

## 1 Scope

As the ISR Task Force pushes the community towards a digital high definition (HD) future, the services and COCOMS maintain that the migration to HD will cause various networks to fail because HD signals require more bandwidth than the current communications channels can support. By pre-processing an HD signal prior to compression using image cropping and scaling operations the resulting compressed signal can meet the bandwidth constraints of a given channel. A properly scaled HD motion imagery signal has the added benefit that it will be of higher quality than imagery from a sensor with that same spatial pixel density.

This EG addresses tradeoffs in latency and image quality of H.264 encoding that meet channel bandwidth limitations. The guidelines are based on image formats and their accompanying data rates as found in the MISM (Motion Imagery System Matrix) tables of the MISB, and subjective evaluations using an industry software encoder and several commercial hardware encoders. Data compression is highly dependent on scene content complexity, and for this reason the evaluation is based on two types of content: 1) pan over a multiplicity of high contrast, fast moving objects (people) and fine-detailed buildings; and 2) aerial imagery of planes, ground vehicles, and terrain typical of UAV collects. While the derived data rates may not reflect all types of scene content, they do serve as practical guidelines. If changes are in order, the suggested changes should be brought to the MISB for incorporation into the MISB and this EG. Vendors are encouraged to validate the practical implementation of the processing methods suggested.

Note on image nomenclature: Image formats discussed include progressive-scan imagery only. For this reason, the “p” generally applied as a suffix when describing progressive-scan formats (for example, 1080p and 720p) is suppressed.

## 2 Informative References

- [1] ITU-T Rec. H.264, Advanced video coding for generic audiovisual service / ISO/IEC 14496-10 *Information Technology - Coding of audio-visual objects Part 10: Advanced Video Coding*
- [2] MISB EG 0802, *H.264 Coding and Multiplexing*, May 14, 2009

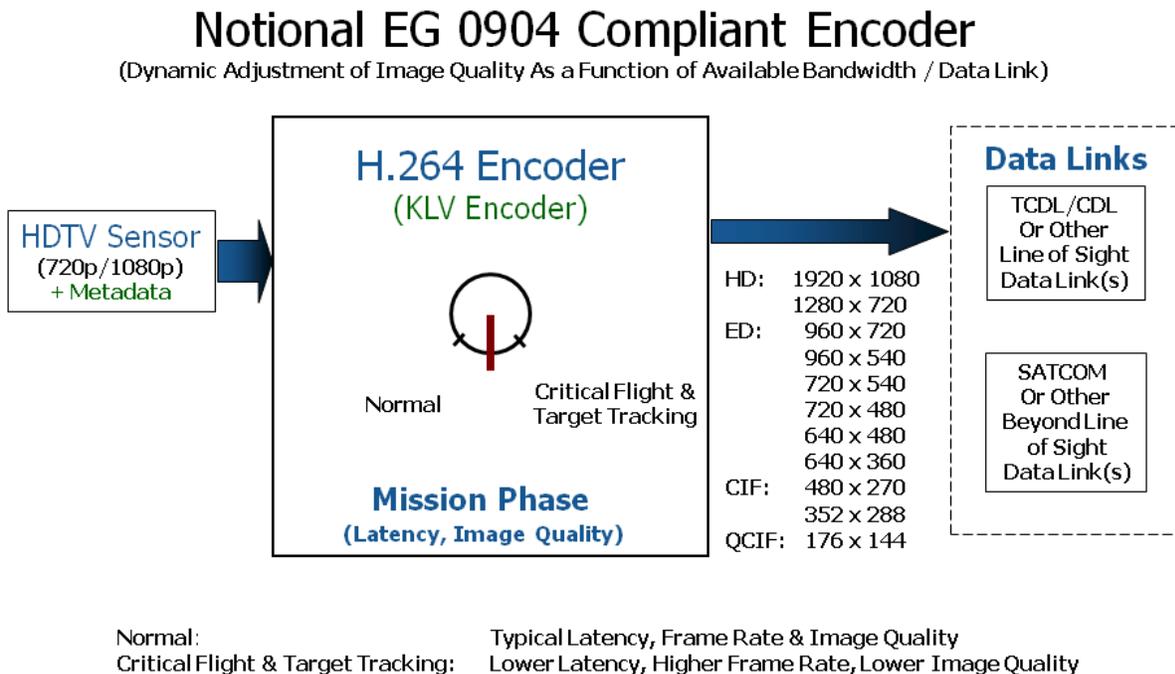
## 3 Acronyms

HD High Definition

KLV	Key-Length-Value
MISM	Motion Imagery Systems Matrix
MISP	Motion Imagery Standards Profile
TCDL	Tactical Common Data Link

## 4 Introduction

Consider an adjustable, image format encoder of Figure 1 designed to accommodate a prescribed data link bandwidth. Here, a high definition sensor produces High Definition (HD) video content of 1920x1080 or 1280x720 format. An operator can set the encoder to meet a specific channel data rate (if the channel rate is known), or the encoder can be set automatically if sensing of the data link rate is provided. In either case, numerous image format changes can reduce the encoded bit rate to be less than that of the HD sensor's signal. To achieve a desired target bit rate the imagery will be spatially and/or temporally modified. Supported spatial formats are denoted as HD, ED, CIF, and QCIF.



**Figure 1: Adjustable Image Format HD Encoder**

These image formats offer a good selection of options for optimizing spatial formats, encoder/decoder latency, and image quality. Figure 1 suggests a structure for dynamically altering the image format and mission requirements as a function of available data link bandwidth. Beyond meeting the channel requirements, this new functionality provides versatility in changing the imagery characteristics based on real-time in-flight mission needs.

Two modes of operation are suggested: a normal mode, which is used for most situations; and a low latency, high frame rate mode, which could be used for critical flight operations and target tracking. Although a “critical imaging” mode was considered, it is recommended that a doubling of the normal mode data rate be used in critical imaging applications for a given image format.

Table 1 specifies levels of encoder capability that should aid both vendors in meeting design goals and those in acquisition making purchasing decisions. Levels are graded from Fully Compliant, which is most desirable and includes the ability to meet the best performance HD can deliver as well as support all the reduced spatial formats that accommodate lower bandwidth data links, to Minimally Compliant, which should allow HD sensor video delivery over the same data link that can support standard definition video today. Not all current encoders may achieve Level A capability today, but will do so within the next several years.

Capability	Compliance Description	Format Support	Encoder Capabilities
Level A	Fully Compliant	1920x1080x60 or 1920x1080x30 or 1280x720x60	Supports HD, ED, CIF & QCIF
Level B	Partially Compliant	1920x1080x30 or 1280x720x60	Supports HD, ED, & CIF
Level C	Less Compliant	1920x1080x30 or 1280x720x60	Supports HD & ED
Level D	Minimally Compliant	1280x720x30	Supports HD & ED

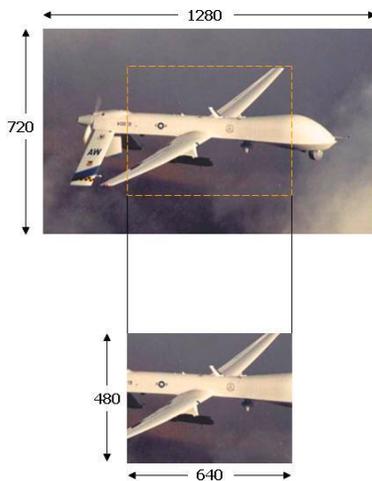
**Table 1 - HD Formats and Encoder Capabilities**

## 5 Spatial Format Reductions

### 5.1 Image Cropping

In image cropping, a smaller sub-area *within* the sensors field of view is extracted for encoding. For example, if the HD sensor field of view is 1280x720 (horizontal pixels x vertical pixels), extracting sub-areas of 640x480 and 320x240 would produce imagery that matches the pixels of the original imagery within the respective sub-area. This reduced-size image represents a reduced field of view with respect to the original. In this case, the 1:1 pixel aspect ratio of the HD source image is maintained, so that geometric distortion does not occur—for instance, circles in the original remain circles in the sub-area image. As indicated in the Figure 2, source image content is lost in cropping.

Cropping is generally not recommended because the image coordinates will change, thereby affecting a re-calculation of metadata describing such. The only cases where cropping is difficult to avoid is in converting a 16:9 HD image format to standard definition and derivative sizes that are normally represented with a 4:3 aspect ratio.



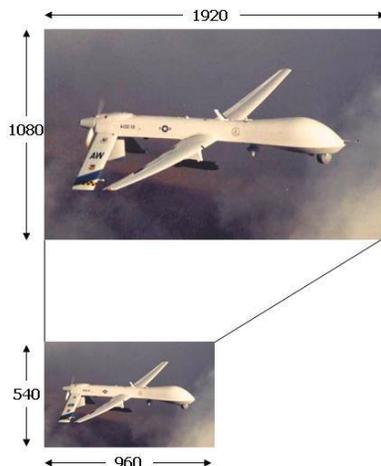
**Figure 2: Image Cropping Example**

Sub-image cut from the full image field, where the pixels within the sub-image are equivalent to those within the original image.

## 5.2 Image Scaling

In image scaling, the sensor field of view is preserved, but possibly at the expense of the spatial frequency content, which may be reduced. For example, if the HD sensor field of view has a format of 1920x1080 pixels a scaling by one-half in each dimension would produce an image with a format of 960x540 pixels. To preserve the image aspect ratio (horizontal to vertical size) each image dimension must be scaled by the equivalent amount. This will ensure that geometric shapes like circles in the original image remain circles in the scaled image. Square pixels are preserved.

While image cropping requires nothing more than a simple remapping of input pixels to those within a target output sub-area, video scaling may require pre-filtering of the image. Simple techniques such as pixel decimation and bilinear filtering can produce image artifacts: in pixel decimation image aliasing can cause false image structure, and thus result in poor compression; bilinear filtering may reduce image fidelity, particularly when used with large scaling factors. More information on image scaling can be found in Appendix 2. The following example illustrates an equal scaling in each dimension of an image. Note that the output image looks identical to the input, except smaller.



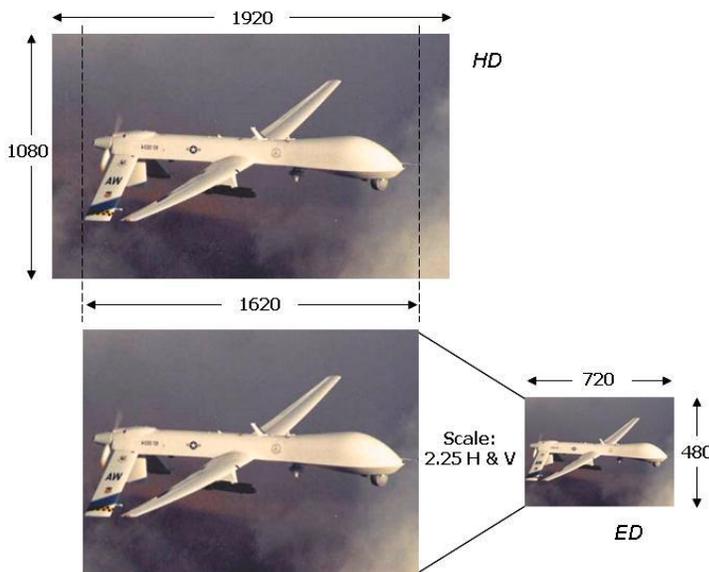
**Figure 3: Image Scaling Example**

New image filtered and scaled, where the original field of view is maintained.

### 5.3 Image Crop & Scale Example

Illustrated next is an example of combining cropping and scaling to convert a 1920x1080 HD format image to a 720x480 ED format image. In this case, the goal is to maintain the square pixel relationship of the original image in the scaled image, so that there is no geometric distortion. To do so necessitates that a certain amount of the original image is cut off; this can be done equally to each side as shown in the example, or taken completely from one side or the other, thereby skewing the image to that side.

This type of conversion is very typical of current home experiences in watching high definition content on a standard definition television receiver. The image on each side is cut off and not visible to those with standard definition receivers.



**Figure 4: Image Crop & Scale Example**

1920x1080 HD image is first cropped to 1620x1080, and then equally scaled by 4/9 both horizontally and vertically to produce a 720x480 pixel ED format image.

## 6 Target Image Formats

### 6.1 HD Format

The most popular High Definition (HD) formats are 1920 horizontally by 1080 vertically, and 1280 horizontally by 720 vertically. Both spatial formats exhibit square pixels (square pixel aspect ratio); this means that the ratio of horizontal to vertical size of each pixel is 1:1.

It is important to understand what format the HD sensor is producing, so that informed tradeoffs can be made for the system CONOPS. Given a high definition sensor source, a high definition H.264 bit stream can be delivered when there is sufficient channel bandwidth. However, in cases where the channel bandwidth is insufficient to communicate the full HD stream, there are several approaches to reduce the data rate to meet the bandwidth constraints. These approaches do impact an encoders design, and not every option may be available from a given manufacturer.

Oftentimes, an encoder may be set to encode the content at a reduced frame rate, or at reduced image fidelity. However, if a reduced frame rate is not sufficient to portray the temporal information appropriately, or if image fidelity is reduced where salient objects are lost then neither solution is sufficient. In these cases, the image spatial format can be changed to accommodate the needed reduction in data rate.

## 6.2 ED Format

Four formats are defined for the Extended Definition (ED) image format.

An ED image of 960 pixels horizontally by 540 vertically can be obtained by simply scaling a 1920x1080 HD image by 1/2 in each dimension, or scaling a 1280x720 HD image by 3/4 in each dimension.

In scaling an ED image to 720x540 pixels the image must be cropped and then scaled to maintain square pixels. A 1920x1080 HD source image is first cropped to 1440 pixels horizontally and then scaled by 1/2 in both the horizontal and vertical directions. A 1280x720 HD source image is first cropped to 960 pixels horizontally and then scaled by 3/4 in both the horizontal and vertical directions.

An ED image of 720 by 480 pixels is closest in pixel density to that of a standard definition television format, but in progressive rather than interlace format. This progressive ED format can be obtained in two ways: 1) horizontally crop a 1920x1080 HD image to 1620x1080 and then scale by 4/9 in both the horizontal and vertical dimensions; or, 2) horizontally crop a 1280x720 HD image to 1080x720 and then scale by 2/3 in both the horizontal and vertical dimensions. Both methods guarantee that square pixels are preserved in the final image. Note that the final image aspect ratio in this case is neither 16:9 nor 4:3, the standard television aspect ratio.

Finally, an ED image of 640x480 pixels can be obtained in two ways as well: 1) horizontally crop a 1920x1080 HD image to 1440x1080 and then scale by 4/9 in both the horizontal and vertical dimensions; or, 2) horizontally crop a 1280x720 HD image to 960x720 and then scale by 2/3 in both the horizontal and vertical dimensions. An ED image of 640 by 480 pixels has both square pixels and a 4:3 image aspect ratio.

## 6.3 CIF Format

Two formats are defined for the Common Intermediate Format (CIF) image format: 480 pixels horizontally by 270 pixels vertically, and 352 by 288 pixels. Here, the uncommon format of 480 by 270 is grouped in the CIF category for convenience. Two methods can be used to create a 480x270 image from an HD original: 1) scale a 1920x1080 HD image by 1/4 in each spatial dimension; or 2) scale a 1280x720 HD image by 3/8 in each spatial dimension.

A 352x288 image is derived as follows: 1) horizontally crop a 1920x1080 HD image to 1320 pixels, and then scale by 4/15 in each spatial dimension, or 2) horizontally crop a 1280x720 HD

image to 880 pixels, and then scale by  $2/5$  in each spatial dimension. Both methods produce images with square pixels.

## 6.4 QCIF Format

One format defines the Quarter Common Intermediate Format (QCIF) image format: 176 pixels horizontally by 144 pixels vertically. This format cannot be derived by scaling either of the HD formats directly; doing so will produce non-square pixels, and induce geometric distortion. Two methods can be used to create a QCIF image from the HD original: 1) horizontally crop a 1920x1080 HD image to 1320 pixels, and then scale by  $2/15$  in each spatial dimension; or 2) horizontally crop a 1280x720 HD image to 880 pixels, and then scale by  $1/5$  in each spatial dimension. Both methods will produce images with square pixels.

## 7 Aspect Ratio

*Pixel aspect ratio*—expressed as a fraction of the horizontal (x) pixel size divided by the vertical (y) pixel size—has already been mentioned. The pixel aspect ratio for square pixels is 1:1. This is not to be confused with *picture aspect ratio*, which is a fraction of total horizontal (x) picture size over total vertical (y) image size, for a stated definition of "image." Standard-definition systems, such as NTSC, have a 4:3 image aspect ratio, while high-definition systems have a 16:9 image aspect ratio.

Standard definition systems do not have square pixels, so unless the proper geometric conversion is done, arbitrary scaling will result in non-square pixels. High definition systems have square pixels, so that as long as scaling is done equally in both the horizontal and vertical dimensions, the scaled sub-area will also have square pixels. Cropping a sub-image from either a 4:3 or 16:9 image preserves the pixel aspect ratio of the original image. So, if a 4:3 original image has non-square pixels, then a cropped sub-image will also have non-square pixels. Likewise, if a 16:9 image has square pixels, then a cropped sub-image will also have square pixels.

When cropping and scaling it is important to understand the effects. A 640x480 image cropped from a high definition image will have square pixels, but have a *different* picture aspect ratio than the original. On the other hand, scaling an HD image to 640x480 will change *both* the picture aspect ratio *and* the pixel aspect ratio. Converting a 1280x720 high definition image to a 640x480 enhanced definition image requires cropping out a 960x720 sub-area from the HD image, and then scaling the sub-area image by  $2/3$  equally in the horizontal and vertical dimensions. This will maintain the pixel aspect ratio (square pixels), while changing the picture aspect ratio from 16:9 to 4:3 (1280:720 versus 640:480).

## 8 Latency and Quality

Users also need to make tradeoffs between latency and image quality. Reducing latency generally results in a lower image quality for a given data rate. When the lowest possible latency is required for a given bandwidth image format reductions may be necessary. In exploring these tradeoffs, the data rates in the MISP RP 9720 provide an initial point of reference. A quality encoder should produce good-to-excellent quality using the numbers from RP 9720. Switching to a low-latency mode will reduce the quality to fair-to-good at that same data rate.

First consider latency. Better H.264 encoder/decoders can produce good quality video with less than 250 milliseconds latency. *Of course, total system latency is dependent on the communications link and other system factors, and can be considerable.* It is outside the goal of this EG to consider latency contributions from other system elements. For this EG, an encoder and decoder pair that produces high-quality motion imagery in a given bandwidth exhibiting a latency of 250 ms or less will be considered low latency, and one that exhibits a latency of greater than 250 ms will be considered normal latency.

For convenience MISP RP 9720e is replicated here in Appendix 1 with the *Data Rate Range* row highlighted. Note that the horizontal pixel count across the various MISM levels extends from 160-176 pixels, to 320-352 pixels, and finally 640-720 pixels. As examples, for a data rate less than 56 Kb/s a format of 160x120x5 can be used; for a data rate within the range of 56-192 Kb/s a format of 320x240x5 can be used; for a data rate within 192-384 Kb/s a format of 320x240x15 can be used, and so forth.

In addition to the normal latency mode, there is the need for low latency at the highest frame rate possible. Example use cases include critical mission takeoff, landing and target tracking. Note that switching to the low latency mode lessens the picture quality. Understand that reducing the spatial pixel count by one-half and doubling the frame rate may not produce quality video at an equivalent bit rate. The relation in trading one dimension of pixel count for another in order to effectively produce the same data rate is highly dependent on scene content, motion, and encoder design.

STANAG 4609 and MISP-compliant systems can only originate progressive (non-interlaced) motion imagery. Efforts were made in Table 2 to favor scaling rather than cropping when possible, and to maintain a square pixel aspect ratio. Data rates of 10.74 Mb/s, 3.0 Mb/s and 1.5 Mb/s are commonly found for TCDL. Table 2 lists data rates that meet these channel rates. For some image formats these target data rates stretch towards the lower end (minimum) where image quality suffers. In time-critical missions a 60 Hz frame rate can be used for the low latency mode. The normal operation may be at 30 Hz as seen from the table..

*Use of interlace-scanned sensors are rejected by the MISP because interlace scanning will introduce temporal artifacts into the imagery. Should there be a need to display a progressive-scanned image on an analog interlaced display, hardware is available that converts from progressive-digital to analog-NTSC for display. With many older analog-interlaced displays now replaced by flat-panel progressive displays, or by computer workstations, a converter is unnecessary.*

**Table 2 - Data Rates for Common Image Formats (H.264)**

<b>Video Data Rate (nominal)</b>	<b>Normal Operations</b>	<b>Critical Mission</b>
	Latency: <i>typical</i> Frame rate: <i>typical</i> Image quality: <i>higher</i>	Latency: <i>lower</i> Frame rate: <i>higher</i> Image quality: <i>lower</i>
10 Mb/s	1920x1080x60	1920x1080x60
6 - 8 Mb/s	1920x1080x30	1280x720x60
4.5 - 6 Mb/s	1280x720x60	960x720x60
3 - 4.5 Mb/s	1280x720x30	960x540x60
2 - 3 Mb/s	960x540x30	720x480x60
1.5 - 2 Mb/s	720x540x30	640x360x60
1 - 1.5 Mb/s	640x480x30	480x270x60
512 Kb/s - 1 Mb/s	480x270x30	480x270x30
384 - 512 Kb/s	352x288x30	352x288x30
256 - 384 Kb/s	352x288x15	176x144x30

**Note:** Image formats shaded in green represent sub-areas cropped from the source image field of view. Scaling directly to these formats will cause geometric pixel-aspect ratio distortion.

## 9 Metadata Considerations

A MISP-compliant stream contains both motion imagery *and* metadata. The metadata carries reference information for the acquired imagery, which must be modified when format changes like cropping are made to the imagery. For example, image corner and center points are in relation to the image format that is encoded, so these must be recomputed based on the image format that will be coded.

### 9.1 Metadata Associated with Dropped Frames

In modes where frames are dropped, for example, changing the temporal rate of 60 frames per second imagery to 30 frames per second, metadata associated with frames that are dropped must still be preserved. It is recommended that for synchronous metadata, where the PTS for the metadata corresponds to a dropped frame video, that those PTS values be modified and made the same as the last valid frame PTS value.

### 9.2 Stream Overhead

Often carried along with the compressed video is metadata and possibly audio. These essence streams of video, metadata, and audio are packaged within a MPEG2 Transport Stream for carriage off-platform. Additional overhead from the metadata, audio, and transport stream

packaging must be considered in any calculation of data rate. Said another way, the effective bandwidth of the data link is decreased for video based on how much metadata and audio is present, and how much overhead the MPEG2 transport stream carrier consumes.

To gain an appreciation of this overhead worst case, assume that the EG0601 local data set (LDS) is the structure used and it is fully populated. In this case, each LDS will be roughly 1200 bytes in size. If one LDS is present for every corresponding frame of video the metadata rate will be 1200 bytes/frame x 60 frames/sec = 576 Kbits/sec. For audio, assume a data rate of 16 Kbits/sec. The transport stream overhead is assumed to add 3-4% of the combined essence streams. Thus, the TS overhead is roughly  $(576 \text{ Kbits/s} + 16 \text{ Kbits/s} + V \text{ Kbits/s}) \times 4\%$ , where  $V$  = the video data rate. As an example, the overhead for the metadata, audio, and TS is 856 Kbits/s for  $V=6 \text{ Mb/s}$  and 616 Kbits/s for  $V=1 \text{ Mb/s}$ . As a percentage of the available bandwidth the 6 Mb/s video will consume roughly 88% of the bandwidth, while the 1 Mb/s video will consume only 62%. See RP 0902 for discussions of the minimum metadata set and an example of minimum metadata at 9.6 Kb/s. These calculations show that at higher video data rates the overhead from metadata, audio, and the transport stream is a smaller percentage of the total data rate. As the video data rate decreases the contribution of the metadata begins to dominate, thus suggesting that a conscious tradeoff in video format versus metadata must be made.

## 10 Conclusions

This EG examines what can be done if an HD sensor is available, but the bandwidth in the communications channel does not support HD. A choice in spatial format and temporal rate can be selected from Table 2 that will meet the data link constraints depending on the application of use. The two categories shown in Table 2 allow for tradeoffs in data rate or bandwidth, latency of the encoder/decoder, and received image quality. The guidelines presented here offer suggested image formats and options based on current knowledge of product capabilities and performance. As the assumptions made here become tested this EG will refine its guidelines accordingly.

## Appendix 1: RECOMMENDED PRACTICE 9720e - MISM, Low Bandwidth Motion Imagery

System Level	MISM							
	L2.2H	L2.1H	L2.1M	L2.0M	L1.3H	L1.2H	L1.1H	L1.0H
Common Description/ Intended Application	Medium / Distribution	Low-Medium / Distribution		Low / Distribution	Low / Distribution	Very Low / Distribution	Very Low / Distribution	Lowest / Distribution
System Attributes: Spatial Definition	Medium	Low - Medium		Low	Low	Low	Low	Very Low
System Attributes: Temporal Definition	Medium	Medium		Medium	Medium	Low	Very Low	Low
System Attributes: Generation Resiliency	Low	Low		Very Low	Very Low	Very Low	Very Low	Lowest
Applicable Standard (Note: Other Profiles /Practices may apply)	H.264 L2.2	H.264 L2.1	MPEG2 MP@ML	MPEG-1	H.264 L1.3	H.264 L1.2	H.264 L1.1	H.264 L1.0
Horizontal Resolution (Nominal)	640 - 720	320 - 352		320 - 352		320 - 352		160 - 176
Vertical Resolution (Nominal)	480 - 576	480 - 576		480 - 576	240 - 288p	240 - 288p		120 - 144p
Bit Depth (bits) (Nominal)	8	8		8		8		8
Frame Rate (FPS)	24 - 30	24 - 30		24 - 30		12 - 15	6 - 7	12 - 15
Compression Ratio (Nominal)	110:1	83:1	55:1	83:1	83:1	83:1	83:1	166:1
Data Rate (Nominal)	1.5 Mb/s	1.0 Mb/s	1.5 Mb/s	1.0 Mb/s	512 Kb/s	256 Kb/s	128 Kb/s	32 Kb/s
<b>Data Rate Range (Kbits/s)</b>	<b>1,024 - 1,500</b>	<b>768 - 1,024</b>	1,024 - 1,500	768 - 1,024	<b>384 - 768</b>	<b>192 - 384</b>	<b>56 - 192</b>	<b>&lt; 56</b>
Candidate Transport Channel (Nominal Rates)	T1/ E1	T1/ E1		T1/ E1	Partial T1/E1	Wireless	Wireless	Wireless
Allowed Transport Protocols	Xon2	Xon2		Xon2	Xon2 RTP/RTSP	Xon2 RTP/RTSP	Xon2 RTP/RTSP	Xon2 RTP/RTSP
Recommended Transport Protocols	Xon2	Xon2		Xon2	RTP/RTSP	RTP/RTSP	RTP/RTSP	RTP/RTSP

## Appendix 2

Note: The following is meant as way of an introduction to the causes and resulting artifacts that may occur when scaling the size of an image.

### A2.1 Image Scaling

Image scaling is a signal processing operation that changes an image's size from one format to another, for example 1280x720 pixels to 640x360 pixels. An image that has a large number of pixels does not necessarily imply a higher fidelity image over one with fewer pixels. For instance, if you focus your camera and take a picture that produces a high fidelity sharp image, and then take that same picture with the lens of the camera defocused both images will be the same size yet have a very different look; one is sharp and one is blurred. Obviously, the equivalent size of the images did not translate into the equivalent fidelity. So, size does not necessarily mean better. What is more important is what the pixels convey.

Images are made up of a number of different frequencies much like that in a piece of music. However, whereas music is a one-dimensional temporal signal, an image is a two-dimensional spatial signal with horizontal (across a scan line) and vertical (top to bottom) frequency components. Video, made from a series of images in time adds yet a third dimension of frequency (temporal frame rate). The combination of horizontal, vertical, and temporal frequency components constituting a video signal is termed the spatio-temporal frequency of a video signal.

To simplify the discussion of frequency as related to the number of pixels consider the horizontal dimension of an image only. Each pixel along a scan line can take on a value independent of its neighboring pixels. The maximum change that is possible from pixel to pixel occurs when sequential pixels transition from full-on to full-off or zero intensity (black) to 100% intensity (white) or vice versa. We call the transition in intensity with adjacent pixels as one complete cycle—black to white or white to black, for example. Conversely, when sequential pixels remain the same value (all pixels are one shade of gray, for example) there is no change across the line and no frequency change as well; this is defined as zero frequency. Thus, across a scan line neighboring pixels can vary between some maximum frequency and zero frequency. Horizontal frequency is specified as a number of these cycles per picture width (c/pw).

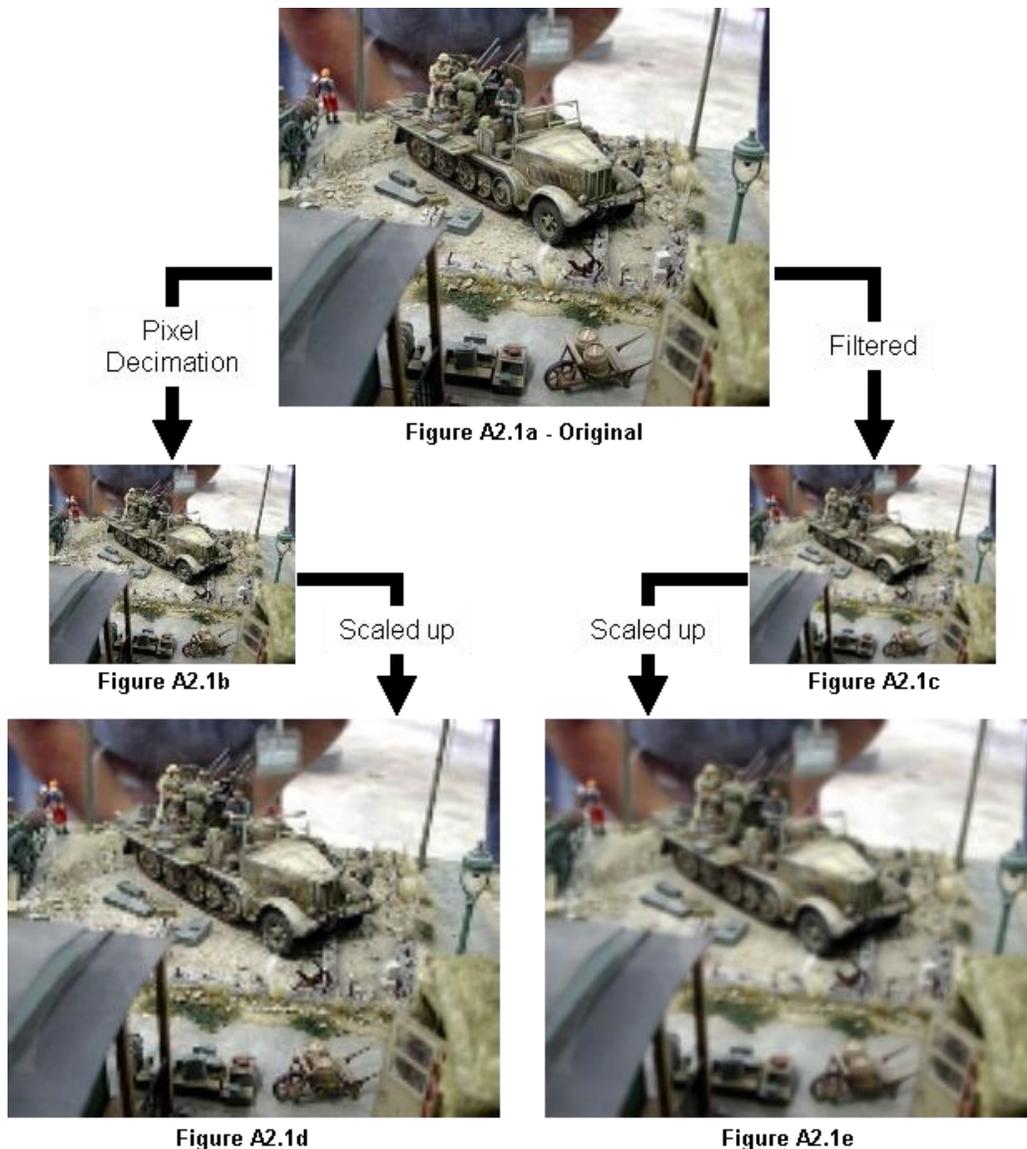
Similarly, the same holds true for vertical pixels within a column of an image. Vertically, frequencies are specified as a number of cycles per picture height (c/ph). In the temporal domain, the maximum frequency is governed by the frame rate, and this is expressed in frames per second, or Hertz.

In the case of our focused picture example above, the pixels within the image will have significant change with respect to one another, while the defocused picture will have much less change amongst neighboring pixels. The lens on a camera acts as a two-dimensional filter, which has the ability to smear the received light from the scene onto groups of pixels on the

image sensor. In effect, this is similar to averaging a neighborhood of pixels and assigning a near constant value to them all.

To gain an appreciation for the artifacts that image scaling can cause consider what would happen in the example above if each successive pixel across a horizontal line changes from zero to 100% intensity. If this were done for every scan line the image would look like a series of vertical stripes each one pixel wide. What would happen if the image is then scaled by one half horizontally, where every other pixel is eliminated? If the eliminated pixels are the zero intensity ones the resulting image would be all white, while if the eliminated pixels are the 100% intensity ones the resulting image would be all black. Obviously, the final scaled image does not resemble the original image. This artifact is called aliasing; so named because the resulting frequencies in the signal are completely of a different nature than what they were originally.

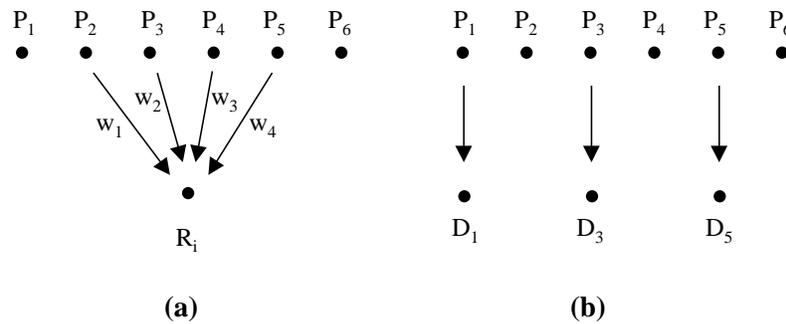
An example of aliasing is shown in Figure A2.1 below.



The original image in Figure A2.1a is scaled by one half in each dimension using pixel decimation (elimination) (Fig A2.1b) and filtering (Fig A2.1c). To emphasize the artifacts induced by both techniques, the images are shown up-scaled by two in Figures A2.1d and A2.1e. Although the filtered image appears less sharp, it has far fewer jaggies and artifacts that will impact compression negatively.

Filters are signal processing operations used to control the frequencies within a signal, so that functions like scaling do not distort the information carried by the original signal. A two-dimensional (2D) filter can remove the spatial frequencies that cannot be supported by the remaining pixels of a scaled image. A 2D low-pass filter, which acts as a defocused lens, is an integrator that performs a weighted average of pixels within sub-areas of an image. This integration prevents aliasing artifacts. How the integration is done is critical in preserving as much of the image frequency content as possible for a target image size. Some types of integration can create excessive blur or excessive aliasing — both undesirable. Blur will reduce image feature visibility, while aliasing will produce false information and reduce coding efficiency.

The number of pixels over which a 2D filter operates may be as few as 2x2 (two pixels horizontally by two pixels vertically), which is simply averaging of the four pixels to produce a new one. Such small filters are computationally efficient, but do a poor job in general. 2D filters that do a better job retain as much image fidelity as possible, and typically include many more neighboring pixels to determine each new scaled output pixel. Figure A2.2a illustrates a collection of weighted pixels  $P_k$  in the horizontal direction that sum to a new output value  $R_i$ , while A2.2b shows a direct scaling by two without any filtering. The weights ( $w_1-w_4$ ) are numerical values that are multiplied by corresponding pixels and the results added to form a new output pixel. For example, in A2.2a, the output pixel  $R_i = w_1 * P_2 + w_2 * P_3 + w_3 * P_4 + w_4 * P_5$ .



**Figure A2.2 (a) Input pixels  $P_k$  Filter taps  $w_1-w_4$  and Filtered Output pixel  $R_i$   
 (b) Direct scaling by one half**

Alternate pixels across the line are eliminated in direct scaling by one half horizontally. In this case, the contributions from pixels  $P_2$ ,  $P_4$ , etc. are completely lost along with valuable information they carried.

## A2.2 Spatio-Temporal Frequency

Video is a three-dimensional signal with spatial frequencies limited by the lens, the sensor's spatial pixel density, and temporal frequencies limited by the temporal update rate. This collection of 3D frequencies constitutes the spatio-temporal spectrum of the video signal. Scaling in the temporal domain, such as changing from 60 frames per second to 30 frames per second, is usually accomplished by directly dropping frames rather than applying a filter first. Our focus, therefore, is filtering as applied in the 2D spatial horizontal and vertical dimensions.

When viewed from the frequency perspective, the image will contain horizontal frequencies that extend from zero frequency to some maximum frequency limited by the number of horizontal pixels, and likewise, vertical frequencies that extend from zero frequency to some maximum frequency limited by the number of vertical pixels. The frequency domain is best understood using a spectrum plot as shown in Figure A2.3. The amplitudes of the individual component frequencies are suppressed in this figure, but would otherwise extend directly outward orthogonal to the page.

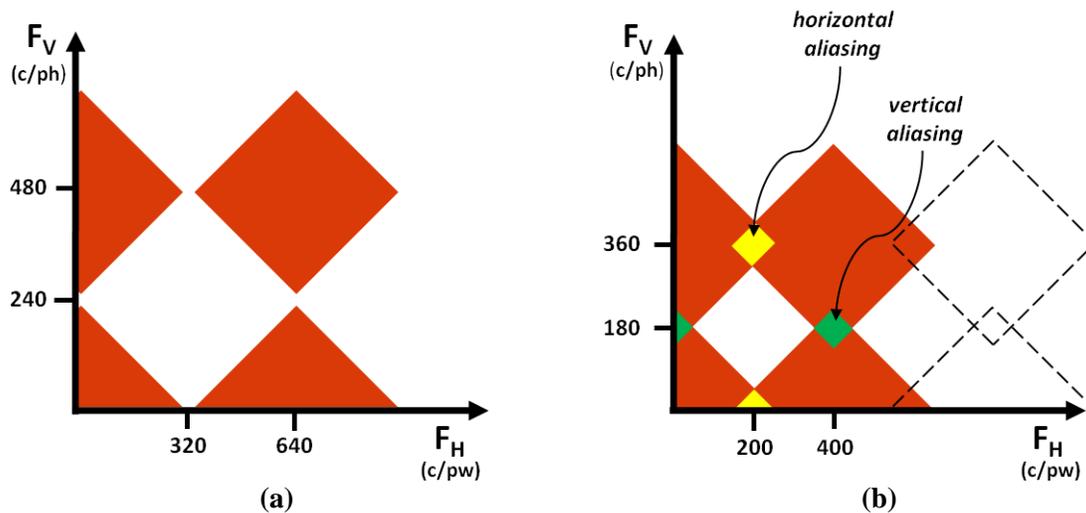


Figure A2.3 (a) HV Spectrum for a 640x480 image; (b) and re-sampled at 400x360

The value in portraying an image in the frequency domain is the ability to identify potential issues when applying a particular signal processing operation such as image scaling. In Figure A2.3a, the frequencies extend from Zero to less than 320 cycles-per-picture width (horizontal frequencies) and 240 cycles-per-picture-height (vertical frequencies). The amplitudes of the frequencies within this quarter triangle depend on the strength of each in the image. Sampling theory dictates that the maximum frequency be no more than one-half the sampling frequency. The sampling frequency for an image is fixed by the number of pixels, and since one cycle represents two pixels the maximum frequency is limited to half the number of pixels in each dimension. A 640x480 image will thus have frequencies no greater than 320 c/pw and 240 c/ph. Most video imagery is limited in spatial frequency extent by the circular aperture of the lens, and so the spectrum is rather symmetrical about the origin.

Sampling theory also indicates that a signal's spectrum is repeated at multiples of the sampling frequency. A digital image spectrum repeats itself at intervals equal to the picture width and picture height—its sampling frequency. For example, the horizontal spectrum of a 640 pixel image will repeat at intervals of 640 c/pw. The vertical spectrum will repeat at intervals of 480 c/ph for 480 pixels (Figure A2.3a).

If the horizontal, vertical, or temporal sample intervals are too close to one another as a result of scaling, or reducing the temporal rate, then the repeat spectra will overlap causing image artifacts. This interference produces cross-modulation frequencies that manifest themselves as aliasing (Fig A2.3b) and flicker. On the other hand, if an image is overly filtered, the image may become blurred because too many higher frequencies are attenuated. Scaling an image to a smaller size will re-position the repeating frequency spectrum's closer because the effective sampling frequency is lowered. A filter will limit the images' frequencies in a particular orientation, so that the image can be scaled with minimal artifacts.

### **A2.3 Rules of Thumb**

Scaling an image will cause artifacts when the resulting pixels can no longer support the frequencies within the image. The number and values of the filter weights determine the final quality of the scaled image. Too few weights may impose excessive blur. For good quality scaling between 100-50% (where 100% is the original image size and 50% is half the size in both the horizontal or vertical directions) five weights in the orientation of scaling is sufficient; nine weights are sufficient for scaling 50-25%. For drastic reductions of 25-12.5% 17 weights are preferred.

These rules of thumb are not required for manufactures to follow. They are only included for guidance. It is to be appreciated that vendors will provide their own value-added solutions.