Never mind the quality – feel the bandwidth

As MPEG video and audio compression becomes increasingly popular, there are signs that the quality is suffering. Here John Watkinson explains why compression artifacts occur and – more importantly – what can be done to reduce them.

John Watkinson, FAES, B.Sc., M.Sc.

In audio/visual program material, the advantages of compression are many – hence its popularity. Compressed material requires less bandwidth which is ideal for broadcasting, where the radio spectrum is under increasing pressure from other mobile services.

Compression also allows the cost of storage or recording to be reduced, although as recording economics improve year on year this may be a transient advantage.

Compression principles

Figure 1a) shows that in all real digital program material the bit rate is the product of the sampling rate and the word length. In practice the overall bit rate is made up of a varying mixture of unpredictable or novel material, known as entropy, and the remainder which could be deduced from first principles, known as redundancy.

An ideal compressor would separate the two perfectly so that only the entropy need be sent. An intelligent decoder would work out the redundancy for itself and reproduce the source signal without loss.

Entropy is a characteristic of the signal and varies. Figure **1b)** shows that if all of the entropy is not sent there is quality loss. The ideal is a variable rate channel which allows constant quality. If a fixed rate channel has to be used, the quality will vary.

In an MPEG-2 transport stream, several compressed signals can be statistically multiplexed together. It is unlikely that all will reach an entropy peak together, consequently a transport stream can be divided into a number of varying bit rate channels.

Provided that the overall bit rate remains constant, individual channels can demand more bandwidth when difficult material is encountered on the assumption that other channels are probably handling easier material at that time.

In the DVD – digital video disk, also known as digital versatile disk – a variable bit rate is supported in a single program stream simply by changing the rate of disk accesses.

Unfortunately the ideal coder of Fig. 1 is infinitely complex and has an infinite processing delay. Practical coders have to constrain both. When either of these constraints are applied, the bit rate has to go up to maintain quality, as Fig. **1c)** shows. **Figure 2** shows that for constant quality the bit rate will reduce as the latency increases.

Video compression

In MPEG-2, the temporal compression is obtained by sending motion compensated difference pictures, and further spatial compression is obtained by transform coding the differences.

Differential coding simply subtracts the previous picture from the current picture and sends the difference. When there is motion differences increase. This is handled by measuring the motion between pictures on a 16-by-16 pixel block, or macroblock, basis

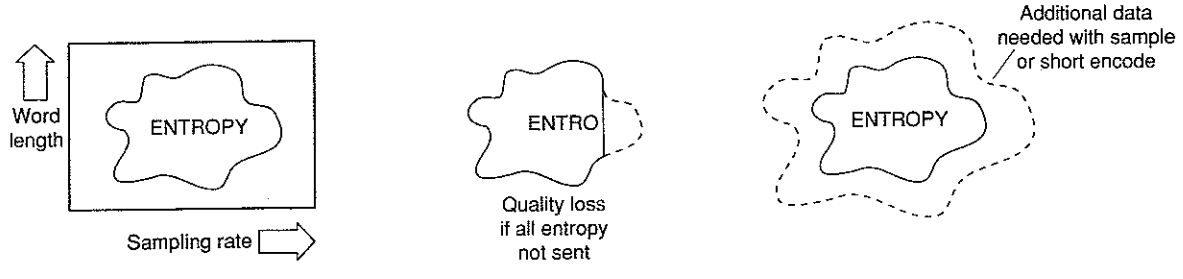

Fig. 1. Entropy is always less than bit rate, a). To avoid quality loss, all entropy must be sent, b). The simpler or faster the coder, the more data that must be sent, c).

and transmitting a vector for each block. The decoder and the encoder both shift the previous picture using the vectors and only the difference between the shifted previous picture and the current picture need be sent.

Pure differential coding fails if there is a transmission error because that error propagates indefinitely. It also makes it hard for the viewer to change channel! In practice periodic whole pictures have to be sent to prevent error propagation and to create decoder entry points. These are known as intra-coded, or I, pictures because they make no reference to any other picture and are only spatially compressed. In between these I pictures differential coding is used.

Moving objects cause problems in differential coding because they reveal background at their trailing edge which is previously unknown.

This is overcome by using information from future pictures. **Figure 3** shows that in bidirectional coding a picture can be decoded using information from pictures before or after. The decoder does not need a crystal ball to obtain the future pictures: instead, pictures are sent out of sequence.

Figure 4 shows that after an I picture, a future picture is differentially coded in the forward direction only. This future P (predicted) picture is sent immediately and stored in the decoder.

Pictures between the I picture and the P picture can now be sent these B (bidirectional) pictures can be created by forward or backward motion compensated differences on an individual macroblock basis.

The pictures and difference pictures are spatially compressed. The process begins by performing a dct, or discrete cosine transform, which expresses an 8-by-8 pixel block as a set of 64 coefficients. In typical video material, many of the coefficients will have zero or negligible values so that only the significant ones need to be transmitted.

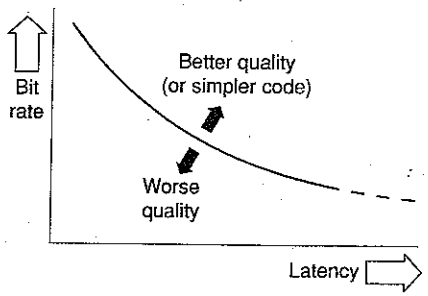

Fig. 2. Shorter latency needs a higher bit rate.

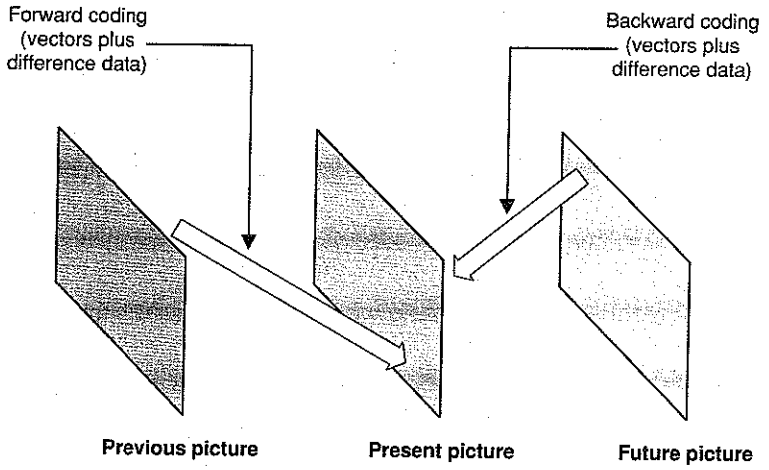

Fig. 3. Bidirectional coding uses information from both past and future picture frames.

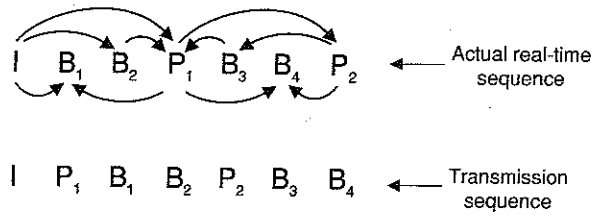

Fig. 4. Bidirectional coding requires pictures to be transmitted out of sequence.

Compression artifacts

Compressors are generally iterative and are driven by a bit-budget measurement. If the output bit rate is too high for the channel the dct coefficients will have to be expressed in fewer bits. In the case of large value coefficients, when low order bits are lost they become less accurate. In the case of small value coefficients, they may be truncated to zero.

This has a number of side effects. Coarse quantising of large value coefficients means that after the inverse dct at the decoder the eight-by-eight pixel block may have considerable errors in the sample values.

While these are not necessarily visible in themselves, the errors in adjacent blocks will mean that there is a discontinuity in the block boundaries so that the blocks become visible as shown in **Fig. 5a)**. If high frequency coefficients have been truncated to zero the block will lack detail and resemble a tile as in **5b)**.

This effect occurs in both the luminance and colour difference paths. In luminance the effect is called contouring whereas in colour the effect is called posterising, where gradual colour changes have been replaced by a limited colour set, as might be available in a box of poster paints. In MPEG the colour posterising can be quite obvious because the chroma blocks are the size of a macroblock and have four times the screen area.

Effects of truncating hf coefficients

Where high-frequency coefficients have been truncated to zero, the effect is to introduce ringing on edges. This is because an edge contains high frequencies and removing them is the equivalent of a sub-optimal low pass filter, hence the ringing. This is particularly noticeable on graphics and captions, less so on natural subjects.

When the prediction of the temporal coding fails, the data in the difference pictures will necessarily increase and this will force the compressor to quantise more heavily, raising the artifact level. This is particularly noticeable on B pictures since they are generally allocated only 10% of the data rate.

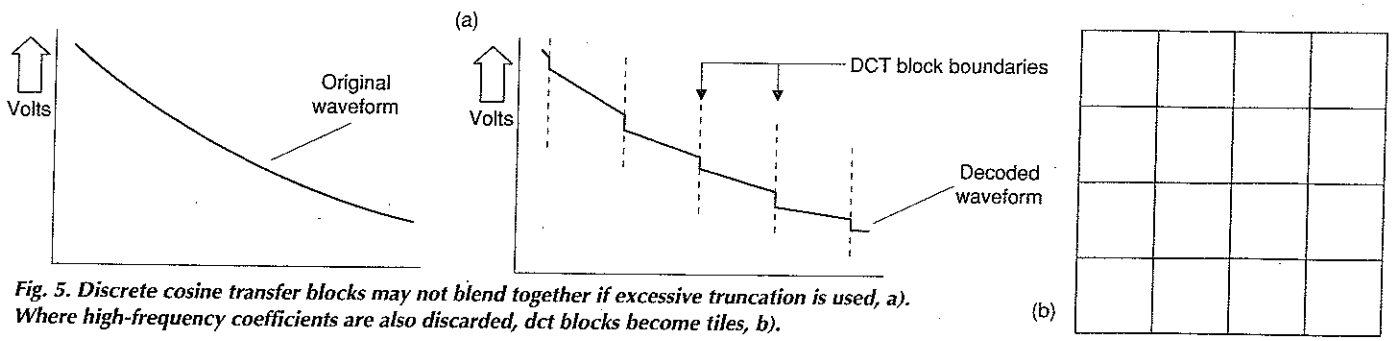

Fig. 5. Discrete cosine transfer blocks may not blend together if excessive truncation is used, a). Where high-frequency coefficients are also discarded, dct blocks become tiles, b).

Temporally difficult material, such as when frequent cuts are made, may overload an encoder. Cuts remove temporal redundancy and defeat bidirectional coding. Following a cut, several pictures may contain serious blocking artifacts.

The real MPEG killer material is video from a press conference where flashguns are firing. Each flash drives up every single pixel value for one picture, and then in the next picture the values come back to normal. This causes temporal chaos and most MPEG coders substitute a picture of the designer's bathroom wall under these conditions.

Pre-processing controls artifacts

The level of artifacts can be controlled by pre-processing. There are three levels at which a pre-processor can operate,

- By removing noise from the source material.
- By removing entropy from the source material.
- By aligning the I pictures in the coder output with the temporal entropy of the source.

Noise in a source pixel block creates more coefficients than a noise free source would. Thus all coefficients have to be truncated more aggressively to carry them, raising artifact level. Noise also increases data in difference pictures. Hence noise reduction will lower artifact levels by reducing spurious coefficients and reducing picture difference data.

If, after other steps, the artifact level is still too high, then the only approach is to restrict the entropy entering the coder. This is done by down-sampling the source images either spatially, so they contain less pixels, or temporally, so there are fewer pictures per unit time, or both.

In source material from telecine, the use of 2:2 and 3:2 pull-down creates what could be called false entropy, because in 2:2 frames are interlaced to make fields, giving a false doubling of picture rate.

The ratio 3:2 gets its name because 24Hz film frames are alternately converted to two and three fields to give a 60Hz output. One in five fields is redundant. Prior to MPEG coding telecine material has to have redundant fields discarded and remaining field pairs are de-interlaced to obtain the original frames.

The largest usage of data in MPEG is the I picture. This is because it does not use any previous information from the source. Consequently it makes no difference if the source I picture is radically different from the ones which went before. In contrast both P and B pictures will require significantly more difference data if there is a cut.

It follows that a significant reduction in artifacts can be obtained if I pictures are temporally aligned with source cuts. The only drawback of this approach is that to do it in real time a great deal of memory is needed to pipeline a stack of frames so that picture type decisions can be taken.

The alternative is to use a time coded source recording and use a two-pass encoding process. On the first pass the cuts are detected and used to design a picture type structure which is stored, and on the second pass the structure is implemented. ■

John is an independent consultant in digital audio, video and data technology and is the author of fifteen books on the subject, including Compression in Video and Audio. He is a Chartered Information Systems Practitioner, a Fellow of the Audio Engineering Society.