

Codecs: Encoding/decoding images and sounds

Adrian Mackenzie

a.mackenzie@lancaster.ac.uk

Codecs perform encoding and decoding on a data stream or signal, usually in the interest of compressing data. They scale, reorder, decompose and reconstitute perceptible images and sounds in information networks and electronic media. They are intimately associated with changes in the “spectral density,” the distribution of energy, transported by sound and image in electronic media. Software such as codecs pose serious analytical problems. On the one hand, they are monstrously complex. Coming to grips with them may entail tedious excursions into technical details. They tax one’s affection for complexity. On the other hand, although they are responsible for displaying most images on screens today, codecs themselves are relatively concealed entities. Only occasionally, when they come to light, it is usually in the form of an error message that something has broken down: the right codec has not been installed. Despite or perhaps because of their convoluted banal obscurity, they are critically important catalysts of people, things, spaces and times in events and forms (Rabinow, 1999, 180).

Codecs: sensation and distribution

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). The standard in fact defines a ‘transport’ system rather than just a codec. Standards, as sociological work has argued, mix physical entities with conventional arrangements (Bowker and Star, 1999, 39). Software implementations of such standards are known as codecs (coder-decoder). Video codecs for different standards (MPEG-1, MPEG-4, H.261, H.263, the important H.264, theora, dirac, DivX, MJPEG, WMV, RealVideo, etc) are strewn across the milieu of sound and image associated with networked electronic media. Because codecs often borrow techniques and strategies of processing sound and image, their genealogy is tangled. 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. Algorithmically it 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). Economically it is a mosaic of intellectual property claims (640 patents held by entertainment, telecommunications, government, academic and military owners according to Wikipedia (Wikipedia, 2006)), and geo-political competition (e.g. China's AVC codec versus the increasingly popular H.264 versus other versions such as Microsoft Windows VC-1 – Windows Media 9). The economic significance of MPEG-2 codecs can hardly be overstated. 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. 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). What appears on screen or what is heard increasingly depends on the techniques of 'lossy compression' that MPEG-2 epitomizes. It generates artifacts (motion blocking, mosquito edging, etc) that affect at a deep level contemporary sensations of movement, color, light and time.

Trading space and time in transforms

The complexity of the codec is such that digital signal processing textbooks caution against trying to program to it at home (which immediately suggests the desirability of trying to) and suggest buying someone else's implementation of the standard (Smith, 2001 , 225)]. Yet the purpose of the MPEG-2 standard set out in ISO/IEC 13818 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 and can be stored on various storage media, transmitted and received over existing and future networks and distributed on existing and future broadcasting channels.
ISO/IEC 13818-2: 1995 (E), vi

How does a generic coding problem 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 and intervenes in a complicated physical process. It allows sounds and images to move further and faster in media networks than they would otherwise. It deeply reorganises relations within and between images and sounds, between things and experience.

Software transforms, compromises and re-organizes the space and time of communication. Codecs mean more pictures, more often, in more places. To do that, however, they trade between space and time at many scales. MPEG's impact on an image illustrates how why this trade-off occurs. Because only a sample or slice of the 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 standards for MPEG-2 are widely described, and contain diagrams, definitions and explanations of coding and decoding the bitstream (Smith, 2001; Wikipedia, 2006). Open-source software implementations of the standard offer one somewhat arduous, but very concrete path into 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). The key areas of interest for the purposes of seeing the trade-offs are the encoding and decoding sections of the software, found in `ffmpeg` libraries called `libavformat` and `libavcodec`.

In an MPEG-2 video stream, images typically arrive at the codec as pixel-level luminance and chrominance values, and then go through several phases of encoding and decoding. These phases probe and re-structure of the image quite deeply, almost to the pixel level. The so-called I-Picture or Intra-Picture coding, the first phase of encoding, is based on spectral analysis. The very name of the software mentioned above, 'ffmpeg', foregrounds the core feature of many video and audio codecs today: their reliance 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). In video and audio codecs, the DCT functions as a primary way of compressing and uncompressing images or sound. Fourier Transform or spectral analysis methods encompass a very wide range of computational problems in which data can be more easily analysed and by analysing signals that vary over time or space into a spectrum of frequencies that can be summed together to reconstitute the original signal. Nearly all video codecs rely on transformations between time or space and frequency domains to remove spatial and temporal redundancy from signals, and to capture those components of a sound or image that are most perceptually relevant to human eyes and ears. (More recent codecs such as Dirac are gradually replacing Fourier-based transforms with 'wavelet'-based transforms because they take less computation on the whole.)

The notion of the transform is mathematical: it is an operation that takes an arbitrary waveform 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. Given that 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 task is to capture 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. 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. In a convergence of thing and sensation familiar in the history of technical media, this the codecs decomposition of a spatial or temporal signal into different frequency components directly correlates with the neurophysiological understandings of human hearing and sight as a kind of spectral analysis.

By discarding high frequency components of the series, the luminance and chrominance values of a block can be compressed, transmitted/stored and decompressed ever sending any pixels bits. The cosine wave coefficients represent amplitudes of different frequency cosine waves. These coefficients can themselves be compressed using quantisation (that is, reduced to a subset of discrete values) and then 'entropy encoded' (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.

What stands out in the ffmpeg code is the densely complex matrix or array manipulation occurring on the thousands of blocks in a picture. 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 deal with grids of blocks in highly counter-intuitive movements. Blocks themselves are not fragments of pictures, but rather distributions of luminosity and chrominance that are packed into the bit stream in sometimes quite counterintuitive order. For instance, because the transform treats blocks as spectra of values, some of which are more significant to human eyes than others, it converts the spectrum values into a sequence in which the most important come first. But this necessitates, for reasons that are slightly too detailed for present purposes, moving through the array of the block values in a 'serpentine pattern' (Smith, 2001, 500).

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 just discussed, 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 since they radically abstract from normal understandings of images. Video codecs are very preoccupied with the relations between pictures ('frames'). Indeed just as pictures themselves are individually treated as tables of luminance and chrominance blocks, the relation between pictures is treated as tables of 'macroblocks' (usually 4 blocks put together) in motion. P and B pictures, the 'pictures' that come after the I-picture in a MPEG-2 bitstream, are really nothing like film frames. There will never be a flicker in an MPEG video because the boundaries between pictures are not fixed in the same way they are in film or even in television with its 'interlaced' scanned images. The working assumption behind production of P and B pictures is that nothing much happens between successive frames that can't be understood as macroblocks undergoing spatial transformations (translation, rotation, skewing, etc). The fact that nothing much happens between frames apart from spatial transformation is used as the basis of the inter-picture compression and the generation of P and B pictures (forward and backward motion prediction respectively). If intra-picture compression is the first major component of MPEG-2, motion prediction between frames is the second. Inter-picture compression relies on forward and backwards correlations, and in particular on the calculation of motion vectors for blocks. In the process of encoding a video sequence, the MPEG-2 codec analyzes for each picture how blocks have moved, and only transmits lists of motion vectors describing the movement of blocks in relation to a reference picture. This fundamentally alters the framing of images. We have already seen that rather than the raw pixel being the elementary material of the image, the block becomes the elementary component. Here the picture itself is no longer the elementary component of the sequence, but an object to be analysed in terms of sets of motion vectors describing relative movements of blocks and then discarded. The 'picture' after encoding is nothing but a series of vectors describing what happens to blocks. Decoding the MPEG stream means turning these vectors back into arrangements of blocks moving between frames.

Motion prediction takes time. The actual combination 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. 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 usually 12 or 15 frames in a sequence such as I_BB_P_BB_P_BB_P_BB_P_BB_. One intra-coded frame is followed by a dozen or so block motion-compensation frames. 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.

From complicated to composite

Many of the complications of the MPEG-2 codecs arise because this software and hardware lies at the intersection of network bandwidth constraints (a commercially marketed service), conventions, embodied cognition, and cultural forms (moving images). Analysing such intersections requires ways of articulating diverse realities. If 'actuality' is in its essence 'composition' (Whitehead, 1958, 162 -163), software like codecs might offer places to begin understanding what this might mean practically. Codecs respond to the economic-technical need to reduce the bandwidth needed to transmit high-resolution digital pictures and sounds. As a convention, the MPEG-2 standard refers implicitly to a great number of physical entities ranging from screen dimensions through network and transmission infrastructures to semiconductor and data storage technologies. The generic method of encoding and decoding images for transmission relates very closely to the constraints and conditions of telecommunications and media networks. Yet the codec more or less performs the function of displaying light, colour and sound on screen within calibrated psycho-perceptual limits. The advent of realtime digital networked media afforded by codecs does not constitute a radical re-ordering of the content of video.

References

Bowker, G. C. and Star, S. L. (1999) *Sorting Things Out. Classification and Its Consequences*, MIT Press, Cambridge MA.

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.

Rabinow, P. (1999) *French DNA : trouble in purgatory*, University of Chicago Press, Chicago, Ill. ; London.

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

Thrift, N. (2004). "Movement-space: the changing domain of thinking resulting from the development of new kinds of spatial awareness." *Economy and Society* 33(4): 582-604.

Virilio, P. (2000) *The information bomb*, Verso, London.

Whitehead, A. N. (1958) *Modes of thought; six lectures delivered in Wellesley College, Massachusetts, and two lectures in the University of Chicago*, Capricorn Books, New York,.

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