John R. Smith
IBM

The VC-1 Video Coding Standard

Hari Kalva
Florida Atlantic University

Jae-Beom Lee
Sarnoff Research Labs

Present-day communication systems rely more heavily on video compression than ever before. As broadband services increase in availability, video services such as user-generated content sites increase in popularity.

Despite these recent advances, video services over the Internet still must choose between hosting lower-quality videos that need less bandwidth or hosting higher-quality video downloads that aren't available in real time. Two broad categories classify the quality of video services and user experience: video over the Internet and broadcast video services.

Broadcasters deliver video over the Internet through Internet Protocol (IP) networks, which isn't the same as video over IP (IPTV) that telecommunications companies deploy as an alternative to cable TV services. IPTV guarantees the quality of service while Internet video services do not.

The nature of the service influences the video compression algorithms used. Online services require low complexity and low bit rate video codecs, while broadcast video services value quality as the most important criterion. Online users widely employ VP6, WMV, Real video, MPEG-4, H.264, and H.263 as the most popular video codecs. However, the MPEG-2 video compression standard dominates broadcast services today with applications that include digital TV, HDTV, and DVD.

The introduction of new video services opened opportunities for other codecs to break into broadcast services—for example, the high-definition DVD standard specifies H.264 and VC-1 in addition to MPEG-2, and new video services such as IPTV are considering codecs such as H.264 and VC-1. We actually discuss VC-1 and H.264 and their application in the next generation of video services in much greater detail elsewhere.[1]

Microsoft based its VC-1 video compression standard on its WMV-9 video standard.[2] WMV-9 supports progressive video and is mainly used for online video services. VC-1 extends WMV-9 and adds features necessary for broadcast services such as interlace support.

The Society of Motion Picture and Television Engineers (SMPTE)[3] has since standardized VC-1, and the DVD forum has adopted it for the high definition DVD standard. We expect that VC-1 will be deployed as a key engine in satellite TV, IP set-tops, and high-definition DVD recorders and players.

Editor's Note

Achieving high-quality online video services requires improved video coding capabilities. The video compression standard VC-1 was developed mainly for online video services and has been adopted by the DVD forum for high-definition DVDs. We can expect to see VC-1 applied to satellite TV, IP set-top boxes, and high-definition DVD recorders.

—John R. Smith

Overview

VC-1 is a hybrid video codec similar to MPEG and other commercially used codecs. It compresses video using a hybrid of motion compensation and transform coding.

The block diagram of the codec shown in Figure 1 is functionally similar to the block diagrams of other hybrid codecs. The codecs encode video one picture at a time and support both progressive and interlaced video coding.

Interlaced video can be coded as a single frame or two fields. Progressive video, on the other hand, can only be coded as a single frame.

VC-1 supports five types of pictures: Intra (or I) pictures, Predictive (P) pictures, Bi-predictive (B) pictures, skipped pictures, and BI pictures. An I picture is coded without using any previously coded pictures and can be decoded independently. The P pictures use a prediction from previously coded pictures and cannot be decoded independently.

The B pictures use two previously coded pictures for prediction, and the BI pictures are B pictures with only I macro blocks (MB). The BI pictures work well when the scene changes and

1070-986X/07/$25.00 © 2007 IEEE

Published by the IEEE Computer Society

the use of a B picture will result in a large number of I macro blocks. For example, a B picture with I macro blocks still has to use a B-picture syntax that uses more bits, and hence is replaced with a BI picture that takes fewer bits for the syntax elements.

When a picture is similar to its reference, the codecs use the skipped pictures. In the decoding process its reference frame replaces the skipped picture. VC-1 does not have a fixed group of pictures (GOP) structure and the number of pictures in a GOP can vary.

Input video to a VC-1 encoder must be in the YUV 4:2:0 format and the output of a VC-1 decoder is also in the YUV 4:2:0 format (where Y stands for the luma, or brightness, component and U and V are the chrominance, or color, components). Unlike the MPEG video coding standards such as MPEG-2 and H.264, VC-1 does not support 4:2:2 and 4:4:4 chroma sampling.

The encoder codes VC-1 bitstreams hierarchically with sequence, picture, entry-point, slice, macro block, and block layers. It codes a sequence as a set of pictures and each picture as a set of slices.

A slice consists of integral numbers of macro block rows and is more constrained compared to the MPEG-2 or H.264 slice structure. The entry point layer is similar to the GOP structure in MPEG video standards. It provides random access into a bitstream. The slice and entry point layers are present only in the advanced profile. The simple and main profiles thus contain only the sequence, picture, macro block, and block layers. In general, the VC-1 design simplifies the syntax compared to H.264.

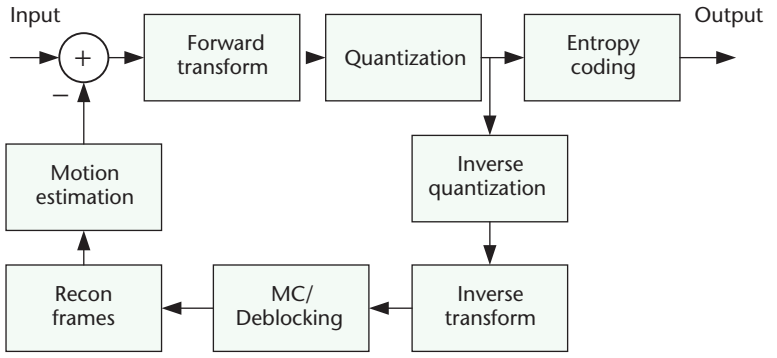

Figure 1. Functional block diagram of a VC-1 encoder shows close similarities to other hybrid codecs. (MC stands for motion compensation.)

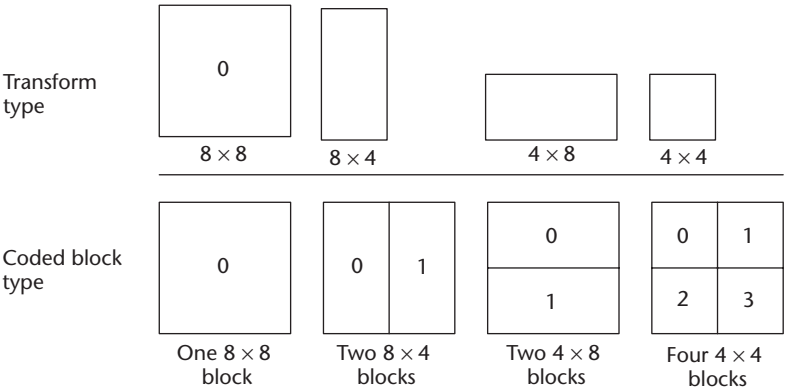

Figure 2. The variable-sized transforms in VC-1, which are more flexible than the two fixed sizes in H.264.

Transform coding

VC-1 uses an integer transform that's an approximation of the discrete cosine transform (DCT). The transform's design lets decoders implement the transform with operations on 16-bit registers. A key difference between VC-1 and the other codecs is that it uses variable-sized transforms. The motion-compensated blocks in Inter MBs are transformed using one of the four available transforms. The Intra MBs always use an 8×8 transform. In Inter MBs, each 8×8 block can be transformed using an 8×8, two 4×8, two 8×8, or four 4×4 transforms. In contrast, H.264 uses a fixed transform of size 4×4 or 8×8. Figure 2 shows the available transforms and their use in 8×8 blocks.

The variable-sized transform is optional and is signaled at the sequence level. When a user enables variable-sized transforms, the chosen transform type can be signaled at the picture level, MB level, or block level.

When signaled at the picture level, all macro blocks in the picture use the same transform size, and when signaled at the macro block level, all blocks in the macro block use the same transform size. I frames only have I MBs, and hence use a fixed 8×8 transform.

Motion in video sequences can cause motion compensation to be more effective in some areas of a macro block while residual in the other areas is still large. In such cases, a variable size transform allows a more effective de-correlation of the residual signal. Because an encoder has to evaluate the best possible transform size, the complexity of the transform coding stage increases.

The transform coefficients go through the quantization stage to reduce perceptual redundancies and bit rates. Simply put, a quantization parameter at the encoder divides transform coefficients, and the quotient is transmitted. The decoder multiplies this quotient by the same quantization parameter to reconstruct the coefficient.

Quantization is a necessarily lossy process and

The VC-1 standard simplifies the motion estimation stage by allowing a maximum of two reference frames and two block sizes.

reduces the bit rates significantly. The user can specify the quantization parameter in VC-1 at the picture level or optionally at the macro block level. I frames use a single QP especially in Simple and Main Profiles, specified at the picture level and the P and B frames let QP be modified at the MB level allowing for a finer control over the bit rates.

Motion compensation

Motion estimation is the most computationally intensive part of video encoding. The motion estimation stage of an encoder finds matching blocks in reference pictures that are similar to the block that's currently being coded.

To compensate for motion in a video, the matched block is subtracted from the current block. The closer the block match, the smaller the difference, and the smaller the amount of information transmitted to a decoder.

The likelihood of finding a good matching block increases as the block size decreases. Block matching also improves with the number of reference pictures. With an increased number of reference frames and block size options, the complexity of an encoder also increases. This tradeoff of quality for complexity has a significant impact on the coded video.

The H.264 standard allows the most flexibility by allowing up to 16 reference frames and seven different block sizes, making the encoder highly complex. The VC-1 standard simplifies the motion estimation stage by allowing a maximum of two reference frames and two block sizes.

VC-1 encodes the Inter MBs using either one-motion vector (MV) or four-MV mode. It uses a 16×16 block in one-MV mode and four 8×8 blocks in four-MV mode. In four-MV mode, each 8×8 block has a separate MV that is transmitted to the decoder. Motion vectors use either 1/2-pixel or 1/4-pixel resolution as signaled at the picture level. In P pictures, MV mode is signaled at the picture level to indicate to the decoder whether 1-MV, 4-MV, or mixed-MV mode is used in the current picture. When mixed-MV mode is used, an MB-level signal indicates the MV type for each MB. VC-1 also allows mixed MBs where a macro block coded in 4-MV mode can have up to three blocks coded in Intra mode. The B frames in VC-1 progressive coding use only 1-MV mode.

The advanced profile of VC-1 allows interlaced video coding and adds more flexibility to the motion compensation stage. The P field-pictures can't exceed two references. The picture layer specifies the use of two fields for references. If it uses two, the actual reference field is described in the MB level and Block level. The B field-picture always employs four references and needs no picture layer selection. The number of B frames between two reference frames in the advanced profile can vary. The picture structure of the entry point layer encodes the new distance value.

VC-1 also allows intensity compensation in addition to motion compensation. When enabled, intensity compensation is applied to pixels in the reference frame before motion compensation. The references are shifted and scaled to create the intensity-compensated pixels that improve the effectiveness of motion compensation.

Loop filter

VC-1 provides two techniques to reduce the blocky effect around transform boundaries: overlapped transform (OLT) smoothing and in-loop deblocking filtering (ILF). OLT is a unique technique based on an accurately defined pre/postprocessing pair. The idea is that forward and inverse operations are defined in such a way that original data are recovered perfectly when operations are serially applied (forward and then the inverse).

The forward transform exchanges information across boundary edges in the adjacent blocks. In a typical case of a block edge, one block has relatively good edge details, while the other block doesn't. The decoder requires an inverse operation to exchange the edge data back again to reduce the blocking effect. Through this approach, good-quality and bad-quality edge pairs diffuse each other thereby improving the visual quality.

The ILF is a more or less heuristic way to reduce the blocky effect. A blocky pattern is high frequency when abrupt value changes occur around block edges. Considering that original data quality might also contain high frequency, this process applies a relatively simple nonlinear

low pass around block edges on the I and P reference frames. Thus, the result of filtering affects only the quality of the next pictures that use the filtered frames as references.

Entropy coding

VC-1 uses simple variable-length coding (VLC) tables for coding coefficients and syntax elements. This improves efficiency by using multiple tables for coding the same syntax elements and signaling the tables used. This approach is less efficient than the context-adaptive VLC used in H.264, but it also has less computational complexity.

Profiles and levels

Profiles and levels of a video codec manage its complexity. It may be unnecessary to implement all possible configurations of a codec depending on its application. For example, we wouldn't expect a video decoder in a mobile phone to play a high-resolution, broadcast-quality video or to implement the features of a broadcast-quality decoder.

Profiles of a video codec define a subset of tools and algorithms used, and levels within a profile place constraints on the parameters that define a particular profile. VC-1 defines three profiles: a simple profile designed for low bit rate video applications, a main profile designed for high bit rate video streaming and higher-quality video delivery over the Internet, and an advanced profile designed for high-quality applications such as digital TV and HD DVD.

Complexity

Table 1 compares the key features of VC-1 and H.264. The H.264 video codec gives a slightly better quality than VC-1 for a given bit rate, but it also has higher complexity. The complexity of H.264 is due to the high degree of flexibility offered by the standard as H.264 is intended for a broad range of applications from low bit rate video to HDTV. The choices made in VC-1 represent quality versus complexity tradeoffs. For example, the use of VLC tables may not give the best performance compared with arithmetic coding but complexity is also low. Especially the hardware implementation of VC-1 is considered easier to develop leading to lower cost implementations.

Conclusion

VC-1 is a new video coding standard developed by Microsoft and standardized by the SMPTE. VC-1 is one of the three video compression algorithms standardized for high definition DVD. With high definition DVD players expecting to support MPEG-2, H.264, and VC-1, end users do not have to be concerned about the coding formats. The VC-1 standard offers a competitive quality-complexity tradeoff compared to H.264, especially for high-definition services. With a diverse digital video market, we can expect to see VC-1 co-existing with H.264 in the next generation of broadband and broadcast video services. **MM**

Table 1. Comparison of features for VC-1 and H.264.

Features	VC-1	H.264
Picture type	I, P, B, BI, Skip	I, P, B, SI, SP
Transform size	Adaptive	Fixed (baseline)
Transform	Integer, similar to discrete cosine transfer (DCT), but with four different transform sizes	Integer, similar to DCT, with 4×4 or 8×8 transforms
Intraprediction	Simple predictor	Directional predictors
Motion compensation	16×16, 8×8	Seven variable block sizes
Reference frames	Maximum of two	Maximum of 16 (in each direction)
Entropy coding	Multiple variable-length coding (VLC) tables	Context-adaptive VLC, arithmetic coding
Deblocking	In-loop filter/overlapped transform smoothing	In-loop filter

References

1. J.B. Lee and H. Kalva, *The VC-1 and H.264 Video Compression Standards for Broadband Video Services*, Springer, 2008.
2. S. Srinivassan et al., "WMV-9: Overview and Applications," *Signal Processing Image Communication*, Oct. 2004, pp. 851-875.
3. Soc. Motion Picture and Television Engs. (SMPTE) Technology Comm. C24 on Video Compression Technology, "Proposed SMPTE Standard for Television: VC-1 Compressed Video Bitstream Format and Decoding Process," *SMPTE 421M*, SMPTE, Aug. 2005.

Readers may contact Hari Kalva at hari@cse.fau.edu and Jae-Beom Lee at jlee@sarnoff.com.

Contact Standards editor John R. Smith at jsmith@us.ibm.com.