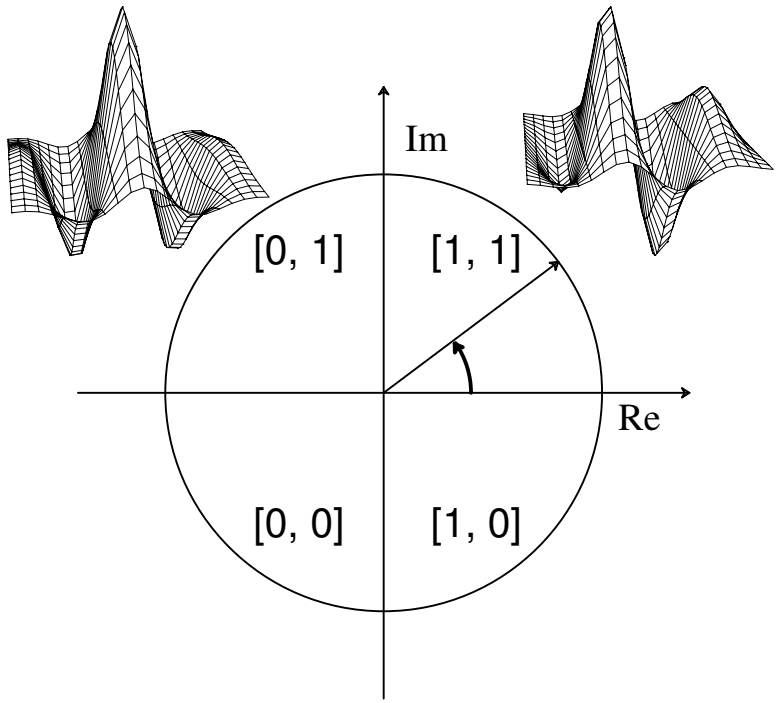
The upper panels show two iris images (acquired in near-infrared light); caucasian iris on the left, and oriental iris on the right.



The lower panels show the images reconstructed just from combinations of the 2D Gabor wavelets spanning 4 octaves seen in the previous slide.

Gabor wavelets are the basis for Iris Recognition systems

Phase-Quadrant Demodulation Code

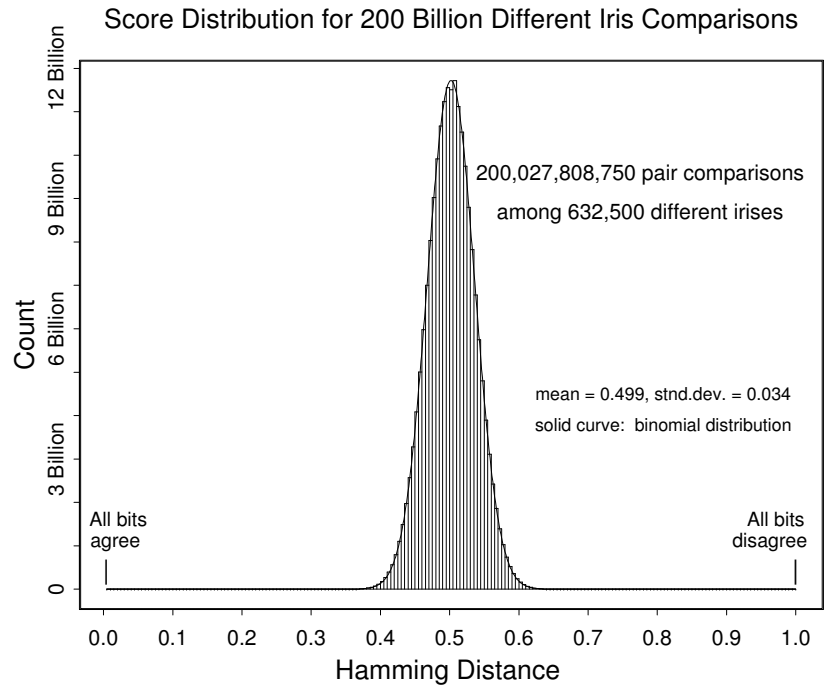


$$h_{Re} = 1 \text{ if } \mathrm{Re} \int_\rho \int_\phi e^{-i\omega(\theta_0-\phi)} e^{-(r_0-\rho)^2/\alpha^2} e^{-(\theta_0-\phi)^2/\beta^2} I(\rho,\phi)\rho d\rho d\phi \geq 0$$

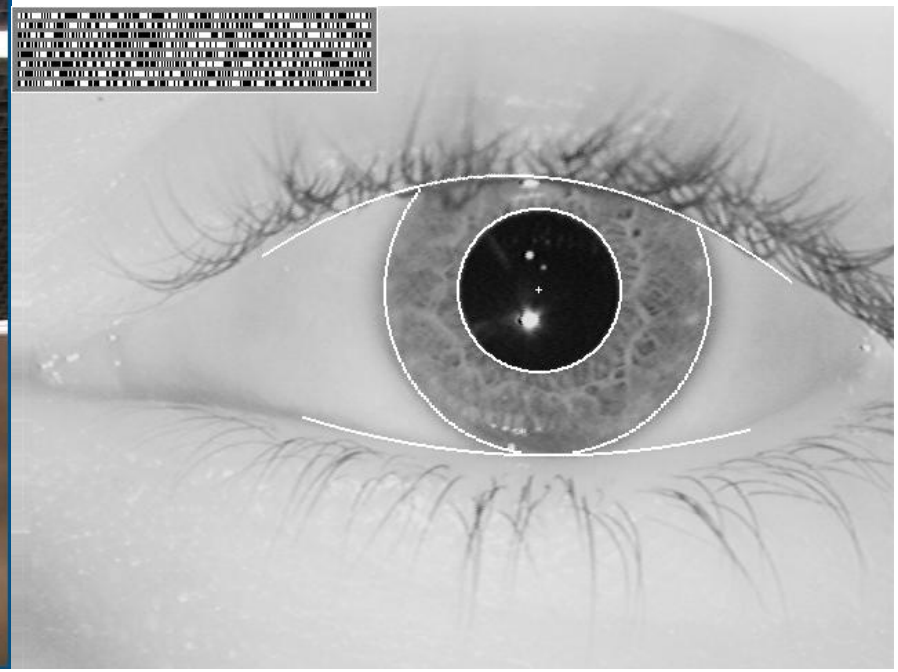
$$h_{Re} = 0 \text{ if } \mathrm{Re} \int_\rho \int_\phi e^{-i\omega(\theta_0-\phi)} e^{-(r_0-\rho)^2/\alpha^2} e^{-(\theta_0-\phi)^2/\beta^2} I(\rho,\phi)\rho d\rho d\phi < 0$$

$$h_{Im} = 1 \text{ if } \mathrm{Im} \int_\rho \int_\phi e^{-i\omega(\theta_0-\phi)} e^{-(r_0-\rho)^2/\alpha^2} e^{-(\theta_0-\phi)^2/\beta^2} I(\rho,\phi)\rho d\rho d\phi \geq 0$$

$$h_{Im} = 0 \text{ if } \mathrm{Im} \int_\rho \int_\phi e^{-i\omega(\theta_0-\phi)} e^{-(r_0-\rho)^2/\alpha^2} e^{-(\theta_0-\phi)^2/\beta^2} I(\rho,\phi)\rho d\rho d\phi < 0$$



Wavelets are much more ubiquitous than you may realize!



At 10 UK airport terminals, the **IRIS** system (Iris Recognition Immigration System) allows registered travellers to enter the UK without having to present their passports, or make any other claim of identity. They just look at an iris camera, and (if they are already enrolled), they cross the border within seconds. Similar programmes exist at border-crossings in many countries. The Government of India is currently enrolling the iris patterns of all 1.2 Billion citizens as a means to access entitlements (the UIDAI slogan is "To give the poor an identity"), and to enhance social inclusion.

Case study in image compression: comparison between patchwise Fourier (DCT) and wavelet (DWT) encodings

In 1994, the JPEG Standard was published for image compression using local 2D Fourier transforms (actually discrete cosine transforms [DCT] since images are real, not complex) on small $[8 \times 8]$ tiles of pixels. Each transform produces 64 coefficients and so is not itself a reduction in data.

But because high spatial frequency coefficients can be quantized much more coarsely than low ones for satisfied human perceptual consumption, a quantization table allocates bits to the Fourier coefficients accordingly. The higher frequency coefficients are resolved with fewer bits (often 0).

By reading out these quantized Fourier coefficients in a low-frequency to high-frequency sequence, long runs of 0's arise which allow run-length codes (Huffman coding) to be very efficient. $\sim$ 10:1 image compression causes little perceived loss. Both encoding and decoding (compression and decompression) are easily implemented at video frame-rates.

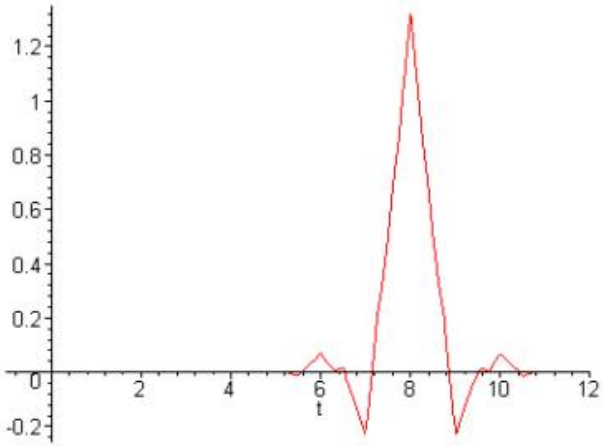
ISO/IEC 10918: *JPEG Still Image Compression Standard.*

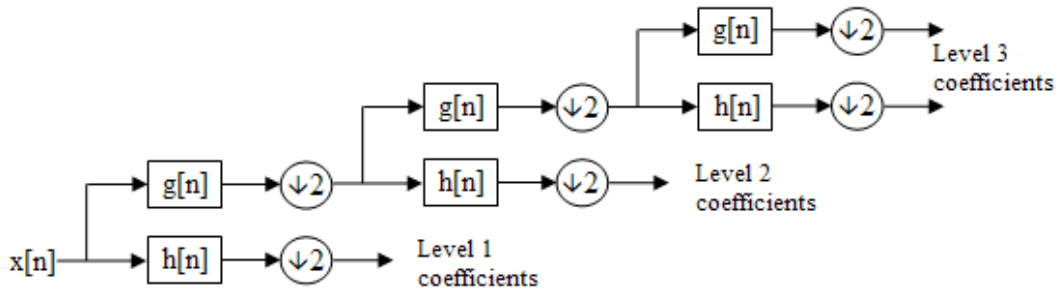
JPEG = Joint Photographic Experts Group `http://www.jpeg.org/`

(Image compression case study, continued: DCT and DWT)

Although JPEG performs well on natural images at compression factors below about 20:1, it suffers from visible block quantization artifacts at more severe levels. The DCT basis functions are just square-truncated sinusoids, and if an entire (8×8) pixel patch must be represented by just one (or few) of them, then the blocking artifacts become very noticeable.

In 2000 a more sophisticated compressor was developed using encoders like the Daubechies 9/7 wavelet shown below. Across multiple scales and over a lattice of positions, wavelet inner products with the image yield coefficients that constitute the Discrete Wavelet Transform (DWT): this is the basis of JPEG-2000. It can be implemented by recursively filtering and downsampling the image vertically and horizontally in a scale pyramid.





ISO/IE

15444: *JPEG2000 Image Coding System.* `http://www.jpeg.org/JPEG2000.htm`

Comparing image compressor bit-rates: DCT vs DWT

Whilst a monochrome `.bmp` image assigns 1 byte per pixel and thus has nominally a greyscale resolution of 8 bits per pixel [**8 bpp**], compressed formats deliver much lower **bpp** rates. These are calculated by dividing the total compressed image filesize (in bit count, not bytes) by the total number of pixels in the image. This benchmark image is uncompressed.

Comparing image compressor bit-rates: DCT vs DWT

Left: JPEG compression by 20:1 (Q-factor 10), **0.4 bpp**. The foreground water already shows some blocking artifacts, and some patches of the water texture are obviously represented by a single vertical cosine in an (8×8) pixel block.
Right: JPEG-2000 compression by 20:1 (same reduction factor), **0.4 bpp**. The image is smoother and does not show the blocking quantization artifacts.

Comparing image compressor bit-rates: DCT vs DWT

Left: JPEG compression by 50:1 (Q-factor 3), **0.16 bpp**. The image shows severe quantization artifacts (local DC terms only) and is rather unacceptable.
Right: JPEG-2000 compression by 50:1 (same reduction factor), **0.16 bpp**. At such low bit rates, the Discrete Wavelet Transform gives much better results.