COMPRESSION, WHY, WHAT AND COMPROMISES

Paul Hightower

Instrumentation Technology Systems
Northridge, CA 91324
phightower@ITSamerica.com

ABSTRACT

Each 1080 video frame requires 6.2 MB of storage; archiving a one minute clip requires 22GB. Playing a 1080p/60 video requires sustained rates of 400 MB/S. These storage and transport parameters pose major technical and cost hurdles. Even the latest technologies would only support one channel of such video.

Content creators needed a solution to these road blocks to enable them to deliver video to viewers and monetize efforts. Over the past 30 years a pyramid of techniques have been developed to provide ever increasing compression efficiency. These techniques make it possible to deliver movies on Blu-ray disks, over Wi-Fi and Ethernet.

However, there are tradeoffs. Compression introduces latency, image errors and resolution loss. The exact effect may be different from image to image. BER may result the total loss of strings of frames.

We will explore these effects and how they impact test quality and reduce the benefits that HD cameras/lenses bring telemetry.

INTRODUCTION

Over the past 15 years we have all become accustomed to having television, computers and other video streaming devices show us video in high definition. It has become so commonplace that our community nearly insists that it be brought to the telemetry and test community so that better imagery can be used to better observe and model systems behaviors.

However, what we see on television, our computers and telephones is a facsimile of what the cameras deliver to the content creators. At a normal viewing distance, we don't notice the effective loss of resolution, introduction of image errors, color shifts, latency, macro blocking and other side effects of compression techniques used to make it possible to deliver content to these devices. When watching from 10 to 20 feet away, even a 1080 TV screen is below the resolution of the eye, so most of these effects are not seen by the viewer. However in using imagery for detailed analysis, close view, single frame images and enlargement of images are essential tools. Not only do these errors become obvious, they can obstruct the view, alter the view and effect the conclusions we draw.

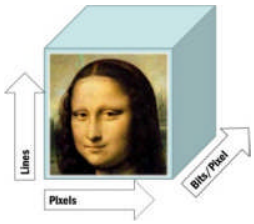
Any time transport of video from camera to display (or recorder) is more than 100 meters, it is nearly certain that compression is used to make the transport possible. Unless you have unlimited storage for your video, it is likely compression is involved. Consequently, all of the original pixel data delivered by the camera is transformed and no longer directly available.

This paper briefly explores the whys and explains the process and describes the compromises one must be prepared for when using compressed video as the imagery source of analysis.

THE WHY

Content creators have been shooting movies in 4K since 2006 nearly exclusively. Although 4K has only been offered in televisions for the past few years, 4K was driven into existence in order to reduce pixel size on a 50 foot screen in a movie theater; 1080 simply was not good enough in a theater showing. So 4K was not driven by TV, it was driven by the need for creators providers to find a means to deliver their content to paying customers. Television has adopted 4K principally as a means to improve the margins when retailers sell TVs to the public. Television broadcasters at present have no way to deliver true 4K content. In general you are watching 1080i video at 30 frames per second. An advanced 4K TV upscales this and estimates what a 4K image might look like derived from the 1080i it received. Even a 4K Blu-ray is a highly compressed source requiring significant math to recreate a facsimile of the 4K video captured by the cinematographer's camera.

Let's talk about storage needs first. The Society of Motion Picture and Television Engineers (SMPTE) manages and maintains the specifications of digital video frames. SMPTE 274 defines the 1080 frame as having 1125 lines (1080 viewable) and 2200 pixels (1920 viewable) per frame. However this is not the total story for resolution. The number of bits that is used to contain the image value (colors and intensity) is the third dimension of resolution in digital video. The International Telecommunications Union (ITU) manages and publishes image sampling specifications. ITU 709 is a standard that applies to high definition (HD) video. There are several sampling techniques, but the most commonly used is 4:2:2 subsampling. By subsampling it means that every two pixels are comprised of two intensity (luma) samples and one color subsample (red and blue)[1]. While originally conceived in analog video 70 years ago, it is used today as a means to reduce the number of bits it takes to represent a complete image. When an image then is specified as 1080 (progressive or interlaced), 4:2:2 10 bits, it is subsampled as described where the intensity element and the two color elements are each 10 bits. This suggests that up to 1024 shades of each color (intensity, red and blue) can be represented. That isn't quite true as SMPTE limits the range of numbers to protect certain bit patterns used for synchronization and ancillary data flags in the video pixel stream.



In general, 4:2:2 subsampling organizes pixels in pairs where two pixels have two intensity samples 10-bits each and one red (R) sample paired with the first pixel intensity and the blue (B) sample paired with the second intensity sample. Therefore, each pixel is 20-bits deep. In total then a single 1080 frame is 1125 lines x 2200 pixels x 20 bits; 6,187,500 bytes. The result? A single second of 1080 60 frames per second video requires 371 MB of storage. It also means that to transport 1080 video at 60 frames per second, the sustained transfer rate must exceed 371 MB/sec; 2.97 Gb/second without any packetizing overhead. These numbers are tough to hit with our existing transport, storage and image processing technologies.

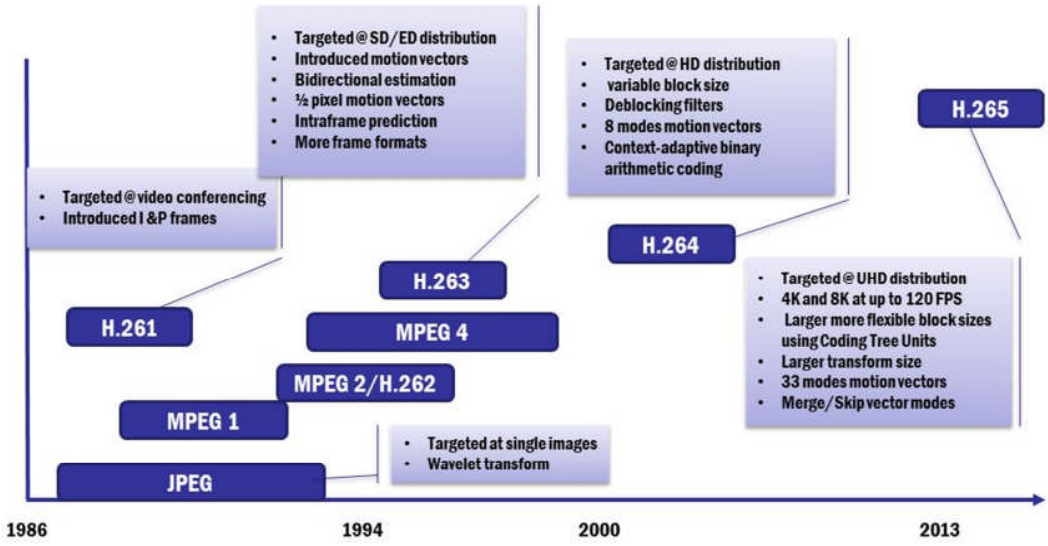
---

[1] A full color image is comprised of a red, blue and green component. The intensity of any pixel is the sum of the RGB content. Black and white images are effectively only the intensity. In the 1940's it was standardized that in order to deliver color television signals and still remain backward compatible with the population of B&W televisions owned by the general public, pixels were divided into intensity (B&W) and color. The third color (green) is derived by subtracting the R and B components of the total sample assuming that the total intensity component is comprised of the sum of all of the color components.

For example, for a movie producer to sell you a Blu-ray disk of a two hour moving in only 1080 resolution they have to find a way to save 2,673,000 MB of data in 50 GB of space. In order to deliver a movie over the best cable channel (500 Gb/sec) one has to fit a 3 Gb sustained data rate (packetized) in that 500 Gb and that channel must only be used for this one stream. The only solution currently available to deliver this content is to use video compression.

THE WHAT

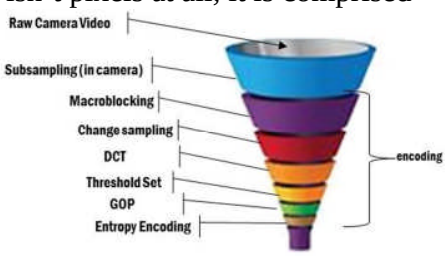
Image compression and its migration to video have been on-going for more than 30 years.



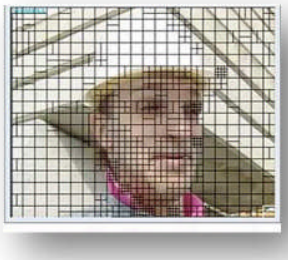
As you can see in the chart above, a pyramid of technologies have been stacked over time to result in the ubiquitous H.264 and the newly introduced H.265 (HEVC). At each stage, evolving computing capabilities, transport mechanisms and reduced storage costs have enabled the implementation of new approaches and techniques. In general all of the motivation is driven by the need to deliver content, not create it. The latest, H.265, has been implemented as the means to deliver 4K movies on Blu-ray disks to consumers. This was the economic drive to bring H.265 to reality.

In fact, the content developers (most notably the cinema community) generates 10s of petabytes (1,000 terabytes is a petabyte) per day in "dailiess"[i]. Cinema only uses uncompressed video from dailies to finished product. Generation errors and resolution loss prohibits the use of compression for editing and overlay special effects graphics. It is essential to use uncompressed to ensure image and color fidelity. It is essential to enable the use of CGI[ii]. Compression in cinema is only used for distribution to theaters, Blu-ray disks, streaming media and television.

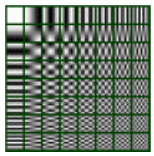
Why is this true? After the compression (lossy) process is complete, the original pixels are no longer present in the data. In fact the data file that is the video isn't pixels at all, it is comprised of a matrix of numbers referring to a matrix of patterns that when multiplied (coefficient x pattern) and added together approximate the shades contour of the pixel pattern from which it was derived.



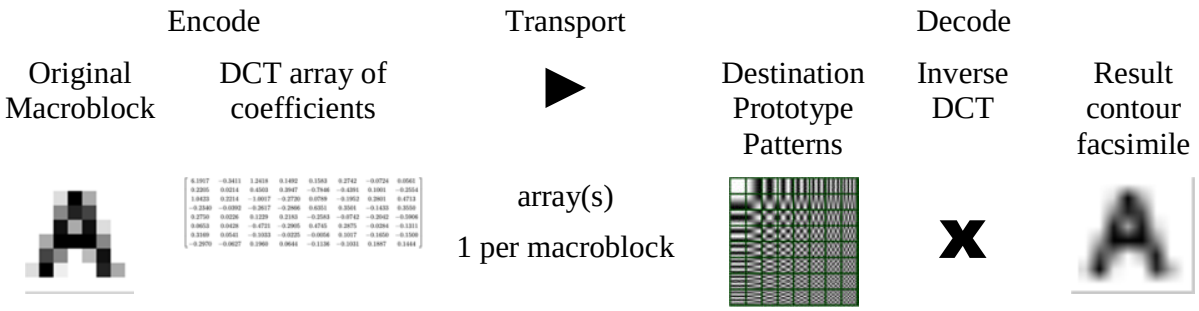
The total process can be depicted as the funnel to the right. We have already described what subsampling is and why it



can be used. After that, each image in the video is analyzed and broken into a mosaic of blocks. The construction worker image here represents one case of how an image may be analyzed and broken in to a mosaic of pixel arrays. Each array of pixels, referred to as a macroblock is processed through a discrete cosign transform (DCT) and correlator that generates a corresponding array of coefficients. Each element points to an array of standard image patterns. Each pattern is represented by a range of shades that is determined by the bit depth of the array itself. At the decode side, each prototype pattern is multiplied by its corresponding coefficient and added together. When complete a facsimile of the original contour of shades is reproduced. This process is similar in concept to a Discrete Fourier Transform (DFT) and how any waveform can be reproduced from a set of prototype sign waves multiplied by corresponding array of coefficients.



Diagrammatically the process end-to-end is depicted below.



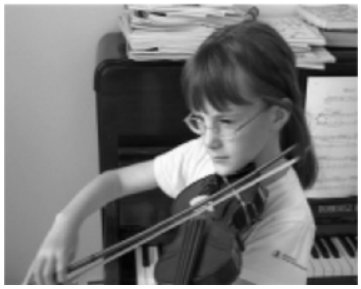
Example by Meisam - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=5789740

As you can see the original samples are replaced by an array then a facsimile of the shades represented by the samples is reproduced at the destination end. A human can see that the source and facsimile images are both “A”s, but they are quite different in detail. There is the rub. In the test and analysis business, the devil is in the details and details are lost in this process. Furthermore, the greater the compression needed to fit into any transport channel (or storage needed) the more alteration of the reconstructed image there is.

The last process in the compression sequence is entropy encoding. This is lossless Huffman encoding. Huffman encoding generates compressed digital streams by taking advantage of long runs of 1s or 0s in the data; run lengths. In preparing for application of entropy encoding, the encoder will look for low value coefficients and force them to zero to maximize the number of zeros in any one array and group of arrays to form long strings of zeros. As the transport channel narrows, the threshold at which a coefficient is set to zero is increased until the encoded output moves through transport without any loss. The higher this threshold, the more image information is permanently lost by the destination. It is not recoverable. While the pixel count in the reconstructed image is unchanged (still 1080x1920 for example), the contour of shades (luma and each color) may be quite different than the original.

This example only shows what effect there is on only one macroblock. Referring back to the construction worker image, a frame is composed of many macroblocks. How many is not known as each encoder evaluates each image differently. Each image analyzed by any one encoder will generate a different macroblock array. The Motion Picture Experts Group (MPEG) defines the

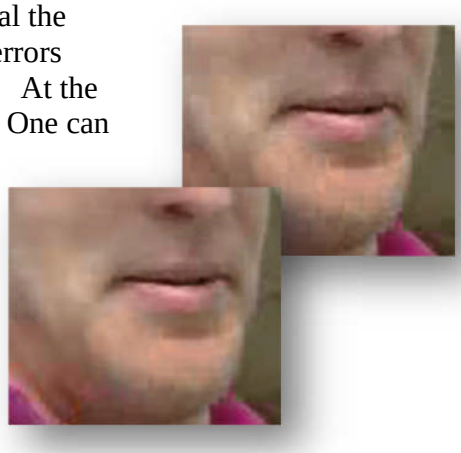
transport protocol for the results; that is the macroblock position, shape, number of pixels and the DCT array for each macroblock that comprises the original image.



In these black and white images, the original image (left) and the coded image (right) are quite different in detail. One can see they are both an image of a girl playing a violin in front of a piano, but many details have been lost or altered.

Look at the sheet music on the piano, her glasses frame, the macro-blocking of her arm.

Since the reconstructed shades contour is a facsimile of the original the reconstructed shades of the adjacent pixels mostly like will have errors making the macroblock edges evident in the reconstructed image. At the right is a portion of the reconstructed construction worker image. One can clearly see the macroblock array caused by mismatches of shades at the edges of the macroblocks. In order to compensate for that and hide the macro blocking, deblocking filters (decoder specific) are used to smudge the edges of adjacent macroblocks to attempt to merge the edges together. The lower image is the result after a deblocking filter is applied. It is a more pleasing image, but is certainly not the original array of pixels delivered by the camera. Details are lost and even modified.



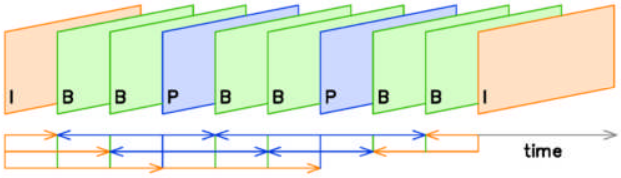
THE COMPROMISES

We have learned that the use of compression is a necessary evil to enable convenient transport and minimize storage of video clips. However in using compression we have learned that the original data gives way to a facsimile that can be quite different in detail than the original. Taking another look at the girl playing the violin, once can see that part of her eyeglasses frame on the left ear is not in the image at all. The sheet music is simply gray and gives no impression of music scales that are seen in the original. The detail the white keys on the piano is all but lost. There is a bright spot on her left cheek under the glasses did not exist in the original image. These changes are permanent after the DCT and entropy encoding is completed and received at the destination end. Is this a loss of resolution? Yes. Anything that results in the loss of detail can be thought of a reduced resolution. Is it aberration? Yes. Details are missing and added.

The prototype array may not be able to represent as many shades as delivered by the camera. That is to say 10-bit samples from the camera correlated to coefficients could be multiplied by an 8-bit deep prototype array at the decoder resulting in a contour being reduced to a contour represented by only 256 shades. Similarly, if the numerical values of the array are only byte

wide, the coefficients themselves effect the accuracy of the matrix math at the decode side. In summary image quality is compromised. To what extent image quality is compromised depends on the images in the scene, how the image analyzer of the decoder evaluates and forms macroblocks, the resolution of its prototype matrix, the resolution of the decoder prototype matrix and how high it was necessary to set the "zeros" threshold to pass through the transport bandwidth. In other words, it isn't predicable. It will have different effects in different areas of an image and different from frame to frame depending on what is happening in the scene that is being captured.

In real time, latency is of concern. As can be seen there is a lot of analysis and math to do in order to create a compressed video stream. The higher the image resolution and the more compression is required to fit into the transport path, the more work has to be done. However, each image is arriving from the imager at the frame rate, 60 frames per second (FPS) for example. Therefore, without regard to the amount of work to be done, the time to accomplish it is 16.66 milliseconds. This is true at the decode side as well in order to reproduce temporal (motion) fidelity to the original scene. Therefore, as one goes from SD to 720, to 1080 and 4K and from 8-bit, to 10-bit to 12-bit sampling the processing power needed increases in proportion at both ends of the path. MPEG H.264 and H.265 encoding relies heavily on differential frames. Differential frames are comprised of only the pixels that change from a reference (I) frame in an MPEG encoded stream. If there are many changes in images between frames, the encoder will create a new I frame. If the changes are small, it will create difference frames (B) which result in highly compressed frames, perhaps 1000:1. Long strings of frames having small changes can create very compressed encoded streams which then enables transport of narrow bands and enables storage of movies on the limited storage available on Blu-ray. MPEG HEVC (H.265) was driven into existence by the desire to distribute 4K movies on Blu-ray. A two hour 4K movie must be compressed by more than 200:1 to fit. If all of a 1,000 Gb/s Ethernet channel were available, it would still require more than 20:1 compression to have a chance at sustained transport.



The techniques used by MPEG require buffering at the encode end in order to create B frames[2]. How many depends on how many differential frames are created before a new I frame is needed. The more B frames (and P), the more buffering. Similarly at the decode side the more B and P frames are in between I frames, the more buffering is needed for the decoder is able to render fully recreated frames at the source frame rate. The data rate available to move the encode signal across the channel and the variability of the available data rate adds the buffer size needed to render smooth video. We have all experienced this with our streaming devices when we see "buffering" symbols on our screens. The delay from source video to displayed video is the latency and can be as little as 50 milliseconds but can be as long as 10 full seconds depending on scene change complexity, transport channel width and the computing power of the encoder and decoder pair in the chain. Therefore, another compromise is the introduction of latency. What latency can be tolerated depends on the mission. As a TV viewer, 10 seconds of latency as long as the displayed frame rate can be supported has no impact. However, if the application is real

[2] A P frame is another type of differential frame where content of a future frame and a previous frame are to used to place pixels in this predictive frame.

time control with a man in the loop, latency must be less than 200 milliseconds. Lower latency can be achieved by a higher bandwidth in the transport channel, high performance equipment and reduced image quality. Compromises are needed to find the best fit for the mission at hand.

Using a smaller field of view (more zoom) can also improve image quality. When zooming in, the shades contour of each macroblock may be simpler and naturally generate more low value coefficients in the DCT array. This will improve decoded image fidelity at the expense of field of view.

Color errors are induced by subsampling and reduced pixel bit depth as compared to the image samples delivered by the source camera. Reduced bit depth reduces the number of shades of intensity and color that each pixel can represent. This can cause visible color gradient bands and reduce the detail that is represented. With fewer colors available to represent, color accuracy is compromised as well.

Subsampling can introduce its own color errors at edges. In this close up of a 4:2:2 subsampled color bar pattern the edges of the color bars are clearly neither the previous bar's color nor the subsequent bar's color. This error is caused by the sharing of one color sample for two pixels and using the color sample to derive the amount of green to use in the green bar and the magenta bar.

It is important to understand these effects when specifying camera, CODEC and transport components. Color errors may negatively impact an investigation or analysis. It is another tradeoff to make.

CONCLUSION

Compression is a necessary evil. It is a tool that enables the use of available transport and storage mediums. However, image quality is impacted by the process. Compression has the effect of reducing the resolution of the source image, a compromise that must be considered when investing in high resolution cameras and lenses.

Compression can eliminate and even add features in the original image. Increased transport bandwidth, careful selection of CODECs and increasing the performance of the encoding and rendering hardware and software can reduce these effects. Reducing the field of view of the camera (more zoom) can also help by reducing the complexity of the shades contour of macroblocks.

Similarly latency can be reduced by careful selection of the CODEC pair, increased bandwidth of transport, the power of the computing platforms at both ends and the target bit rate of the encoded output.

Using the compression tool is necessary, but it adds many compromises to consider when designing and specifying an image capture system for testing, analyzing and observing the behavior of things.

BIBLIOGRAPHY

[i] From Wikipedia: **Dailies**, in filmmaking, are the raw, unedited footage shot during the making of a motion picture. They are so called because usually at the end of each day, that day's footage is developed, synced to sound, and printed on film in a batch (or telecined onto video tape or disk) for viewing the next day by the director, some members of the film crew, and some actors. Dailies serve as an indication of how the filming and the actors' performances are progressing.https://en.wikipedia.org/wiki/Dailies - cite_note-1 However, the term can be used to refer to any raw footage, regardless of when it is developed or printed.

[ii] From Wikipedia: Computer-generated imagery (CGI) is the application of computer graphics to create or contribute to images in art, printed media, video games, films, television programs, shorts, commercials, videos, and simulators. The visual scenes may be dynamic or static and may be two-dimensional (2D), though the term "CGI" is most commonly used to refer to 3D computer graphics used for creating scenes or special effects in films and television.