Recent Advances in Video Compression Standards

by *Guy Côté and Lowell Winger,*
VideoLocus, Waterloo, ON

Abstract

Video compression is a critical component of many multimedia applications available today. For applications such as DVD, digital television broadcasting, Satellite television, Internet video streaming, video conferencing, video security, and digital camcorders, limited transmission bandwidth or storage capacity stresses the demand for higher video compression ratios. To address these different scenarios, many video compression standards have been ratified over the past decade. This article first discusses the general structure and components of a standards-based video coding system. An overview of the emerging video coding standard H26L, currently being developed jointly by the ITU and ISO standard bodies, is then presented, highlighting key differences with its predecessor standards, such as MPEG-2, MPEG-4, and H.263.

Sommaire

La compression vidéo fait partie intégrale de plusieurs applications multimédia disponibles aujourd'hui. Pour certaines applications, par exemple les lecteurs DVD, la transmission de télévision numérique, la télévision par satellite, la transmission de vidéo par l'Internet, la vidéo conférence, la sécurité vidéo, et les caméras numériques, une bande de transmission limitée ou de la mémoire limité contribuent une demande pour des rapports de compression vidéo plus élevés. Pour adresser ces différents scénarios, plusieurs standards de codage vidéo ont été ratifiés durant la dernière décennie. Cet article discute en premier lieu la structure générale et les composantes d'un système de codage vidéo standard. Une description du standard émergent H.26L, qui est présentement en développement par les groupes de standardisation ITU et ISO, est présentée, soulignant les différences clés avec les standards précédant, tel que MPEG-2, MPEG-4 et H.263.20.

1.0 Introduction

Video compression is necessary for transmission of digital video over today's band-limited networks, or for storage constrained applications. For example, the transmission of digital video at 24 bit per pixel raw sampled at 720 by 480 spatial resolution and 30 frames per second (fps) temporal resolution[1] would require a bit rate of above 248 Mbps!

Compression of digital video without significant quality degradation is usually possible because video sequences contain a high degree of: 1) spatial redundancy, due to the correlation between neighbouring pixels, 2) spectral redundancy, due to correlation among the colour components, 3) temporal redundancy, due to correlation between video frames, and 4) psycho-visual redundancy, due to properties of the human visual system (HVS).

Removal of temporal redundancies in video signals accounts for a significant percentage of the achieved compression. Therefore, advanced techniques for the coding of the residual signal usually provide little additional compression as compared to traditional techniques, and additional complexity often does not justify this improvement. Rather than improving residual coding itself, most effective techniques attempt to reduce the residual to be coded, by improving the prediction of motion in a video sequence.

In recent years, interest in multimedia has generated a lot of research in the area of video coding in academia and industry alike and several successful standards have emerged, e.g. ITU-T H.261 [1], H.263 [2], ISO/IEC MPEG-1 [3], MPEG-2 [4] and MPEG-4 [5]. These standards address a wide range of applications having different requirements in terms of bit rate, picture quality, complexity, error resilience and delay, as well as improved compression ratios.

Here we first describe the block-based hybrid motion compensated and transform video coding method used by all video standards today. We briefly describe each component of such a system. The emerging H.26L [6] video coding recommendation is then described. We present the H.26L coding tools that differ significantly from previous video coding standards, and describe the performance of this standard in comparison to previous standards such as MPEG-2 and MPEG-4.

2.0 Block-based motion compensated and transform video coding

In the hybrid motion compensated and transform video coder, motion compensated prediction first reduces temporal redundancies. Transform coding is then applied to the corresponding difference frame to reduce spatial redundancies. For highly correlated sources, such as natural images, the compaction ability of the Discrete Cosine Transform (DCT) is very close to that of the optimal transform, the Karhunen-Loeve Transform (KLT). Moreover, the DCT, unlike the KLT, is data independent. This has made the DCT the most popular transform for image coding, as evidenced by its use in the JPEG still image international standard. Moreover, although motion compensated prediction difference frames are poorly correlated, the DCT is still the most popular transform for coding such frames. In fact, the DCT is used in all current video-coding standards.

In addition to removing temporal and spatial redundancies, psycho-visual redundancies are typically reduced as well. The most significant measure is a reduced resolution of colour detail in comparison to luminance detail to better match the characteristics of human perception. Video frames consist of three rectangular matrices of pixel data representing the luminance signal (luma Y) and two chrominance signals (chroma Cb and Cr) that correspond to a decomposed representation of the three primary colours associated with each picture element. Eight bits and 4:2:0 sub-sampling is the most common format used in video compression standards: the two chroma components are reduced to one-half the vertical and horizontal resolution of the luma component.

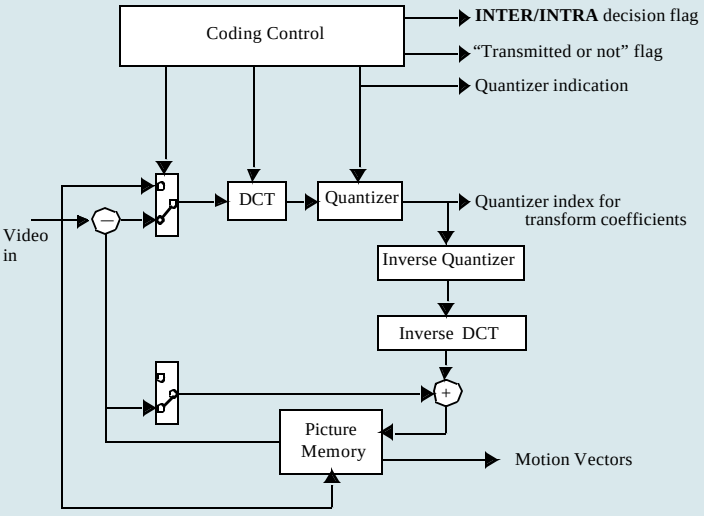

Figure 1: Block diagram of a block-based motion compensated and transform video encoder.

[1]: *720 x 480 pixels at 30 frames/second is the typical format used for broadcast television.*

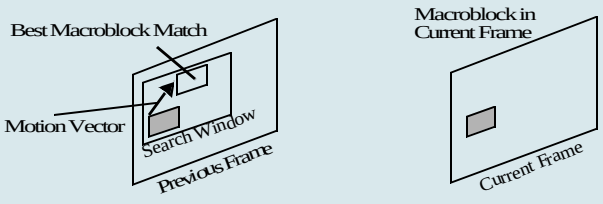

Figure 2: Example of motion estimation search window for a block matching algorithm.

A block diagram of a typical motion compensated prediction and transform video encoder is presented in Figure 1. In the next sections, we describe the building blocks of this video encoder.

2.1 Motion estimation and compensation

Each video frame is divided into macroblocks of equal size. Motion compensated prediction assumes that a block of pixels within the current picture can be modeled as a translation of a block from a previous picture, as shown in Figure 2.

Each block is normally predicted from the previous frame. This implies an assumption that each pixel within the block undergoes the same amount of translational motion. Two-dimensional displacement vectors or motion vectors represent this motion information. Due to the block-based picture representation, many motion estimation algorithms use block-matching techniques that obtain the motion vector by minimizing a cost function measuring the mismatch between a candidate block and the current block.

Although several cost measures have been introduced, the most widely used in motion estimation algorithms is the sum-of-absolute-differences (SAD), which computes the sum of pixel differences between the candidate reference block and the original block. To find the block producing the minimum mismatch error, we need to calculate the SAD at several locations within a search window. The simplest, but the most compute-intensive search method, known as the full search or exhaustive search method, evaluates the SAD at every possible pixel location in the search area. To lower the computational complexity, several algorithms that restrict the search to a few points have been proposed [7].

One motion vector per block is usually allowed for motion compensation. Sub-pixel motion estimation algorithms can provide a substantial improvement in reproduction quality. Most recent video coding standards allow both horizontal and vertical components of the motion vectors to be of half pixel accuracy. The range of representable motion vector values often limits the search window used in motion estimation. A positive value of the horizontal or vertical component of the motion vector represents a block spatially to the right or below the block being predicted, respectively.

Macroblocks can be predicted from previous frames only (P-macroblock), or from previous and/or future frames (B-macroblocks). The compression performance of B-macroblocks is superior to that of P-macroblocks, given the additional coding options. However, additional decoding delay is incurred, since the future P-frames must be decoded before temporally preceding B-frames can be decoded. A typical Group of Pictures is shown in Figure 3.

In current standards, motion compensation is usually performed on block sizes of 16x16 or 8x8 for P and B-macroblocks, and followed in the encoder by transformation as detailed in the next section.

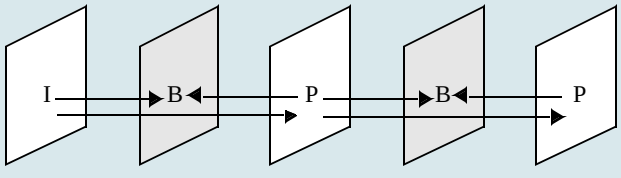

Figure 3: Typical arrangement of a Group of Pictures.

2.2 Transform

The purpose of the 8x8 DCT employed in all current video-coding standards is to de-correlate the 8x8 blocks of original pixels or motion compensated difference pixels and compact their energy into as few coefficients as possible. Besides its relatively high de-correlation and energy compaction capabilities, the DCT is efficient and amenable to software and hardware implementations. The most common algorithm for implementing the 8x8 DCT is 8-point DCT transformation of the rows followed by 8-point DCT transformation of the columns.

Although exact reconstruction of the original data can be theoretically achieved (inversion of the DCT transformation), it is often not possible using finite-precision arithmetic. While forward DCT errors can be tolerated, inverse DCT errors must meet a minimum level of precision in order to avoid IDCT mismatch between the reconstructed frames at the encoder and decoder.

The 8x8-DCT transform results in one DC coefficient and 63 AC coefficients. The DC coefficient is the mean or average of the transformed samples, representing the coarsest detail of the image block (the lowest spatial frequency). The AC transform coefficients represent finer image details (higher spatial frequencies).

In the encoder, transformation is followed by coefficient quantization, the stage at which loss of video detail is traded-off against the video compression ratio.

2.3 Quantization

The human viewer is more sensitive to reconstruction errors related to low spatial frequencies than those related to high frequencies. Slow linear changes in intensity or colour (low frequency information) are important to the eye. Sharp, high frequency changes can often not be seen and may be discarded. For every element position in the DCT output matrix, a corresponding quantization value is computed by dividing each DCT coefficient by a quantization parameter. The quantization parameter is the primary means by which the amount of compression and corresponding reduction in fidelity of the compressed video are controlled. The net effect is usually a reduced variance between quantized coefficients as compared to the variance between the original DCT coefficients, as well as a reduction of the number of non-zero coefficients, which improves the efficiency of the entropy coding, described in the next section.

2.4 Entropy coding

Entropy coding reduces the average number of bits used to represent the compressed video through the use of means such as variable length codes (VLCs). VLCs are often generated with Huffman codes such that shorter codewords are used to represent more frequently occurring symbols (such as small coefficient values). Arithmetic coding can also be used as means of entropy coding. Other information, such as prediction types and quantizer indication, is also typically entropy coded by means of VLCs or arithmetic codes.

Prior to entropy coding, motion vectors are usually predicted by component values set to the median value of three neighbouring already transmitted motion vectors: the motion vectors of the blocks to the left, above and above right of the current block. The difference motion vectors are then entropy coded.

Prior to entropy coding, the quantized DCT coefficients are arranged into a one-dimensional array by scanning them in zig-zag order. This rearrangement places the DC coefficient first in the array and the remaining AC coefficients are ordered from low to high frequency. The rearranged array is coded using a 3-dimensional run-length VLC table, representing the triple (LAST, RUN, LEVEL). The symbol RUN is defined as the distance between two non-zero coefficients in the array. The symbol LEVEL is the non-zero value immediately following a sequence of zeros. The symbol LAST is equivalent to the End-of Block flag also employed in 2-dimensional run-length coding, where “LAST = 1” means that the current code corresponds to the last coefficient in the coded block. This coding method produces a compact representation of the 8x8 DCT coefficients, as a large number of the coefficients are normally quantized to zero. Ideally, the re-ordering results in the grouping of long runs of consecutive zero values.

2.5 Coding control

The two switches in Figure 1 represent the intra/inter mode selection.

Such a selection is usually made at the macroblock level. The performance of the motion estimation process, usually measured in terms of the associated distortion values, can be used to select the coding mode. The coding mode where temporal prediction is used is called the inter mode. This mode is selected if the motion compensation process is effective, and only if the prediction error macroblock - the difference between the original macroblock and the motion compensated predicted macroblock - need be encoded. If temporal prediction is not used, the corresponding coding mode is called the intra mode. If a macroblock does not change significantly with respect to the reference picture, an encoder can also choose not to encode it, and the decoder will simply repeat the macroblock located at the subject macroblock's spatial location in the reference picture. This coding mode is referred to as skip. More sophisticated coding mode selection algorithms based on rate-distortion (RD) optimization methods can also be used, as discussed in the next section.

2.6 Rate-Distortion Optimized Video Coding

A key component in high-compression lossy video coding is the operational control of the encoder, through the motion estimation process, quantization step size selection, and the video coding mode selection. The process of selection between different possible representations with varying rate-distortion efficiencies can be optimized using Lagrangian minimization techniques based on rate-distortion theory [8], which are briefly described in this section. At the source coding level, rate-distortion theory sets limits on the achievable output distortion for a given coder output rate, or conversely, sets limits on achievable output rate for a given output distortion.

In video coding, the coding modes of operation are generally associated with signal-dependent rate-distortion characteristics, and rate-distortion tradeoffs are inherent in the coding parameters selection process. The optimization task is to choose, for each image block, the most efficient coded representation in the rate-distortion sense. This task is complicated by the fact that the various coding options show varying efficiency at different bit rates and with different scene content. For example, inter coding is efficient in representing key changing content in image sequences. However, intra coding may be more efficient in a situation where the block-based translational motion model cannot accurately represent the image sequence changes. For low activity regions of the video sequence, simply using the skip mode may be preferred. By allowing multiple modes of operation, we expect improved rate-distortion performance if the mode selection method is applied judiciously to different spatio-temporal regions of a video sequence.

The goal of the video compression system is to achieve the best fidelity (or the lowest distortion D) given the capacity of the transmission channel, subject to the coding rate constraint R(D). This optimization process can be solved using the Lagrangian multiplier method where the distortion term is weighted against a rate term. The Lagrangian formulation of the minimization problem is such that we minimize:

$J = D + \lambda R$, for a particular Lagrangian parameter λ. Each solution for a given value of the Lagrangian parameter λ should correspond to a locally optimal solution for a given rate constraint. A given value of λ represents a specific point on the operational rate-distortion curve. It is possible to obtain an approximate relation between the quantizer step size Q, which controls the output bit rate, and the optimal value of λ. This is particularly useful when a rate control method is used to achieve a particular video encoder bit rate.

3.0 Commonalities in the Emerging H.26L Recommendation

The elements common to all video coding standards that are discussed in the preceding sections are also present in the emerging H.26L recommendation, which is anticipated to become the newest international video-coding standard in early 2003. In summary, the following elements are present in the current H.26L recommendation: macroblocks are 16 lines by 16 pixels; luminance is represented with higher resolution than chrominance with 4:2:0 sub-sampling; motion compensation and block transforms are followed by scalar quantization and entropy coding; motion vectors are predicted from the median of the motion vectors of neighbouring blocks; bi-directional B-pictures are supported that may be motion compensated from both temporally previous and subsequent pictures; and a direct mode exists for B-pictures in which both forward and backward motion vectors are derived from the motion vector of a co-located macroblock in a reference picture. In the following sections, coding blocks of the emerging H.26L recommendation are compared and contrasted with other recent standards.

3.1 Intra prediction

H26L provides means to spatially predict intra-coded macroblocks. With these advanced prediction modes, the performance of intra-frame compression in H26L is similar to that of the recent still image compression standard, JPEG-2000. H263 and MPEG-4 also provide intra prediction. The differences between H26L and H.263 (and MPEG-4) are that the prediction is in the pixel domain, as opposed to the frequency domain, and sub-block level prediction modes are available, as opposed to only macroblock modes.

Intra coded macroblocks (in intra- or inter-frames) may use either 16x16 or 4x4 spatial prediction modes for luma. Three sub-modes are available with 16x16 prediction. A 16x16 macroblock can be predicted from the previously adjacent decoded pixels that are available due to the raster order (from the top-left with left-to-right swaths) decoding of macroblocks: vertical prediction from pixels above, horizontal prediction from pixels to the left, and plane prediction by spatial interpolation between the two sets of pixels.

Nine sub-modes are available with 4x4 prediction. A 4x4 sub-block can be predicted from the previously adjacent decoded pixels that are available due to the raster order decoding of each 8x8 block within a macroblock, and the nested raster order decoding of each 4x4 sub-block with each 8x8 block. Due to this decoding order, not all of the 4x4 prediction modes will always have decoded pixel data available in their desired prediction direction. In this case, the closest available decoded pixel data is used. The intra prediction modes are the following: DC prediction from the mean of adjacent pixels above and to the left, vertical (down) prediction from pixels above, horizontal (left) prediction from pixels to the right, diagonal (down-left) from pixels above and pixels to the right, diagonal (down-left) from pixels above and to the left, and four off-diagonal modes (+/- 22.5 degree predictions: left-of-vertical, right-of-vertical, up-from-horizontal, and down-from-horizontal).

3.2 Motion estimation and compensation

The H.26L recommendation supports the use of multiple different reference pictures from which prediction of inter macroblocks and blocks can be made. Multiple reference pictures may help prediction of transitionally covered background and periodic non-translational motion.

As in MPEG-4, ¼-pel motion compensation is used for temporal prediction. Six-tap interpolation filtering for the ½-pel positions is followed by bi-linear interpolation to derive the ¼-pel positions. A new feature is the existence of a funny position that is filtered more heavily to support instances in which only the coarse details, and not the high spatial frequencies, of current picture are accurately predicted by the reference picture. In addition, eight-tap interpolation filtering for 1/8-pel positions is optionally available.

As with. H.263 and MPEG-4, the model for motion compensation is variable-size block translation with motion vectors that may extend outside the picture boundaries by extending boundary pixel values to outside the frame. However, a larger variety of block sizes are now available for motion compensation. Each 16x16 macroblock may be divided horizontally and/or vertically for the purpose of motion compensation. If a macroblock is partitioned both horizontally and vertically, resulting in four 8x8 blocks, then each of those 8x8 blocks may also be partitioned horizontally and/or vertically. In this way, up to 16 motion vectors may be transmitted for a macroblock. The common partitioning of a 16x16 macroblock and of an 8x8 block is shown in Figure 4.

Macroblocks in inter-frames (P-frames or B-frames) may be coded as skipped, direct mode (B-frames only), intra 4x4 spatial prediction, or motion compensated with up to 16 motion vectors and with the possibility of optionally coding each 8x8 sub-block with intra 4x4 prediction.

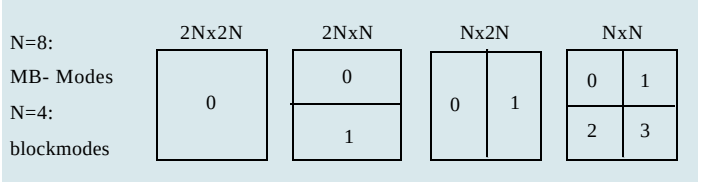

Figure 4: Available macroblock partitionning in H.26L.

Each 8x8 sub-block may predict from a different reference frame.

Post-filtering for the removal of blocking and ringing artefacts is known to be a critical element for obtaining high perceptual quality with current standards. With the H.26L recommendation, the ringing artefact is alleviated through the use of a shorter 4-point transform. Adaptive deblocking filtering is moved into the coding loop such that temporal prediction is based on the superior filtered reconstructed (decoded) images.

Conventional television is broadcast in interlaced format where a picture frame is divided in two picture fields that are displayed at a different time interval. The temporally first picture field (called top or odd field) is displayed on every odd line of a picture frame, and the second picture field (called bottom or even field) is displayed on every even line of a picture frame. The two fields form the picture frame. Interlaced material benefits from separate motion compensation and transformation in the different video fields when large motion is present. In H.26L, adaptive switching between separate and combined fields is supported at the picture level, rather than at the macroblock level, as previously supported by MPEG-2.

3.3 Transform & Quantization

A 4x4 integer "pseudo-DCT" transform replaces the previously common 8x8 DCT transform for de-correlating the pixel prediction residuals. The shorter transform length becomes more competitive with the longer transform due to the improved prediction modes for pixel luma. The benefits of the new transform are the complete elimination of inverse transform mismatch, which would lead to encoder/decoder mismatch in previous standards, improved perceptual quality, and lower complexity.

The 4x4 transform is expanded to a more traditional 8x8 transform for chroma and 16x16 luma predicted blocks through the use of a second 2x2 transform acting on the DC coefficient of 4x4 transformed blocks in a 4:2:0 macroblock. Similar to H.263 Annex T, a smaller step size is used for the quantization of chroma samples to improve chroma fidelity.

3.4 Entropy Coding

In contrast to other recent video coding standards, one universal table of variable length codes is available for VLC entropy encoding. Simplification is achieved by mapping each symbol to the VLC codeword that is appropriate given its frequency in the bitstream, rather than the more common approach of constructing separate VLC tables for each of the symbols sets (motion vector prediction residuals, run-levels, (macroblock modes, etc.).

As a higher complexity, higher performance alternative to VLC entropy coding, context-based adaptive arithmetic coding (CABAC) may be used. An arithmetic code is more efficient than a VLC for symbol probabilities that are much greater than 50%, since it permits a symbol to be represented with less than one bit. Adaptive codes reduce the inefficiency of non-stationary symbol statistics caused by mismatch between static codeword lengths and symbol probabilities that change due to bitrate, type of motion present in the source, and other factors. Context modelling provides estimates of the conditional probabilities of the symbols. The sophistication of the contexts defined with CABAC substantially improves upon the syntax-based arithmetic coding (SAC) optionally available with H.263.

3.5 Coding Control

Improvements in video compression with recent standards have often been fairly predictably achieved through the use of a larger number of choices. A much larger number of possible coding modes are available in the H.26L standard than in previous standards. As the number of coding choices increases, searching and rate-distortion optimization as tools for decision-making in the encoding process, as discussed in Section 2.5, become increasingly important.

4.0 Relative Performance and Conclusion

This survey of the emerging H.26L recommendation has compared key differences that lead to increased compression performance in comparison to previous standards. Compression improvement of up to 50% over the best previous standards is the primary motivation for advancing the new H.26L recommendation. Although storage and bandwidth are continuously growing, increasing demand for higher resolutions and more simultaneous streams with existing and emerging communications channels and storage media will continue to fuel the demand for greater compression performance.

5.0 References

[1]. ITU-T Recommendation H.261: "Video Codec for Audiovisual Services at px64 kbit/s", Geneve 1990.

[2]. ITU-T Recommendation H.263, Version 2: "Video Coding for Low Bitrate Communication", Geneve 1998.

[3]. ISO/IEC 11172-2:1993 Information Technology - Coding of Moving Pictures and Associated Audio for digital storage media at up to 1.5 Mbits/s. Part 2.

[4]. ISO/IEC 13818-2:2000 Information Technology - Generic Coding of Moving Pictures and Associated Audio Information. Part 2: Video.

[5]. ISO/IEC 14496-2:2001 Information Technology - Coding of audio-visual objects. Part 2: Visual.

[6]. ISO/IEC JTC1/SC29/WG11, ITU-T VCEG: "Working Draft Number 2 of Joint Video Team Standard", latest document publicly available at: ftp://ftp.imtc-files.org/jvt-experts/draft_standard/ and ftp://standards.pictel.com/video-site/0201_Gen/JVT-B118r2.zip

[7]. Peter Kuhn, "Algorithms, Complexity Analysis and VLSI Architectures for MPEG-4 Motion Estimation", Kluwer Academic Publications, 1999.

[8]. T.Berger, "Rate Distortion Theory", NJ: Prentice Hall, Inc. 1971.

About the authors

Guy Côté holds a Ph.D. in Electrical and Computer Engineering from the University of British Columbia and a B.A.Sc. in Electrical Engineering from the Royal Military College of Canada.

Guy is a co-founder and VP R&D at VideoLocus. Prior to VideoLocus, Guy developed video coding algorithms for PixStream Inc., which was acquired by Cisco Systems in December 2000. His work at PixStream/Cisco included research on different aspects of video coding standards including MPEG-1, MPEG-2, MPEG-4, and H.263.

He is an active participant in the ITU-T Video Coding Experts Group (VCEG) and a voting member of the International Standardization Committee ISO/IEC/JTCI/SC29 (JBIG, JPEG, MPEG). Guy has published over 25 papers and standards contributions and has four patents pending in the area of video compression.

Lowell Winger holds a Ph.D. in Electrical and Computer Engineering from the University of Toronto, a M.A.Sc. in Systems Design Engineering and a B.A.Sc. in Systems Design Engineering from the University of Waterloo, and is a licensed Professional Engineer.

Lowell Winger is a co-founder and currently CTO at VideoLocus Inc. Before founding VideoLocus, Lowell provided technical direction and oversight for development of a second generation, multi-pass MPEG-2 encoding platform and an architecture for a flexible video processing platform at PixStream Inc.

He is also currently an Adjunct Professor for the Dept. of Systems Design Engineering at the University of Waterloo, an active participant in the IEEE (Institute of Electrical and Electronics Engineers), and a voting member of the International Standardization Committee ISO/IEC/JTCI/SC29 (JBIG, JPEG, MPEG).