

CONTRIBUTION TO IEEE STANDARDS SUBCOMMITTEE

COMMITTEE: G-2.1.6 Compression and Processing Subcommittee

TITLE: Overview of Picture Quality Measurement Methods

EDITOR: David Fibush

SOURCE: Tektronix

CONTACT: David Fibush
Tektronix Inc., MS 50-353
P.O. Box 500
Beaverton, OR 97077
(503)627-6289
(503)627-1707 (facsimile)
davef@tv.tv.tek.com

DATE: May 6, 1997

DISTRIBUTION: G-2.1.6 Compression and Processing Subcommittee

ABSTRACT: This contribution gives an overview of video quality testing definitions, subjective testing methods, objective testing methods, a system approach to objective testing and the description of a measurement instrument for application of the human visual system to objective measurements.

Overview of Picture Quality Measurement Methods

INTRODUCTION

Increasing use of compressed video methods in the generation and distribution of television programs has led to a requirement for objective picture quality measurement methods. Although traditional, indirect, signal quality measurements are still required for evaluation of the uncompressed part of the system, they are not adequate for evaluation of the compression-decompression process. Subjective testing methods are complex and time consuming and are only applicable for development purposes. They do not lend themselves to operational monitoring, production line testing or trouble shooting. Considering the variety of video compression methods and growing ease of data interconnection, measurement of resulting video quality must not be limited to a particular system and must be applicable for concatenation of several compression-decompression processes. This paper describes the feature-extraction and picture-differencing objective picture quality measurement methods with reference to on-going research developments. Emphasis is placed on application of the human visual system model to picture-differencing as the preferred objective method. System requirements for application of the model are defined in conjunction with a practical implementation of a measurement set.

DIGITAL COMPRESSION METHODS

Digital video became a reality in 1973 with the invention of the composite-based digital time base corrector for video tape recorders. In the early 1980s a worldwide digital component video standard was developed requiring 216 Mb/s or 270 Mb/s depending on the use of 8-bit or 10-bit sample values. This standard is commonly known as Rec 601. It is the dominant sampling structure for digital television and its use is showing rapid growth for all types of applications. Since the approval of the Rec 601 standard much research and development has been directed towards digital video data rate reduction resulting in a variety of video compression methods. Each of these compression methods has its own advantages, disadvantages and picture degradation characteristics. It will be important for any general purpose picture quality measurement instrument to provide a result that is independent of the compression method used. Some of the more important methods for individual picture compression are briefly described below.

Differential Pulse Code Modulation (DPCM) is a technique where only the changes in signal level are coded generally using a non-linear scale. A full pixel value is sent on a regular basis to reset the system in case of error. Some systems employ a "leakage" of the actual value to further reduce potential errors.

Sub-band Coding uses a bank of filters to separate the image into different frequency bands in two dimensions. Each frequency band may have a separate quantization and coding strategy. Selective decimation of certain frequencies and recursive application of single lowpass/highpass filter pairs are common techniques.

Wavelets are a generalization of Fourier theory. Inputs are decomposed into a basis which is all translations and dilations of a single function. The dilation (scaling) property is often used to create a pyramid of different spatial resolutions of each image. The decomposition process is implemented by filter banks (much like sub-band coding). Different quantization and coding strategies are applied to the different resolutions.

Fractals are mathematically generated descriptions (images) which look like complex patterns found in nature (e.g., the shoreline and topographic elevations of a land mass as seen from an aerial photograph). The coded image is built up from a number of different fractal equations. The key property of a fractal is self-similarity over different domain regions.

Vector Quantization is a technique where a vector (usually a square of samples of one color component of an image) are represented by a single number. This number is an index into a codebook by which the vector is reconstructed. Major issues are calculating a robust codebook and choosing the best codebook entry for a given vector.

Discrete Cosine Transform (DCT) coding is similar to wavelets and sub-band coding except that the DCT operates on blocks of pixels (generally 8 by 8) instead of the whole picture. The transform losslessly converts spatial data into spatial frequency data for the block. Compression is accomplished by using less bits (coarser quantization) for the higher frequency DCT coefficients. Many standardized compression systems for television and video conferencing use the DCT technique.

VIDEO QUALITY

There are three key testing layers for the modern television system; video quality, protocol testing of compressed data, and transmission system testing. Picture quality assessment is the new paradigm in video quality testing and must be seen as an integral part of complete compressed system evaluation. Several dimensions of video quality measurement methods are summarized in Table 1. *Subjective* measurements are the result of human observers providing their opinion of the video quality. *Objective* measurements are performed with the aid of instrumentation, manually with humans reading a calibrated scale or automatically using a mathematical algorithm.

Direct measurements are performed on the material of interest, in this case, pictures and are also called *picture quality* measurements. *Indirect* measurements are made by processing specially designed test signals in the same manner as the pictures and are also called *signal quality* measurements. Subjective measurements are only done in a direct manner since the human opinion of test signal picture quality is not particularly meaningful. (Of course, expert viewing of full field test signal pictures is useful as a way to determine signal distortions not for their aesthetic value.)

In-service measurements are made while the program is being displayed, directly by evaluating the program material or indirectly by including test signals with the program material. *Out-of-service*, appropriate test scenes are used for direct measurements and full field test signals are used for indirect measurements.

	In-Service	Out-of-Service
Subjective Direct (Picture Quality)	Program Material	Test Scenes
Objective Direct (Picture Quality)	Program Material	Test Scenes
Objective Indirect (Signal Quality)	Vertical Interval Test Signals	Full Field Test Signals

Table 1. Dimensions of video quality measurement

Although composite encoding represents a modest amount of compression, the NTSC and PAL systems are considered uncompressed in today's terminology. Signal quality (objective indirect) measurements are a reasonably good way to determine the picture quality for such uncompressed systems. That is, there is a good correlation between subjective measurements made on pictures from the system and objective measurements made on a suite of test signals using the same system. This is shown in the diagram of Figure 1. The correlation is not perfect for all tests. There are distortions in composite systems, such as false color signals caused by poorly filtered high frequency luminance information being detected as chroma. These distortions are not easily measured by objective means. Also there are objective

measurements which are so sensitive they don't directly relate to subjective results. However, such objective results are often very useful because their affect will be seen by a human observer if the pictures are processed in the same way a number of times. An example would be multiple generations using an analog video tape recorder.

The reason signal quality measurements work with analog and full bandwidth digital systems is that uncompressed systems are essentially linear. That is, the system behavior is time invariant, signal independent and superposition applies. Signal quality measurements are made with a suite of test signals whose resulting distortions will determine transmission channel or video processing characteristics. These test signals can be as short as one line in the vertical interval. Signal quality of the uncompressed video remains critical in systems that use compression for several reasons listed below. This leads to a strong requirement for testing the entire video chain including the analog and full bandwidth digital portions as well as the sophisticated compression and transmission systems.

- The input to a video compression codec must be accurate, in compliance with appropriate standards, and of as high a quality as possible to provide for efficient encoding.
- Video processing such as adding titles and special effects can not be accomplished in the compressed domain.
- Production facilities will not be fully compressed due to the cost and quality of compression codecs.
- The only way for different compressed formats to be interchanged is in uncompressed form.

With the advent of compressed digital video systems the situation has become more complex. Signal quality testing will not work for the compression encoder/decoder part of the system. Traditional test signals are relatively simple compared to a natural scene and are easily compressed with little distortion or loss. Due to the ease of compression, these signals do not evaluate the encoder/decoder process. As an example, signal-to-noise ratio is not a reliable measure of picture quality, is not a constant for a given system and can give completely misleading results. Therefore picture quality measurements require a direct method, using natural scenes, or an equivalent thereof, which are much more complex than traditional test signals. These complex scenes stress the capabilities of the encoder resulting in non-linear distortions that are a function of the picture content.

Use of digital compression has expanded the types of distortions that can occur in the modern television system. Due to this increased complexity and the desire to optimize program distribution both technically and economically, the field of subjective measurement has expanded. Since signal quality measurements will not do the job, objective picture quality measurements are needed leading to the measurement diagram shown in Figure 2. The total picture quality measurement space has increased due to subjective measurements which now include multi-minute test scenes with varying program material and variable picture quality. Some of the subjective measurements even include an element of program quality. Objective picture quality measurements are unlikely to cover that entire area. The new objective measurement methods must also have strong correlation with subjective measurements and cover a broad range of applications. They will be able to detect much of the degradation due to the compression process. They should also be able to detect most of the picture defects now measured with signal quality methods albeit with less ease of measurement or resolving power. It is expected that picture quality distortions too small for the human to see will be measured and provide an indication of the performance of concatenated systems.

Expanded types of signal quality measurements are not appropriate to cover the new subjective methods. In fact, with the increased ideas for subjective evaluation it may be true that the traditional signal quality measurements no longer have such a strong correlation with subjective requirements. There does not appear to be any plan to expand or re-qualify the signal quality measurement methods since there is so much work to do in developing objective picture quality methods.

Another aspect of objective picture quality measurements is the immediacy of the results. As will be discussed later, a great deal of computation is required for such measurements. While real-time or continuous measurement results would be preferred, sampled measurements will be more practical in the near future. As an example, two seconds of video measured out of each minute of program material.

Subjective Testing

Television programs are produced for the enjoyment or education of human viewers so it is their opinion of the video quality which is important. Informal and formal subjective measurements have always been, and will continue to be used to evaluate system performance from the design lab to the operational environment. Even with all the excellent objective testing methods available today for analog and full bandwidth digital video, it is important to have human observation of the pictures. There are impairments which are not easily measured yet are obvious to a human observer. This situation certainly has not changed with the addition of modern digital compression. Therefore, casual or informal subjective testing by a reasonably expert viewer remains an important part of system evaluation or monitoring.

Formal subjective testing, as defined by Rec 500, (1) has been used for many years with a relatively stable set of standard methods until the advent of digital compression. In brief, a number of non-expert observers are selected, tested for their visual capabilities, shown a series of test scenes for about 10 to 30 minutes in a controlled environment and asked to score the quality of the scenes in one of a variety of manners. Subjective testing is used for both quality assessment, system performance under optimum conditions, and impairment assessment under non-optimum performance due to transmission limitations. In a modern television system that incorporates compression, the picture quality is not a constant over time. Picture quality is a function of the complexity of the program material and, in the case of statistical multiplexing, the moment to moment operation of the transmission system. Considering this time varying property and the number of new impairments, the defined and proposed measurement methods have grown in recent years. In addition to selection of the measurement method there are a number of other procedural elements for which alternate approaches are available. These are such things as; viewing conditions, choice of observers, scaling method to score the opinions, reference conditions, signal sources for the test scenes, timing of the presentation of the various test scenes, selection of a range of test scenes, and analysis of the resulting scores. Selection of the parameters for each of these elements is related to the intended application of the television system and leads to a complex maze of possibilities. In the most recent version of Rec 500 there are five major subjective testing methods defined with another two being proposed in the annex. Some of the methods include two or more scoring procedures. A description of the various subjective measurement methods provides some insight.

Double Stimulus Impairment Scale (DSIS) — Observers are shown multiple reference-scene, degraded-scene pairs. The reference scene is always first. Scoring is on an overall impression scale of impairment: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying. This scale is commonly known as the 5-point scale with 5 being imperceptible and 1 being very annoying.

Double Stimulus Continuous Quality Scale (DSCQS) — Observers are shown multiple scene pairs with the reference and degraded scenes randomly first. Scoring is on a continuous quality scale from excellent to bad where each scene of the pair is separately rated but in reference to the other scene in the pair. Analysis is based on the difference in rating for each pair rather than the absolute values.

Single Stimulus Methods — Multiple separate scenes are shown. There are two approaches: SS with no repetition of test scenes and SSMR where the test scenes are repeated multiple times. Three different scoring methods are used:

Adjectival — the 5-grade impairment scale, however half-grades may be allowed.

Numerical — an 11-grade numerical scale, useful if a reference is not available.

Non-categorical — a continuous scale with no numbers or a large range, e.g. 0 - 100

Stimulus Comparison Method — Usually accomplished with two well matched monitors but may be done with one. The differences between scene pairs are scored in one of two ways:

Adjectival — a 7-grade, +3 to -3 scale labeled: much better, better, slightly better, the same, slightly worse, worse, and much worse.

Non-categorical — a continuous scale with no numbers or a relation-number either in absolute terms or related to a standard pair.

Single Stimulus Continuous Quality Evaluation (SSCQE) — A program, as opposed to separate test scenes, is continuously evaluated over a long period, 10 to 20 minutes. Data is taken from a continuous scale every few seconds. Scoring is a distribution of the amount of time a particular score is given. This method relates well to the time variant qualities of today's compressed systems, however it tends to have a significant content of program quality in addition to the picture quality. In one evaluation, Rec601 video which has been considered to be essentially perfect for the past fifteen years, was given a quality rating above 90% for only 14 minutes out of a 20 minute program.

Advantages of subjective testing are; valid results are produced for both conventional and compressed television systems, a scalar mean opinion score (MOS) is obtained, and it works well over a wide range of still and motion picture applications. Weaknesses of subjective testing are; a wide variety of possible methods and test element parameters must be considered, meticulous setup and control are required, many observers must be selected and screened, and the complexity makes it very time consuming. The net result is subjective tests are only applicable for development purposes. They do not lend themselves to operational monitoring, production line testing or trouble shooting.

OBJECTIVE TESTING

The need for an objective testing method of picture quality is clear, subjective testing is too complex and provides too much variability in results. However, since it is the observer's opinion of picture quality that counts, any objective measurement system must have good correlation with subjective results for the same video system and test scenes. As with subjective testing, most objective testing methods do not claim to measure picture quality directly but provide an indication of how a degraded picture or scene compares with a reference picture or scene.

Over the past few years a wide variety of methods have been investigated for objective testing of picture quality in compressed video systems. The methods proposed may be roughly divided into two categories, feature extraction and picture differencing, each of which may be implemented in a variety of ways.

Feature extraction uses a mathematical computation to derive characteristics of a single picture (spatial features) or a sequence of pictures (temporal features). This usually results in an amount of data per picture (say, a few hundred bytes) that is considerably less than used to transmit the compressed picture. The calculated characteristics of the reference and degraded pictures are then compared to determine an objective quality score.

Picture differencing uses a matrix-based mathematical computation to process each picture or sequence of pictures. The resulting data represents a filtered version of the pictures containing an amount of data similar to the original pictures. Usually, the pixel-by-pixel difference between filtered versions of the reference and degraded pictures is used to determine an objective quality score. In some cases it may be the difference between the reference and degraded pictures that is filtered.

Applications of how the two basic methods might be used in an objective measurement system are shown in Figure 3a for feature extraction and Figure 3b for picture differencing. The advantage of the feature extraction method is the calculated characteristics of the reference (input) picture may be sent through the transmission channel along with the compressed picture for objective scoring at a remote location. Because

of this advantage the feature extraction method has been vigorously pursued, sometimes in combination with the picture differencing method. However, research at Tektronix and other laboratories has shown that certain picture differencing methods provide objective scores that correlate best with subjective results.

It is important to note, neither of these methods can be guaranteed to always give the correct polarity of the change in pictures although virtually all systems produce picture degradation. There are examples where a picture with noise or other artifacts is improved by filters at the input to a compression system resulting in a net picture improvement through the compression/decompression process.

Some of the concepts of the feature extraction method are codified, *for luminance only*, in a recently approved American National Standards Institute (ANSI) standard. (2) The standard defines methods for calculation of spatial information based on the statistics of spatial gradients in the vicinity of image pixels and temporal information based on temporal changes to the image pixels in the video scene from one frame to the next. Both scalar (total content) and vector (content as a function of angle) computations are used in the calculation of spatial information. Blockiness or tiling which may occur in a DCT-based system is detected by the change in spatial gradients as a function of angle. As an example: if a scene contains a fairly constant amount of spatial gradients for all angles, a blockiness impairment would increase the spatial gradient content in the horizontal and vertical directions. A comparison of this feature for the reference and degraded picture provides an objective measure of that type of impairment. Other feature extraction measurements in the ANSI standard are; maximum added edge energy, maximum lost edge energy, average edge energy difference, maximum lost motion energy and average motion energy difference. There are other methods defined in the standard including a picture differencing method, peak signal to noise ratio (PSNR). The ANSI standard may be considered a tool box of objective measurement methods providing a set of performance parameters where each parameter or combination of parameters is sensitive to some unique dimension of video quality or impairment type. Different compression systems have different artifacts. In fact, the same compression system may exhibit different artifacts depending on bit rates and encoder parameter settings. Therefore, the scope of the standard states “Discrimination between two or more similar systems is beyond the accuracy of the objective measurements defined in this standard at this time”. Further work by the members of the ANSI committee has been reported by Cermak and Wolf (3) indicating that a combination of feature extraction and picture differencing methods give the best results. Even with these extensions the methods (tools) to be used are expected to be chosen depending on the application to provide the best correlation between subjective and objective scores.

Another significant approach to feature extraction has been developed and reported in the latest proposed revision to the international subjective testing standard, Rec 500, as appendices “Picture-content failure characteristics” and “Composite failure characteristics of program and transmission conditions”. They introduce the concept of “criticality” which is a measure of the complexity of the pictures to be compressed. The idea is that pictures with more criticality (complexity) will be more difficult to compress and will result in lower picture quality. A measurement method for criticality is not defined however several methods are being investigated by a number of research laboratories, Ardito and Visca (4), Yuyama (5). One method uses some of the feature extraction calculations similar to those defined in the ANSI standard. Another method uses certain parameters derived in the compression encoding process either from a reference encoder or bit stream data from the actual system encoder. Using a reference encoder with no rate control, the output bit rate would be a measure of picture complexity. An example of using the bit stream data would be to monitor an MPEG encoder quantizer_scale parameter which is a variable throughout the picture but will have high peak and average values for more complex pictures. One extension of that approach is to divide the quantizer_scale parameter by the square of the spatial information content. This calculation can be made at the decoder with no direct data based on the reference scene.

Where objective measurement algorithms, such as criticality, require a knowledge of the compression method the results will be dependent on that compression method. Different methods will produce a different picture quality for the same calculation of criticality. Even the same method, for instance MPEG-2, will produce different results if parameters are changed such as group of picture length or relative number of bits allowed for luminance verses chrominance or perhaps simply a change in the quantization matrix.

The method of calculating criticality will be dependent on the specific system and its application much as the feature extraction methods are when applying the ANSI techniques.

As previously stated, certain picture differencing methods provide better objective picture quality measurement correlation with subjective results. The most obvious picture differencing method is to simply subtract the two pictures without any filtering or processing. If the difference is zero, the pictures are identical. When the pictures are different a mean square error (MSE) can be calculated on a pixel by pixel basis, a larger MSE indicates a greater difference between reference and degraded pictures. Another way to express this direct picture difference is peak signal to noise ration (PSNR) which computes the log of the ratio of the square of the peak signal (255_{hex} in an 8-bit system) to the MSE, much as is done for signal to noise ratio (SNR) in an analog system. This method has some practical uses and some significant failings. For a very constrained system, say bit rate change only on a single test sequence, MSE will increase with decreasing picture quality. Also, designers may find it useful to view the pixel value differences in picture form when looking for design problems. However, it is well known that MSE can give a completely false indication. An example would be comparison of two degradations; addition of a small amount of random noise, say five quantizing levels, or addition of somewhat less blockiness, say two quantizing levels. The latter impairment will have a smaller MSE value, however observers will consider the noisy picture to have little degradation where the blockiness will be quite apparent as a significant degradation. An example of this measurement is shown in Figure 4. (Note: due to various software applications and printing processes used, these pictures do not always show the subjective differences as described.) Both pictures are degraded and the reference is not shown. Codec A provided an output image with a MSE value of 21.26 but a significant amount of blockiness whereas codec B provides a much better looking picture with a small amount of added noise, however the MSE is worse with a value of 27.10. Therefore, MSE is not an appropriate picture differencing method for objective picture quality measurements.

Although use of feature extraction parameter calculations as the processing for picture differencing methods improves on the basic ANSI approach, the result is not application or technology independent. A recent ITU-T SG12 contribution by Beerends (6) emphasized the need for measurement techniques that do not require knowledge of the compression process. Numerous researchers have indicated that the way to achieve technology independence and provide good correlation between subjective and objective measurements is to have the test instrument perceive and measure video impairments in the same manner as a human observer. In other words, filtering for the picture differencing method should use a model of the human visual system (HVS). Application of such a model will provide an image quality metric that is independent of video material, types of impairments, and the compression system used.

Application of the Human Visual System

Researchers at the David Sarnoff Research Center (Sarnoff Corporation) have devoted significant resources over a number of years to studying the human visual system and applying it to television display and picture quality evaluation. Based on this work the JNDmetrix™ (JND = just noticeable difference) methodology has been developed for automatically and accurately assessing the perceptual magnitude of differences between a test and reference sequence.

Figure 5 shows an overview of the JNDmetrix architecture. The inputs are two sequences of arbitrary length which are separately processed (filtered) to the “difference metric” box near the bottom of the diagram where the differences between the processed sequences are used to develop the JND maps and JND numeric values. An example is shown in Figure 6. Image A is the reference and image B is the degraded picture. Image C is the JND map. Note the distortion of the numbers on the trolley car and the corresponding bright area in the JND map. Also note the solid line on the ground to the left of the trolley car which has become a dotted line in the degraded picture. In the JND map a series of dots shows the noticeable difference between the two pictures.

For the JND image quality metric calculation each field of the sequence is represented as a trio of RGB images. In the first stage labeled Front End Processing the voltage units are transformed to light output units to obtain luminance (Y), and then to the psychophysically defined quantities of the CIE $L^*u^*v^*$

uniform color space to obtain the two channels (u^* , v^*) of the model's chrominance pathway. In the next stage of the model, labeled Pyramid Decomposition, each sequence is filtered and down-sampled using a Gaussian pyramid operation to efficiently generate a range of spatial resolutions for subsequent filtering operations. Next, the Normalization stage sets the overall gain with a time-dependent average luminance, to model the visual system's relative insensitivity to overall light level and to represent such effects as the loss of visual sensitivity after a transition from a bright to a dark scene.

After normalization, three separate contrast measures are calculated; oriented, flicker and chromatic. In each case the contrast is a local difference of pixel values divided by a local sum, approximately scaled as a function of pyramid level so the result is 1 when the image contrast is at the human threshold. This establishes the definition of 1 JND which is passed to subsequent stages of the model. (The JND unit of measure is functionally defined such that 1 JND corresponds to a 75% probability than an observer viewing two images multiple times would be able to see the difference.)

In the Contrast Energy Masking stage, each contrast image is subjected to a point non-linearity, the gain of which is controlled by the response across other resolutions and channels. This gain-setting is included to model visual masking effects such as the decrease in sensitivity to distortions in busy image regions. The parameters of the point non-linearity at this stage are fit according to contrast discrimination data in which the contrast increment needed to detect the change in contrast is measured as a function of the contrast from which the change is made.

At the Difference Metric stage, outputs from the test and reference sequences are combined via a simple difference operator and then summed across pyramid levels and channels to return the number of JNDs in both luma and chroma. Separate JND maps for luma and chroma can be pooled into one map and summary statistics can be obtained. Such statistics would be mean JND, max JND and Q-norm which allows a generalized approach to mean and max calculations.

The JND image quality metric provides all the facilities required for a robust objective picture quality measurement method. It includes the three necessary dimensions for evaluation of dynamic and complex motion sequences; spatial analysis, temporal analysis and full color analysis. By using a model of the human visual system in a picture differencing process results will be independent of types of the compression process and resulting artifacts. This is particularly important in concatenated television systems which are expected to involve several different compression methods as discussed by Dalton (7). Objective measurement methods that rely on a model of the compression codec or evaluate specific types of artifacts will have very limited application in concatenated systems. In addition to being appropriate for overall system measurement, it is expected that combining the results of the JNDmetrix for separate parts of a concatenated system will provide a useful indication of overall performance.

SYSTEM APPROACH TO OBJECTIVE TESTING

An overall block diagram for application of the JNDmetrix is shown in Figure 7. A reference sequence is supplied to the system under test from a source such as a video recorder or other picture generating equipment. Objective measurements of picture quality including temporal aspects of the human visual system are possible with about two seconds of video sequence. The test sequence source should provide about five seconds of continuous video so the measurement can be made after the compression system has adapted to the type of test material. (Scenes with cuts could be one of the tests.) In addition, the longer reference will allow informal subjective assessment by an expert viewer with the scene repeated or palindromed if required for viewing. (For proper subjective measurements, ten to twenty seconds of continuous reference scene is preferred.) At the system output the degraded image is captured in the picture quality measurement instrument which also has a copy of the reference sequence. Reference and degraded picture filtering, data differencing and data pooling is accomplished with extensive compute power and the results made available by an appropriate human interface or computer data connection.

In order to make objective picture quality measurements it is necessary to insure that both the reference and degraded video sequences are presented in a similar manner to the image quality metric calculation process. This is much the same as required for subjective tests where room lighting, viewing distance and monitor

set up must be accurately duplicated. Therefore, three aspects of the measurement process that must be considered in system design are; format conversions, signal changes due to processing in the non-compressed part of the system, and picture alignment.

Virtually all modern compression methods operate on a component version of the video, however the equipment being used may provide only composite (PAL or NTSC) inputs or outputs. Therefore format conversion may be required as part of the presentation process. Since composite encoding and decoding produces artifacts in the picture which are independent of the compression system (although they may well affect operation of the compression encoder) there are three format related requirements for the picture quality measurement instrument; composite encoded reference scenes, an excellent quality composite decoder and component reference scenes that include the composite artifacts. Experiments conducted at Tektronix indicate the picture quality testing where composite encode and decode processes are included will tend to mask measurements of compression systems with small amounts of degradation. An example would be MPEG-2 main profile at main level (mp@ml) with long groups of pictures and bit rates in the 12 Mb/s to 15 Mb/s range. This appears to be a reasonable result since those bit rates represent the highest quality of entertainment video, either excellent NTSC/PAL or good component video. Systems that don't incorporate composite signals and provide a Rec 601 input/output can be evaluated for very small picture degradations (suitable for studio program production contribution quality) based on the JNDmetrix.

Linear signal changes, such as gain or dc level, may occur in the non-compressed part of the system. This is most likely where part of the system to be tested includes analog (generally composite) processing. In order to make the picture quality measurement (either objective or subjective) these variations must be removed. Therefore, one function of the objective picture quality measurement system will be to perform an appropriate set of traditional signal quality measurements on the degraded test scene and provide an appropriate output to the user. This will allow the user to adjust the equipment under test. Alternately, automatic adjustment (with appropriate warnings) can be made by the measurement instrument.

Most important for the JNDmetrix is the requirement for temporal alignment and very accurate spatial alignment. These two requirements are due to the differencing process between video frames. As an example consider the MSE method. If there is no degradation, and the pictures are accurately aligned, the difference will be zero. However, if they are off by one pixel, the difference will be large even though there is no degradation. Spatial alignment to one-twentieth of a pixel is required for a measurement resolution appropriate for the overall accuracy of the JNDmetrix. It is important to note there are some degradations that may occur in analog systems, such as picture stretching with a non-timebase corrected VTR, which make the measurement impossible. This type of degradation must be detected by the measurement instrument and reported to the user.

TEST SCENES

Input to the system under test is a number of short reference sequences used in a direct measurement technique. That is, actual pictures are used rather than test signals. Multiple test stimulus is also the approach for analog or full bandwidth digital systems which use a number of test signals for indirect measurements. For picture quality measurements the different reference sequences will represent various applications for the system and types of program material. Some examples are; sports following the action with background moving, sports stationary camera with the action moving, scenes with high detail, panning and zooming on high detail scenes, rotary motion with colors not easily handled by some compression systems, subtle skin tones and lighting, and scenes with variable amounts of noise content. One requirement is the test material be such that the system being measured is working at or near the limits of its capabilities. This has always been done with traditional analog measurements (an example would be use of the 2-T pulse) and is even more important to stress the non-linear characteristics of video compression systems. Although scenes that break either the compression system or measurement method will be of some interest to find the outer limits of the system, they are not appropriate for repeatable and consistent measurements.

Studies which compare subjective and objective picture quality measurements generally conclude there is a moderately wide variation in subjective results. This conclusion is often emphasized by one or more scenes whose subjective quality does not provide good correlation with objective measurements. Certainly it would be desirable to develop an objective method with no algorithm-breaking scenes, however standardization of well behaved and truly representative scenes should provide very useful results. Considering that some program material does not correlate with signal quality test results in today's analog systems (striped shirts near the subcarrier frequency) and that objective tests for compressed video systems are predicted to be only 90% to 95% accurate, it would seem appropriate for the industry to agree on a set of standardized motion sequences for objective measurement of picture quality. This will allow development of very useful, if not perfect, picture quality measurement equipment.

There are a variety of test scenes being used by research laboratories worldwide. Everyone has their favorites and there are many types of program material being delivered by compressed systems. In order to have consistent, repeatable and predictable results for a variety of transmission and measurement systems it will be necessary for standards bodies to select a limited number of scenes. In addition to being appropriate the scenes must be available for all to use for any purpose including commercial applications. This requirement places a roadblock for many scenes due to copyright considerations. Therefore, it is suggested a subset of the CCIR test scenes referenced in ITU-R BT.802-1 be used since they are stated to be in the public domain. If additional scenes are needed, public domain use of appropriate copyrighted scenes should be obtained or new scenes should be produced in a copyright-free manner. Once defined, digital tapes of the scenes could be produced and certified by an appropriate standards organization such as the SMPTE.

Use of specific reference scenes means that testing will be out-of-service. This paradigm for video testing will not be popular with those who have, for many years, used vertical interval test signals (VITS). Although in-service testing with the actual program material would be logistically possible in some applications (monitoring a direct broadcast satellite system at the up-link location), it might not provide meaningful results for a majority of the program material which does not stress the system. Once a measurement method is standardized and fully validated, its application directly to program material may prove to be useful.

There is another concern regarding the use of specific reference scenes. It may be possible for a compression encoder to be tuned for best results on those scenes while performing poorly on some program material. On the other hand, selection of the correct set of test scenes may exercise all key aspects of the encoder such that good measurement specifications are reflected in its performance on general program material. This will be an on-going discussion until a reasonable body of experimental results are obtained.

A PICTURE QUALITY MEASUREMENT INSTRUMENT

Tektronix and Sarnoff Corporation are cooperating on the development of a picture quality measurement product based on the JNDmetrix and the signal processing required for system considerations described above. A functional system block diagram is shown in Figure 8. The instrument consists of a server class computer with two plug-in modules for measurements and two optional modules for reference scene generation. Two of the modules are identical (labeled Compute Engine) although they are shown differently for functional purposes. Each compute engine has 128 MB of SDRAM and two Texas Instrument C80 digital signal processors (DSPs) supplying the computational power.

Measurements are made by capturing the appropriate two seconds of the five-second test scene. Input formats are Rec 601 (component 270 Mb/s), S-VHS (analog component) and PAL/NTSC composite. The latter two are translated into Rec 601 by the decoder module. Reference scene data is supplied from the computer hard disk which will have been previously loaded from the CD-ROM or ethernet port. Dotted-line boxes shown in the compute engine are not hardware but represent software DSP processes. First, signal measurements are made on the degraded scene and reported to the user. Automatic adjustments are made to the degraded scene for appropriate presentation to the JNDmetrix algorithm. Most important of

these adjustments are the spatial and temporal alignment. Signal quality type parameters (e.g. gain) may be manually adjusted in the system under test. The main computational process is calculation of the JND map and value. Luminance values are always calculated, and chrominance values are an option due to the increased time required to make the calculations. There are two copies of each 2-second reference scene, one with composite artifacts included and one without. The latter is used in conjunction with system configurations which do not include composite encoding.

The JNDmetrix provides a scale of values, large signifying large observable differences and small for less observable differences. At the low end of the scale picture, degradation may not be observable, however the JND values are useful for evaluating concatenated systems. Prior to pooling the data to form a single JND value, a map of each picture is available to show which areas have the greater noticeable differences. These maps as well as the reference and degraded scenes may be displayed on the external Rec 601 video display. Although MSE or PSNR may not be a reliable measures of picture quality, as described above, error maps are available as a trouble shooting tool for equipment designers.

One source of test scenes is the optional generator which consists of a compute engine module and an encoder to provide an analog PAL/NTSC output. Scenes may be output at the same time the degraded scene is being captured thus providing full test-set capabilities. Test scenes for loading into the instrument are available on CD-ROM. There are test scenes for both 525-line and 625-line systems. Each test scene is 5 seconds in duration and contains one of two different calibration stripe methods.

Calibration stripes are used in order to reduce the amount of time required to make a picture quality measurement. The limiting factor on measurement time is compute power for the signal measurement process in addition to the picture adjustment process and JND calculation. Use of the calibration stripes significantly reduces the time required for the signal measurement process, particularly the accurate resolution of any picture shift. The target specification is to make a luminance-only measurement in less than 60 seconds. For systems with no frame dropping, stable gain and stable picture position, a 1-second header stripe is included only in the first second of the test scene with the picture quality measurement made in the third and fourth seconds. Where a system is believed to be unstable in these aspects the striping is included in the portion where the picture quality measurement is made.

Figure 9 shows the function of the calibration stripes. (Drawing not to scale.) The stripe serves several purposes; luminance gain/alignment calibration, Cr/Cb gain/alignment calibration, descriptive data, and two of the four cropping calibration blocks. When used as header (first second of video) the location is approximately as shown and the cropping calibration blocks at the top and bottom are included. Descriptive data is the sync block and scene number. When used for continuous calibration (third and fourth seconds) the stripe is at the top of the picture, the two separate cropping calibration blocks are not included, the data is inserted between the chrominance alignment blocks, and the data contains only field count.

It is expected that future advancements in compute power and measurement algorithm optimization will allow in-service testing for applications where the reference (input) and degraded (output) video is available at the measurement instrument. This is important for statistically multiplexed encoding systems where bit rates are shared between programs with the potential that any part of program could be stressful to the encoding process due to the bit rate allowed.

Measurement is accomplished by user interaction with a computer display in much the same manner as for a dedicated instrument. The input format (analog composite, analog S-VHS, or Rec 601 digital component) is selected. Degraded scene capture can be a specific test sequence or simply the first recognized sequence. While the measurement calculations are being made the JND maps are displayed. When results are complete, average JND values for the 2-second sequence are displayed. Alternate results displays are a field-by-field list of values or a graph of the values. All maps and scenes may be displayed on the computer monitor, however the Rec 601 output supplied to a studio quality picture monitor will provide better subjective viewing.

CONCLUSION

There is a continuing need for traditional test methods for the program production process and to insure the best quality video is delivered to the compression process. Although formal and informal subjective picture quality assessment has been used to develop and test today's compressed television systems, it is too complex, time consuming and expensive for most design, manufacturing and operational applications. Therefore the need for objective measurement of picture quality (degradation with respect to a reference) is well established and immediate.

Researchers have developed a number of objective picture quality assessment techniques in two general categories, feature extraction and picture differencing. Application of human visual system model filters in a picture differencing mode provides results that are well correlated with subjective measurements. In addition, such a technique does not depend on knowledge of the compression method or types of degrading artifacts and is expected to work well for systems using a concatenation of compression processes. Although a wide variety of test scenes are being used for subjective evaluation it will be necessary to standardize a limited number in order to provide consistent, repeatable and predictable results for a variety of transmission and measurement systems.

The objective measurement methods discussed in this paper represent an exciting new paradigm for the television and telecommunications industries. Tektronix and Sarnoff Corporation are cooperating in development of an objective picture quality measurement instrument utilizing the human visual system based JNDmetrix. The instrument being developed provides the compute power required for JND calculations, a convenient control system, picture and calculation results displays, and the ancillary picture processing and measurement system necessary for real world applications.

REFERENCES

1. ITU-R BT.500-7, Methodology for the Subjective Assessment of the Quality of Television Pictures.
2. ANSI Standard T1.801.03-1996, Digital Transport of One-way Video Signals, Parameters for Objective Performance Assessment.
3. Cermak, G. and Wolf, S., 1996. Objective and Subjective Measures of MPEG Video Quality. ANSI committee document T1A1.5/96-121.
4. Ardito, M. and Visca, M., 1996. Correlation Between Objective and Subjective Measurements for Video Compressed Systems. SMPTE Journal, December 1996, pp. 768 to 773.
5. Yuyama, I. 1996. Objective Measurement of Picture Quality for the Digital Television Broadcasting. Proceedings of the Made to Measure '96 Symposium, Montreux, Switzerland, November, 1996.
6. Beerends, J. G., 1997. Objective Measurement of Picture Quality" ITU-T committee document Com7-12-E February 1997.
7. Dalton C., 1996. Why is Objective Evaluation Needed for Compressed Digital Video. Proceedings of the Made to Measure '96 Symposium, Montreux, Switzerland, November, 1996.

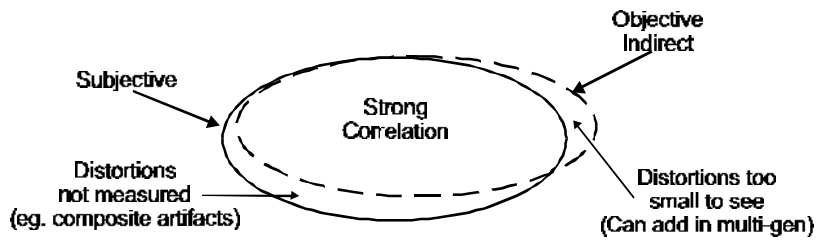


Figure 1. Traditional video measurements

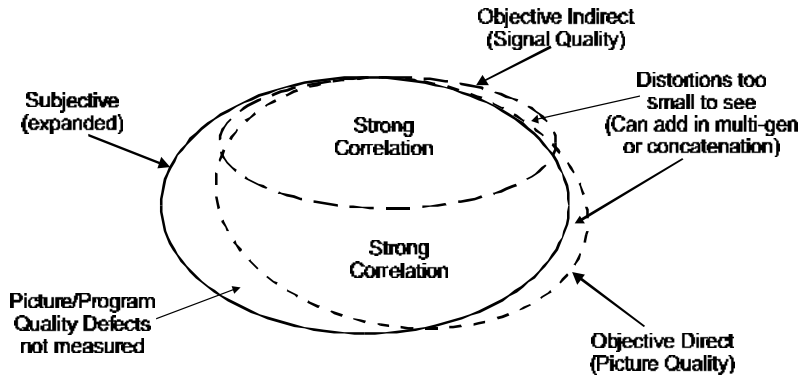


Figure 2. Modern video measurements

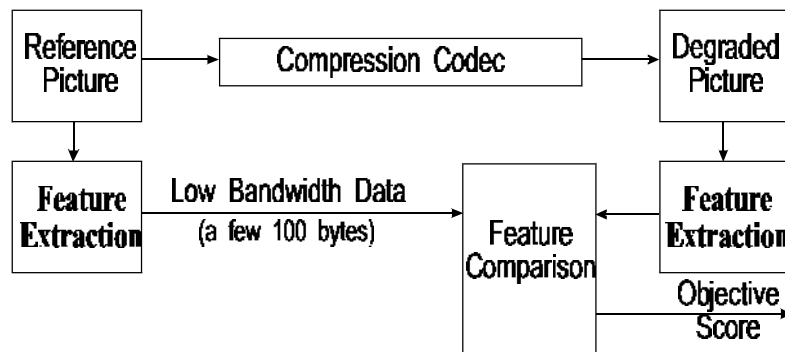


Figure 3a. Feature extraction method

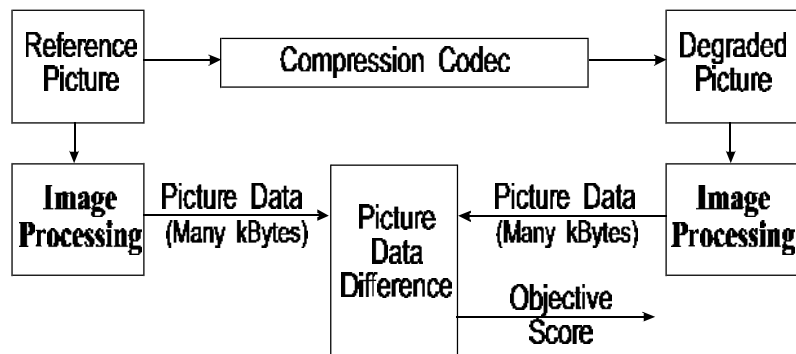


Figure 3b. Picture differencing method



Figure 4a. Codec A, MSE = 21.26



Figure 4b. Codec B, MSE = 27.10

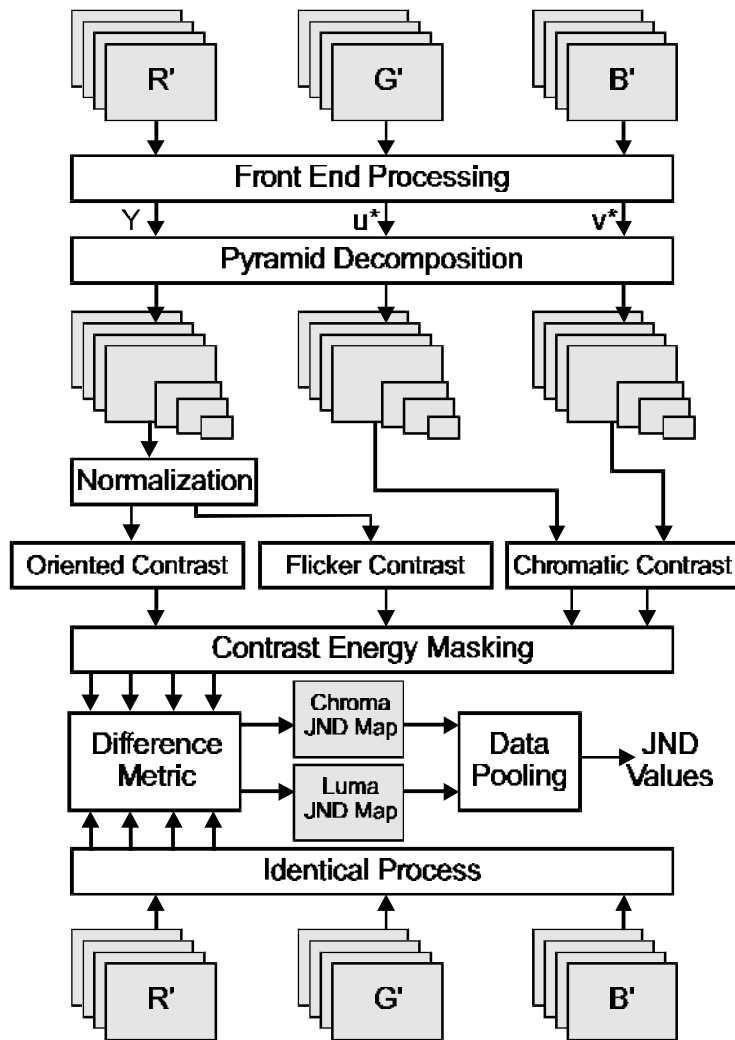


Figure 5. JNDmatrix measurement system



Figure 6a. Original picture



Figure 6b. Degraded picture

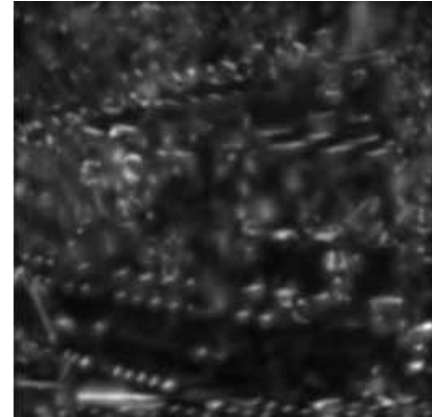


Figure 6c. JND Map

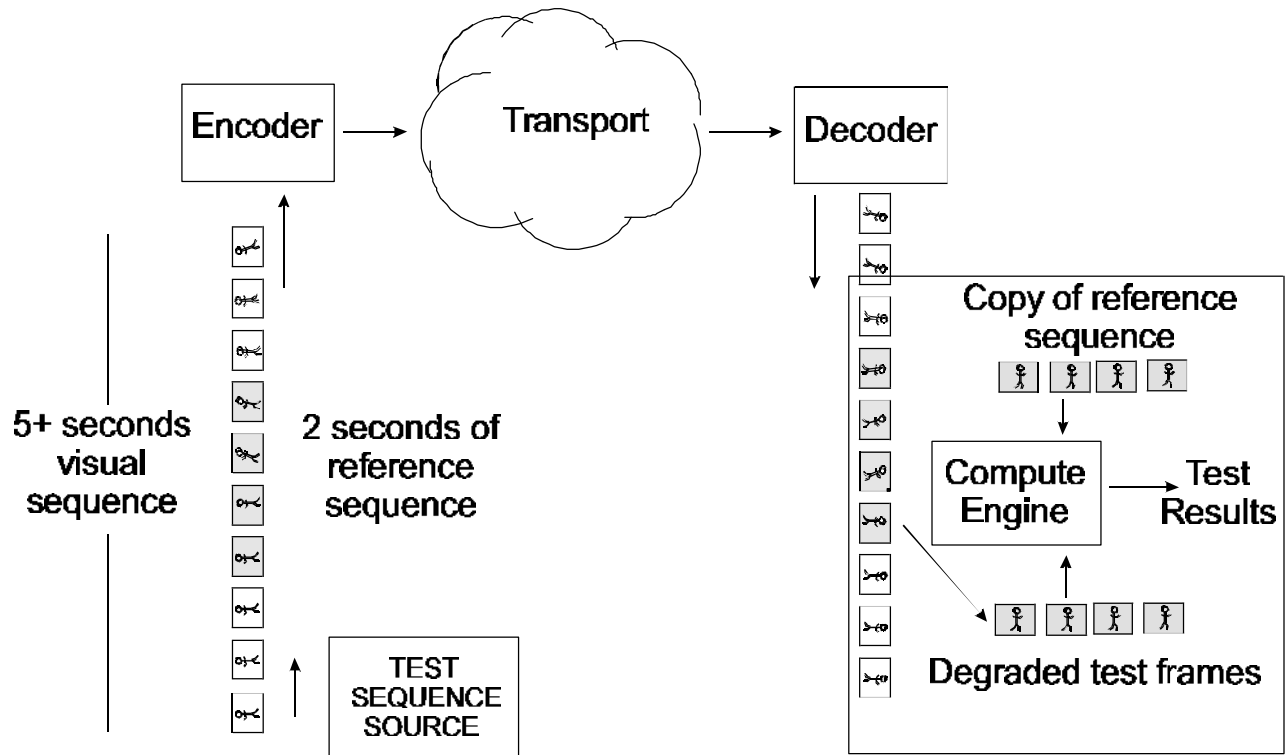


Figure 7. JNDmetrix measurement system

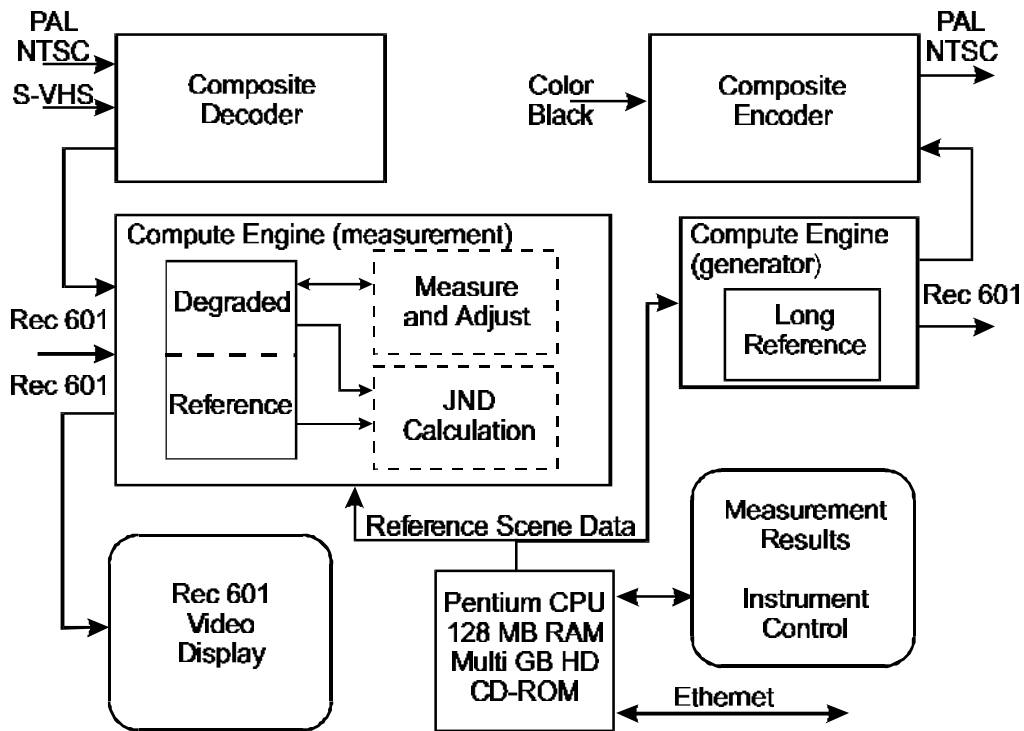


Figure 8. Picture quality measurement, functional diagram

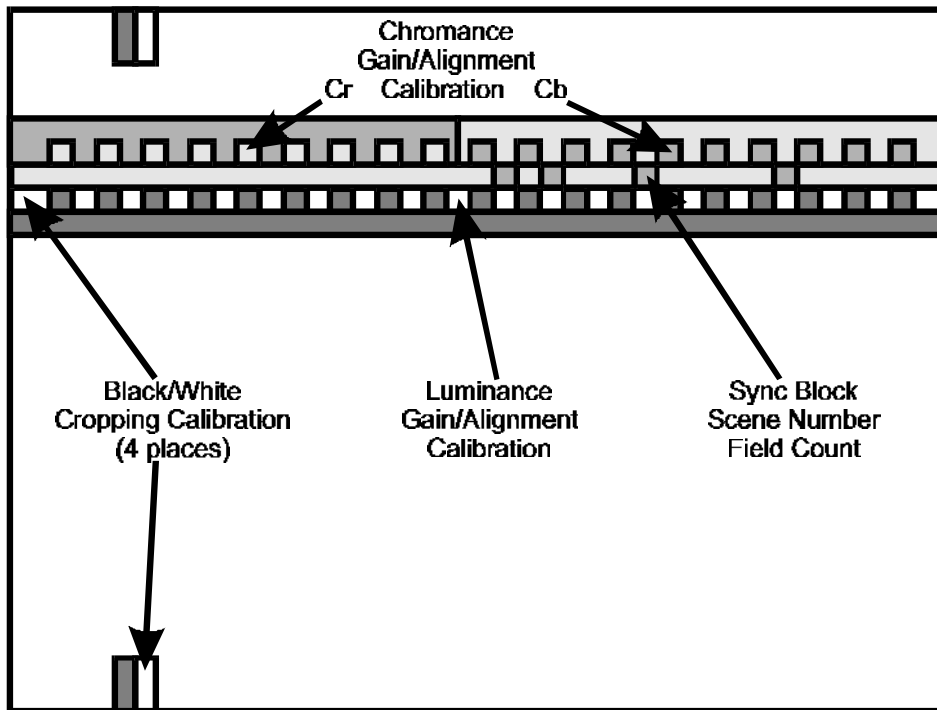


Figure 9. Calibration stripes (scene not shown)

