

# Codecs: Encoding/decoding images and sounds

Adrian Mackenzie

[a.mackenzie@lancaster.ac.uk](mailto:a.mackenzie@lancaster.ac.uk)

October 2006

Codecs (coder-decoder) perform encoding and decoding on video, speech, music and text. They scale, reorder, decompose and reconstitute perceptible images and sounds so that they can get through information networks and electronic media. Codecs are intimately associated with changes in the “spectral density,” the distribution of energy, of sound and image in electronic media.

Codecs pose numerous analytical problems for software scholars. They are mathematically deep. Coming to grips with them may entail lengthy immersion in technical details. Although they are responsible for displaying many moving images on screens today, codecs themselves often hide in hardware and lower-level code. They come to light occasionally, usually in the form of an error message saying that something is missing: the right codec has not been installed. Despite or perhaps because of their convoluted obscurity, they catalyze new relations between people, things, spaces and times in events and forms.

## ***Patent pools and codec floods***

Video codecs such as MPEG-1, MPEG-4, H.261, H.263, the important H.264, theora, dirac, DivX, XviD, MJPEG, WMV, RealVideo, etc are strewn across networked electronic media. Roughly a hundred different audio and video codecs are currently available, some in multiple implementations. Because codecs often borrow techniques and strategies of processing sound and image from each other, they have tangled genealogies.

Leaving aside the tangle of relations between different codecs and video technologies, even one codec, the well-established and uncontentious MPEG-2 coding standard, is extraordinarily complex. MPEG-2 (a.k.a. H.262) designates a well-established set of encoding and decoding procedures for digital audio and video formalised as a standard (ISO/IEC 13818) in the mid 1990s. The standards for MPEG-2 are widely described. Many diagrams, definitions and explanations of coding and decoding the bitstream are available in print and online (Smith, 2001; Wikipedia, 2006). Open-source software implementations of the standard offer another way to examine their implementation. For instance, ffmpeg, 'is a complete solution to record, convert and stream audio and video' (ffmpeg, 2006). It handles many different video and audio codecs, and is widely used by many other video and audio projects (VLC, mplayer, etc).

Economically MPEG-2 is a mosaic of intellectual property claims (almost 700 patents held

by entertainment, telecommunications, government, academic and military owners according to Wikipedia (Wikipedia, 2006)). The large patent pool attests to the economic significance of MPEG-2 codecs. As the basis of commercial DVDs, the transmission format for satellite and cable digital television (DVB and ATSC), as the platform for HDTV as well as the foundation for many internet streaming formats such as RealMedia and Windows Media, MPEG-2 forms a primary technical component of contemporary audiovisual culture. It figures in geo-political codec wars (e.g. China's AVC codec versus the increasingly popular H.264 versus other versions such as Microsoft Windows VC-1 – Windows Media 9).

Many salient events in the development of information and digital cultures (for instance, MP3-base file-swapping, or JPEG-based photography) derive from the same technological lineage as MPEG-2 (lossy compression using transforms). At an embodied, what appears on screen or what we hear is coloured by the techniques of 'lossy compression' that MPEG-2 epitomizes. Codecs affect at a deep level contemporary sensations of movement, color, light and time.

### ***Trading space and time in transforms***

The MPEG standard is complex. Digital signal processing textbooks caution against trying to program to it at home (which immediately suggests the desirability of trying to). They suggest getting someone else's implementation of the standard (Smith, 2001, 225)]. Where does this complexity come from? The purpose of MPEG-2 as set out in the standards document is generic:

This Part of this specification was developed in response to the growing need for a generic coding method of moving pictures and of associated sound for various applications such as digital storage media, television broadcasting and communication. The use of this specification means that motion video can be manipulated as a form of computer data. (ISO/IEC 13818-2: 1995 (E), vi)

How does a 'generic coding method' end up being so complex that 'it is one of the most complicated algorithms in DSP [digital signal processing]' (Smith, 2001, 225)? MPEG-2 defines a bitstream that directly addresses the complicated psychophysical and technocultural processes of seeing. Codecs put more pictures, more often, in more places. It moves sounds and images further and faster in media networks than they would otherwise. It deeply reorganises relations within and between images and sounds, between things and experience.

However, to do that, video codecs trade between space and time at many scales. Algorithmically, MPEG-2 combines several distinct compression techniques (converting signals from time domain to frequency domain using Discrete Cosine Transforms, quantisation, Huffman and Run Length Encoding, block motion compensation), timing and multiplexing mechanisms, retrieval and sequencing techniques, many borrowed from

the earlier, low-bitrate standard, MPEG-1 (ISO/IEC 11172-1:1993). This tradeoff impacts heavily on images. It changes what becomes of them. Only a sample of the many processes in the codecs can be discussed here. I will concentrate on what happens at the lowest levels of the picture, the block (8 x 8 pixels). The key areas of interest for the purposes of seeing the trade-offs are the encoding and decoding sections of the software.

Digital video typically arrives at the codec as a series of frames. Each frame comprises arrays of pixel-level luminance and chrominance values. Each frame then undergoes several phases of encoding and decoding. These phases probe and re-structure of the image quite deeply, almost to the pixel level. Digital video pictures are composed of 2D arrays of pixels that have much spatial redundancy (that is, many adjacent pixels will be very similar). One priority is to express the spatial distribution of luminance (the brightness or amount of light emitted) and chrominance (the two signals that encode color information) as efficiently as possible. The so-called I-Picture or Intra-Picture coding, the first phase of encoding, is based on spectral analysis. Many video and audio codecs today rely on Fourier Transforms or, because it is easier to program and concentrates the energy of the signal into a smaller number of coefficients, on a particular variant of the Fourier Transform, the Discrete Cosine Transform (DCT). Fourier Transform or spectral analysis methods encompass a very wide range of computational problems. These methods decompose signals that vary over time or space into a spectrum of component frequencies that can be summed together to reconstitute the original signal. Nearly all video codecs transform spatially extended images into sets of frequencies. This allows them to isolate those components of a sound or image that are most perceptually relevant to human eyes and ears. (More recent codecs such as the British Broadcasting Corporation's open-source Dirac are gradually replacing Fourier-based transforms with 'wavelet'-based transforms because they take less computation on the whole.)

There is something quite counter-intuitive in transform compression. The notion of the transform is mathematical: it is an operation that takes an arbitrary signal and analyses it as a series of sinusoids of different frequencies and amplitudes. Added together, these sine or cosine waves re-constitute the original signal. For the coding of a given picture, the I-Picture results from division of the luminance and chrominance arrays into 8 x 8 blocks. The DCT applied to each of these spatial blocks in isolation produces a series of coefficients (or multipliers) of different frequency cosine waves that range from low to high frequencies. The cosine wave coefficients represent amplitudes of different frequency cosine waves. The coefficients sketch distributions of light in the image. This means that the luminance and chrominance values of an image are compressed, transmitted/stored and decompressed ever sending any information about individual pixels. The codec discards low value coefficients that describe individual pixel differences. It keeps the high value coefficients that express more energetic components of the signal. It subjects these coefficients to further compression using quantisation (that is, reducing them to a subset of discrete values) and 'entropy encoding' (that is, using standard compression techniques such as Huffman coding). When the block is decoded (for instance, during display of a video frame on screen), the coefficients are re-attached to corresponding cosine waves, and these are summed together to re-constitute arrays of values for luminance and chrominance comprising the block.

In the source code for codecs such as theora, the density of matrix or array manipulation stands out in the transform phase of the encoding. Thousands of blocks in each picture undergo DCT. In contrast to film's use of linear sequences of frames, or television and video's interlacing of scan-lines to compose images, transforms such as DCT subject images to highly intricate reorderings. Since blocks themselves are not fragments of pictures, but rather distributions of luminosity and chrominance, they are put into the bit stream – the data that flows out of the encoder - in sometimes quite strange order, an order that bears little direct relation to the displayed image.

### ***Motion prediction - forward and backwards in time***

Software has long been understood as closely linked to ideation or thought, particularly mathematical thought. Despite the mathematical technicalities of the transform compression, the thinking present in software cannot be reduced to mathematical thought, or at least, mathematical thought as it is usually conceived. Codecs perhaps bear a closer relation to cinematic thought and memory. In their handling of images, they deviate radically from normal understandings of representation. Video codecs are very preoccupied with the relations between pictures ('frames'). Indeed just as pictures themselves are individually analysed as distributions of luminance and chrominance values, video codecs relate pictures to each in terms of motion vectors.

MPEG video never flickers. It sometimes gets a bit 'blocky.' This is because the boundaries between pictures are not fixed in the same way they are in film frames or even in television with its 'interlaced' scanned fields. P (forward prediction) and B (backward prediction) pictures come after the transform-encoded I-picture in a MPEG-2 bitstream. However, this succession is cinematic. The working assumption behind the production of P and B pictures is that nothing much happens across successive frames that can't be understood as macroblocks (usually 4 block together) undergoing linear geometric transformations (translation, rotation, skewing, etc). The fact that nothing much happens between frames apart from geometric transformation is used as the basis of the inter-picture compression. Intra-picture compression of the space of the image is the first major component of MPEG-2. Motion prediction between frames, or time compression, is the second.

Inter-picture compression relies on forward and backwards correlations between the unencoded frames. It calculates motion vectors for every part of the image. For each frame, the MPEG-2 codec analyzes which parts have moved in comparison to the previous or later frame. It only transmits lists of motion vectors describing the movement of blocks in relation to a keyframe or reference picture. This fundamentally alters the character of frames (which is why the MPEG standard calls frames 'pictures'). We have already seen that rather than the raw pixel being the elementary material of the image, the block becomes the elementary component. In motion prediction, the frame is no longer the elementary component of movement, but an object to be cut up and sorted into sets of

motion vectors describing relative movements of blocks. The 'picture' after encoding is nothing but a series of vectors describing what happens to blocks. At the decoding end, an MPEG decoder turns the streams of vectors back into arrangements of blocks moving between frames.

### ***From complicated to composite***

All this is a bit complicated. Motion prediction takes time. The ratio of intra-frame and inter-frame pictures in a given bitstream depends on where the encoding is done and the bandwidth of the expected transmission channel (DVD, 3G cellphone, satellite digital TV, HDTV, the internet, etc). In an MPEG data-stream, the precise mixture of different frame-types (I, P-forward and B-backward) is defined at encoding time in the Group of Pictures (GOP) structure. It is typically 12 or 15 pictures in a sequence such as I\_BB\_P\_BB\_P\_BB\_P\_BB\_P\_BB\_. One intra-coded picture is followed by a dozen or so block motion-compensation picture. Their order in the bitstream does not correlate directly with the order of display. The combination of forward-prediction and backward-prediction found in the GOP means that the MPEG bitstream effectively treats images as massive doubly-linked list (Knuth, 1997, 280). The ratio of different frame types to each other affects the encoding time because motion compensation is much slower to encode than the highly optimised block transforms. Codecs must make direct tradeoffs between computational time and space. The tradeoffs sometimes result in artifacts visible on screen as such as blocking and mosaic effects. At times, motion prediction does not work. A change in camera shot, the effect of an edit, might mean that no blocks are shared between adjacent frames. In that case, the codec falls back on intra-frame encoding.

Many of the complications of the codecs arise because they link very different scales of technological infrastructure, markets and embodied cultural practice. Codecs connect network bandwidth constraints (a commercially marketed service), conventions of spectatorship, embodied cognition, and media-historical forms. Codecs respond to the economic-technical capping of bandwidth in telecommunications markets. Their teeming patent pools reflect high estimates of their value. The 'generic method' of encoding and decoding images for transmission relates very closely to the constraints and conditions of telecommunications and media networks. As a convention, the MPEG-2 standard cites explicitly or implicitly a great number of physical quantities ranging from screen dimensions, resolution and colour models through to network and transmission infrastructures to the clock rates and memory sizes of semiconductor and data storage technologies. Yet the codec must propagate light, colour and sound on screen within calibrated psycho-perceptual limits. Analysing such intersections requires ways of articulating diverse realities. Software like codecs might offer places to begin understanding what happens when passages between different scales and orders multiply.

## **References**

ffmpeg (2006), FFmpeg Multimedia System, <http://ffmpeg.sourceforge.net/index.php>, [accessed 4 Feb 2006]

ISO/IEC 13818-1, I. I. (1995). Information technology - Generic coding of moving pictures and associated audio information: Systems

ISO/IEC 13818-2 (1995) Information technology - Generic coding of moving pictures and associated audio information: Video

Kittler, F. (1993) Draculas Vermachtnis Technische Schriften, Reclam Verlag, Leipzig, pp. 182-207.

Knuth, D. E. (1997) The art of computer programming, Addison-Wesley, Reading, Mass.

Smith, S. W. (2001) The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing.

Wikipedia (2006), MPEG-2, <http://en.wikipedia.org/wiki/MPEG-2>, [accessed 12 Jan 2006]