

A Guide to Picture Quality Measurements for Modern Television Systems



Introduction

Convergence is a popular term these days. It has many definitions and many factors depending on one's perspective and technology base. From a generalist's point of view, convergence may be defined as "the coming together of communication, computer and television technologies to provide information of any kind to any location." One of the major focal points of convergence is the need for a complete new technology for the evaluation of modern television systems. In this guide, the aspects of video testing are presented based on an understanding of the complete television system including production, compression, decompression and the display or reuse of the original program. The need for continuing application of traditional video testing methods is explained along with their limitations for the identifying the artifacts introduced by video compression. With the variety of video compression methods in use and

being developed, there is a requirement for picture quality assessment methods which are independent of the compression algorithm and its related artifacts. An overview of subjective testing, which uses a panel of observers, is presented as it has been the mainstay for video compression system development. Due to the complexity and variability of subjective testing there is a strong requirement for an objective measurement instrument much as we use today for traditional television systems. Advantages and limitations of proposed objective testing algorithms are presented leading to the conclusion that a method based on a representation of the human visual system is required for best results. To complete the guide, implementation of an objective picture quality assessment algorithm in a practical measurement instrument is shown to require a combination of traditional video technology and modern computer techniques.

Compressed Television Systems

Compression is Nothing New. There are two reasons to compress television video signals, practical limitations of processing speed (bandwidth) and cost of transmission or storage resulting from the required bandwidth. Today, the availability of high-speed semiconductors and integrated circuits make the latter reason most important in nearly all applications. Virtually all video compression methods utilize the limitations of the human visual system to remove the less visible picture information that might otherwise be present.

As broadcast television was being developed, display rates of 50 or 60 pictures per second were considered necessary. To provide sufficient visual information each picture was judged to need about 500 display elements (now called pixels) in each direction.¹ To generate and transmit such a sequence of pictures in analog form would require a processing speed and transmission bandwidth of about 10 MHz which was difficult for the available technology

and excessive for the available radio frequency spectrum.

The first practical television broadcast systems used a form of two-to-one bandwidth reduction, or compression, called interlace. Instead of sending 50 or 60 frames per second, each frame is divided into two fields containing half the total number of lines. The lines in the first field are every other line from the frame, say lines 1, 3, 5... and the lines in the second field fill in the missing lines during the second field as shown in Figure 1. Picture degradation due to interlace is in the form of an artifact known as inter-line twitter, however the quality is quite satisfactory for entertainment video viewed several picture heights away from the display device.

In the 1950s, color television was developed. A single color picture requires three images, specifically red, green, and blue (RGB) for light emitting tubes (CRT). Starting from the full progressive scan picture, this would require a 30 MHz bandwidth to provide the desired picture rate.

Again, interlace is used to reduce the bandwidth to 15 MHz for an analog RGB system. Within a studio the signals are carried on three separate cables at 5 MHz or more bandwidth each, as shown in Figure 2. A fundamental compression scheme used in color television is to translate the three color signals into the color-difference domain where the picture is represented by a luminance (equivalent to the earlier monochrome) picture and two color difference pictures, R-Y and B-Y. Another name for this system is YUV, Y for luminance and U, V for the two color difference signals. Again using the limitations of the human visual system, in this case less color than luminance visual acuity, the bandwidth of the color difference signals is reduced by 50% for a total YUV bandwidth requirement of 10 MHz. Today, YUV signals are used in both analog and digital forms and have very little visible degradation compared to interlaced RGB video. Both forms are known as component video with YUV being used for most applications.

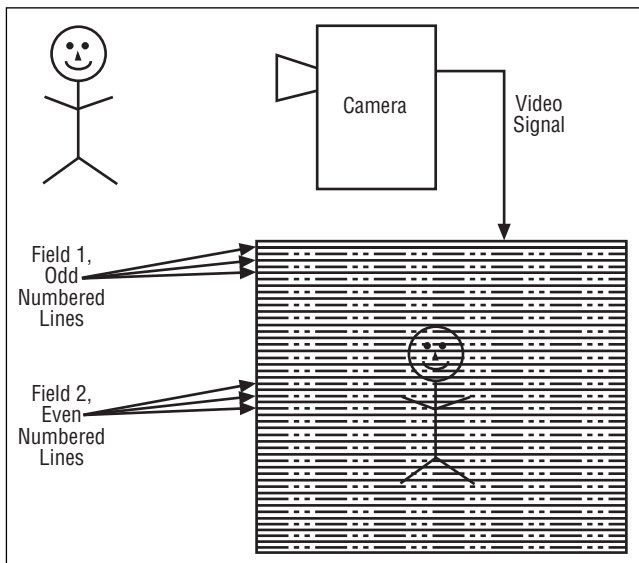


Figure 1. Interlaced scanning.

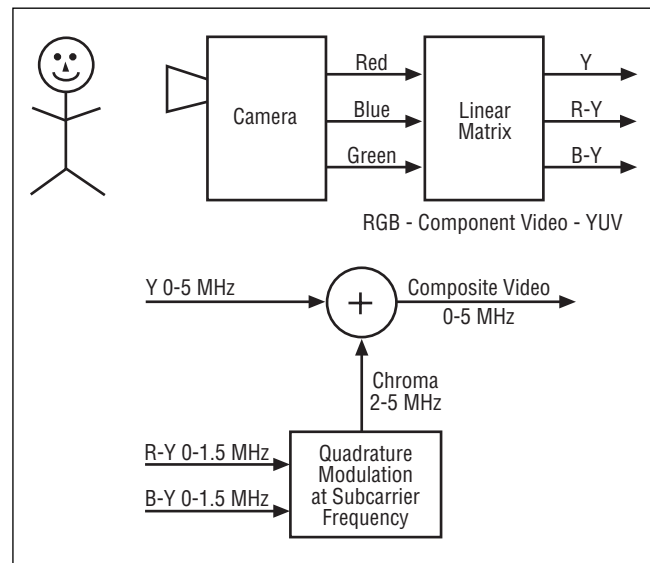


Figure 2. Analog television compression.

¹The Americas, Japan and Korea use the 525-line, 60 field/s system while most of the rest of the world uses the 625-line 50 field/s system.

Color television in the 1950s, and until recently, required further compression both to fit in the allocated 6 MHz bandwidth of transmission channels and to be compatible with the installed base of monochrome television sets. To accomplish this task the two color difference signals are further reduced in bandwidth to about 1.5 MHz each and quadrature modulated on a subcarrier which is subsequently added to the luminance signal producing composite video. Composite NTSC and PAL² produce very good entertainment quality color video in a 5 MHz bandwidth, a compression ratio of six to one from the ideal progressive scan RGB video. The final two to one compression from component to composite does bring with it noticeable picture degradation including chroma information seen as incorrect luminance and visa versa.³

Today, using modern digital compression methods, four or more excellent quality digital component television signals can be delivered to the home within the same 6 MHz transmission channel. If derived from a high quality digital component

source, the resulting multiple television signals have a noticeable quality improvement over the single 6 MHz bandwidth composite video signal.

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. 601⁴. 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. The compression method that is becoming dominant today is called MPEG-2, defined by the Motion Picture Experts Group and standardized by both the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC). MPEG-2 is based on the Discrete Cosine Transform (DCT) method in combination with powerful temporal compression techniques.⁵ Although some applications may be best served by other compression methods, MPEG-2 is expected to be the most widely used method in the foreseeable future. This is because it is an agreed standard that is either optimum or good enough for a wide variety of applications, a large amount of effort is going into the development of chip sets for lower cost encoders and decoders, and the forthcoming large installed base will be attractive for many equipment manufacturers and application developers.

²NTSC (National Television System Committee) is used in most 525-line countries and PAL (Phase Alternating Line) is used in many, but not all, 625-line countries.

³For a more complete description of basic digital television see "A Guide to Digital Television Systems and Measurements", Tektronix literature number 25W-7203.

⁴Rec. ITU-R BT.601, "Encoding Parameters of Digital Television for Studios." Originally it was CCIR Recommendation 601 but has been changed to Recommendation ITU-R BT.601. Rec 601 is used throughout this guide (and is much easier to say).

⁵See "MPEG-2 Fundamentals for Broadcast and Post-production Engineers" for a brief description of the MPEG compression method. Tektronix literature number 2AW-1061.

The Modern Television System.

A simplified block diagram of a modern compressed television processing and transmission system is shown in Figure 3. Television nominally consists of audio and video; however, the system may include data and control signals (not shown in the figure) hence may be thought of as a multimedia system. One-direction transmission is shown. A multiplicity of methods are depicted, particularly in transmission, making this diagram an overview of many types of applications. No specific compression method is shown, however the MPEG-2 transport stream is shown in the transmission area since it can be considered a general purpose multiplexing scheme capable of carrying any type of compressed video and audio.

Analog RGB video is produced in the camera and processed into one or more of several possible formats; analog composite, digital composite, analog component or digital component. Full-bandwidth digital

video is an extremely important part of the television system today. Program production processes must be full-bandwidth digital (or analog) in order to manipulate the picture to produce desired artistic results.

Following program production, the television signal may be compressed for storage, efficient transmission or intra-facility interconnection in digital form. Typically this will be MPEG-2 compression resulting in an MPEG transport stream (MTS) that may be multiplexed with other MPEG transport streams for transmission or interconnection.

New systems for RF transmission of television signals use digital modulation schemes which are generally more robust for the same transmitted power and provides the digital channel for compressed television signals. It is important to note, that even compressed digital video broadcasting to the home often starts with full bandwidth digital video to drive the bit-rate efficient, statistically-multiplexed compression system.

The broadband telecommunication system provides a variety of transmission methods. Traditionally these have been voice channel oriented with special data mapping for digital television signals. Although direct mapping of the MTS into the digital telecommunications hierarchy is in the process of being standardized, it is expected that asynchronous transfer mode (ATM) will become the preferred method of inter-facility video transfer.

Video testing in this modern television system is not just a matter of developing new techniques to evaluate the effects of compression. The significant portion of the system utilizing analog and full-bandwidth digital signals requires application of traditional analog and recently developed digital test methods. To determine picture quality impairments caused by compression, a video measurement system must take into account the various signal format changes affecting the video throughout the system.

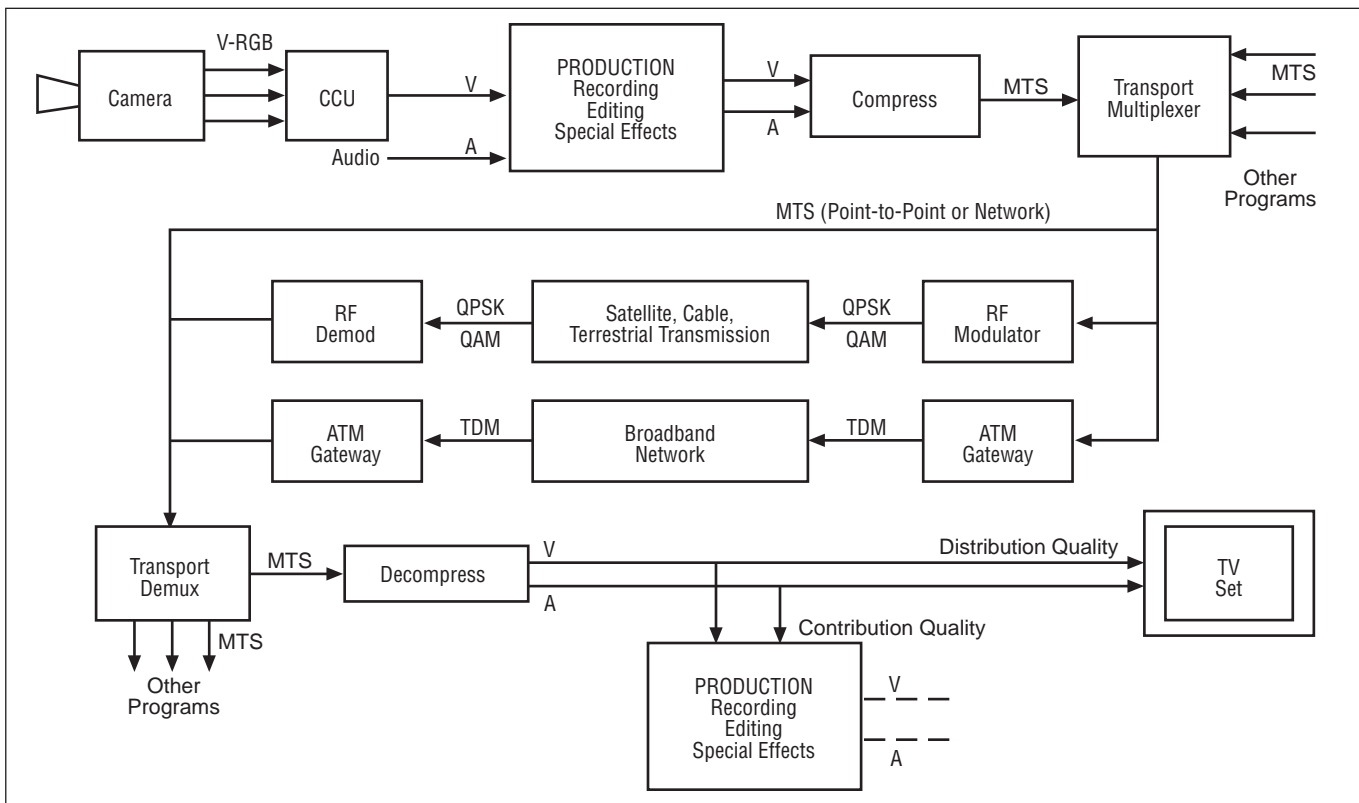


Figure 3. Modern television system.

Video Testing Concepts

Growth of Functional Layers. Over the half-century of widespread television use there has been a relatively simple model for analyzing analog video systems. Figure 4 shows a basic block diagram of the analog video system, its functional layers and test methods. Testing is performed at one interconnection generally carrying a composite PAL or NTSC signal. A single measurement instrument can analyze both the operational aspects, such as signal level or color balance, and the data formatting which is the synchronizing signal part of the same video waveform. This analysis of the signal quality through the transmission path using a suite of test signals does an adequate job of characterizing resulting picture quality. The idea of a suite of test signals is important. No one test signal will characterize the system and some expert interpretation as well as visual inspection of the resulting pictures is required. For intra-facility transmission of signals on coaxial cable, a separate piece of test equipment, the time domain reflectometer, is used to ensure the continuity of the physical layer. Long range transmission is by amplitude or frequency modulation on a carrier, however the resulting channel characteristics for the video are still determined by analog measurements such as those specified in ANSI T1.502.⁶

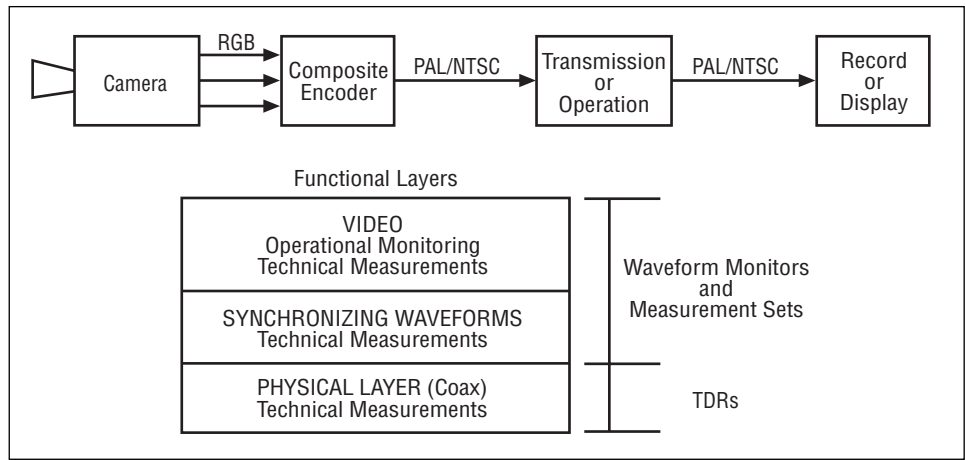


Figure 4. Analog video system.

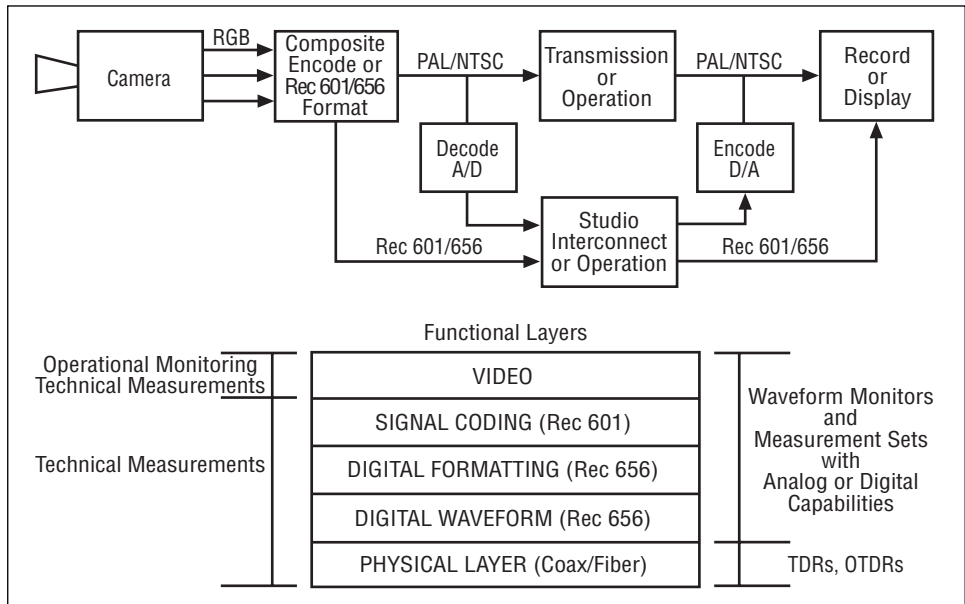


Figure 5. Hybrid digital/analog video system.

⁶ANSI Standard T1.502 "System M-NTSC Television Signals Network Interface Specifications and Performance Parameters."

With the advent of digital television over the past 15 years, a more complex system block diagram and set of functional layers has been required as shown in Figure 5. The analog signal is converted to digital in accordance with a sampling standard such as Rec. 601. Formatting and studio interconnection of the digitized signal follow a related standard, Rec. 656,⁷ leading to an extension in the functional layers and the variety of tests to be performed. For operational purposes, the monitoring of analog video signal properties is still key; however, this signal must be processed from the digital data.

Where testing of the analog signal required only that various parameters be measured on a single waveform, digital testing requires analysis of the digital waveform, digital data formatting and digital signal coding in addition to the resulting analog signal. Again a suite of test signals is required, expanding the suite needed for analog-only testing. Although all those measurements can be performed with a signal instrument, such as the Tektronix WFM601M, there is significant processing between each pair of layers with different analysis methods for each layer as well. Prior to the advent of digital compression

techniques, transmission of this higher quality signal was handled by compression back to the composite analog domain. The analog-to-digital and digital-to-analog conversion does introduce some signal quality degradation beyond that of the basic NTSC or PAL analog signal.

With the convergence of television and telecommunication, not only are there many more functional layers for the test engineers to consider but there are various possible paths with different layers. Figure 6 shows a few of the possible functional paths and layers.

Serial Digital Interconnect (SDI) is the Rec. 656 worldwide standard used for serial digital video. The SMPTE Working Group on Packetized Television Interconnections is developing a method of carrying packetized data over the same cabling and switching hardware called SDDI (for Serial Digital Data Interconnect). A networking type interconnect for the television facility, being considered by ANSI and SMPTE, is Fibre Channel which provides high speed, large packet sizes and reasonably priced hardware. The Synchronous Digital Hierarchy (SDH) telecom methods are well established worldwide and can directly carry the MPEG-2 transport stream with simple data formatting, although there is presently no standard. Looking toward the future, ATM is the expected method for transmission of packetized data, certainly for long distances and perhaps within a studio.

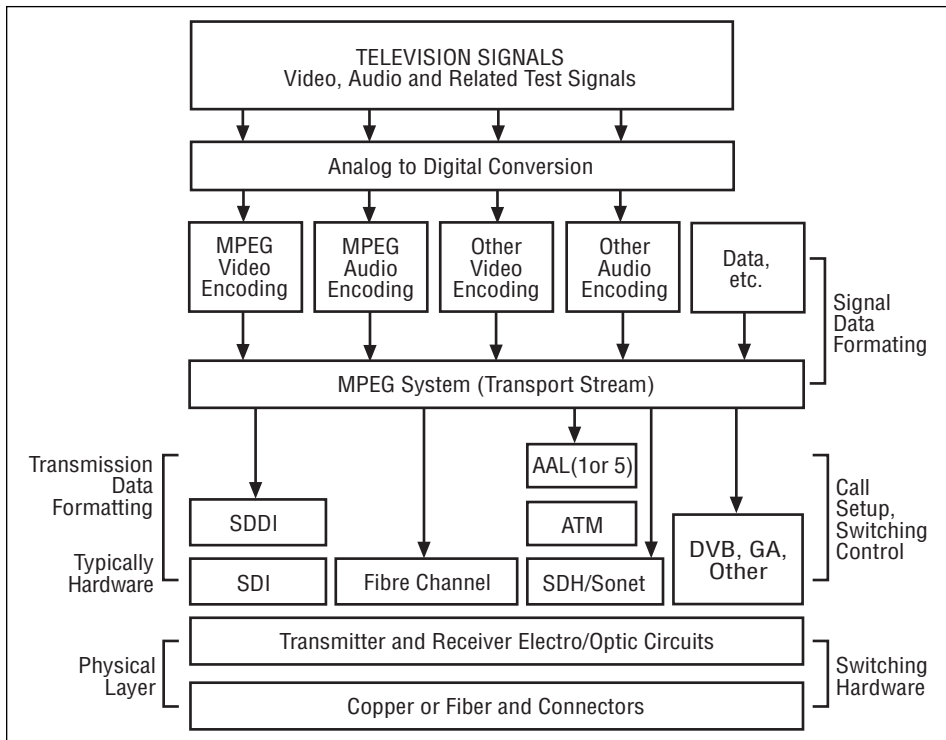


Figure 6. Modern television functional layers.

⁷Rec. ITU-R BT.656, "Interfaces for Digital Component Video Signals in 525-line and 625-line Television Systems Operating at the 4:2:2 level of Recommendation 601."

Three key testing layers can be defined for the modern television system as shown in Figure 7. Each has its own subset of more detailed testing layers. Video quality for compressed television systems is a much more complex matter than just using the indirect measurement methods for uncompressed video. This will be covered in detail in subsequent sections of this guide. Once the picture has been compressed, the resulting data is formatted for intra-facility connections. Examples for the use of such connections are: program interchange between video disk servers or several video/audio encoders sending single program transport streams to a multiplexer to produce a multi-program transport stream for satellite broadcasting. This is an appropriate layer for protocol testing because the data formatting can be quite complex and is relatively independent of the nature of the uncompressed signals or the eventual conversion to inter-facility transmission formats. For a majority of the television transmission systems the MPEG-2 transport stream is the common denominator at the compressed data level. The syntax and semantics for both the compressed data and the transport stream are well defined. Typical protocol testing equipment, such as the Tektronix MTS 100, will be both a source of known valid, or specifically invalid, signals and an analyzer which locates errors with respect to a defined standard and determines the value of various operational parameters for the stream of data. There are a number of possible inter-facility transmission methods as previously described. Many are well established, such as SDH/Sonet and cable television, with a variety of effective test equipment available. ATM is an emerging technology with new test equipment on the market

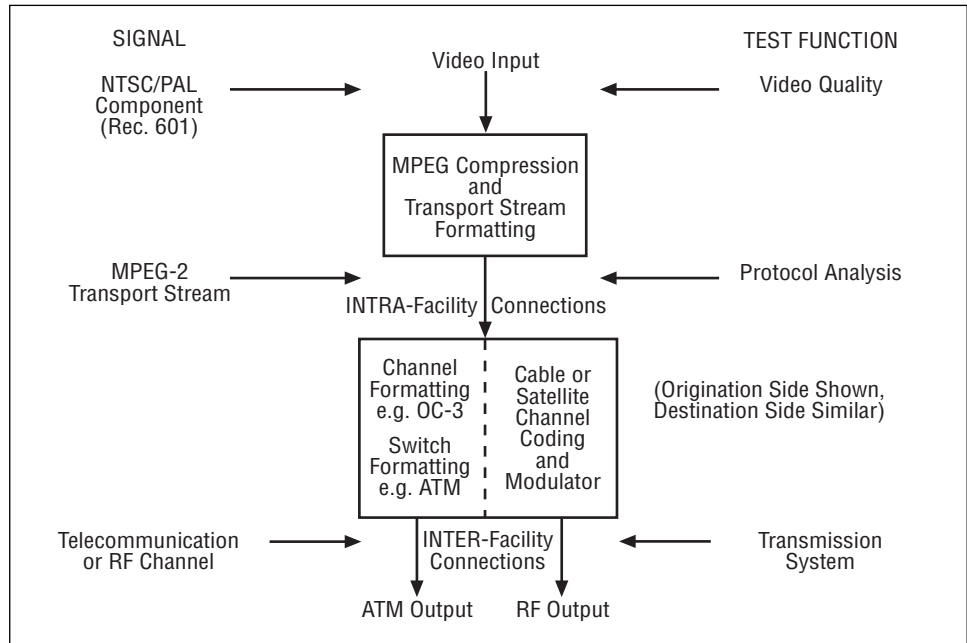


Figure 7. Functional testing layers.

and under development. Adaptation of traditional communication test equipment to analyze or interconnect with MPEG-2 transport streams is on the horizon.

Video Quality. There are several dimensions of video quality measurement methods that need definition. These are summarized in the table below. 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 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

Although there is a modest amount of compression applied to the NTSC and PAL composite systems, they 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 strong mathematical correlation between subjective measurements made on pictures from the system and objective measurements made on a suite of test signals using the same system. The correlation is not perfect for all tests. There are distortions in composite systems, such as false color signals caused by high frequency luminance, which 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 effect 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 uncompressed systems are linear.⁸ 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 very short, as an example, one line in the vertical interval. Signal quality of the uncompressed

video remains critical in systems that use compression for several reasons:

- 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 at the full bit-rate level.

This leads to a strong requirement for testing of the analog and full bandwidth digital portions as well as the sophisticated compression and transmission systems.

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, it is not a constant for a given system and it 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. Quantization noise which is also present in full-bandwidth digital systems is often increased by the compression system bit rate reduction process. Blockiness is a checkerboard pattern that may occur in DCT-type compression systems. Loss of resolution is common because the compression systems use the human visual system limits of acuity as a guide for removing information from the picture. Therefore, greater compression generally means less resolution. Although human acuity is less for chroma, the uncompressed picture has already used some of that latitude and compression systems often squeeze the chroma even more than the luminance. Edge busyness is another effect of quantization since more information is removed from the high-resolution parts of the picture producing noise on edges. When that noise is displaced by the compression processing into nearby flat areas it is sometimes called mosquito noise. Motion related artifacts, such as jerkiness or misplaced blocks of pixels, are present in systems which use temporal compression either based on sophisticated motion compensation or simply dropping frames because there are not enough bits available in low bandwidth systems.⁹

⁸Analog systems are not perfectly linear, however they are quite good and sensitive objective testing can be used to determine the small amounts of non-linearity.

⁹A list of impairment terms and other definitions may be found in ANSI T1.801.02 iDigital Transport of Video Teleconferencing, Video Telephony Signals - Performance Terms, Definitions and

With the broader range of distortions to measure and the desire to optimize program distribution both technically and economically, the field of subjective measurement has expanded. Some of the subjective measurements even include an element of program quality as well as picture quality as will be discussed in detail later in this guide. Since signal quality measurements will not do the job, objective picture quality measurements are needed. 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 as strong a correlation with subjective requirements. There does not appear to be any plan to expand or re-test the signal quality measurement methods since there is so much work to do in developing objective picture quality methods. Such picture quality measurement methods must, also, have strong correlation with subjective measurements and cover a reasonably broad range of subjective considerations. 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.

Picture Quality Testing

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 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 has been used for many years with a relatively stable set of standard methods until the advent of digital compression subjective testing described in Rec. 500.¹⁰ 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. 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.

¹⁰The standard for subjective measurements is ITU-R BT.500 "Methodology for the Subjective Assessment of the Quality of Television Picture". First issued in 1974 and formally known as CCIR Rec. 500, version 7 of this document covers all of the past and proposed methods for subjective testing.

- **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, Rec. 601 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.

In addition to these defined methods, there are two new approaches that start to bridge the gap between subjective and objective picture quality measurements. They are "picture-content failure characteristics" and "composite failure characteristics of program and transmission conditions." These will be discussed in the section on objective measurements.

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 observers' 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, nearly all 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. Such comparisons tend to eliminate the aspect of program quality from the measurements. 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.

Figure 8 shows how the two basic methods might be used in an objective measurement system. The advantage of the feature extraction method (8a) 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 (8b) 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.¹¹ The 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. 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 reported¹² indicating that a combination of feature extraction and picture differencing methods give the best results. Even with these extensions, the methods to be used should 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.

This approach is compression method dependent as well as application dependent. Different compression methods will produce a different picture quality for the same input 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 versus chrominance. The method of calculating criticality will be dependent on application much as the feature extraction methods are when applying the ANSI techniques.

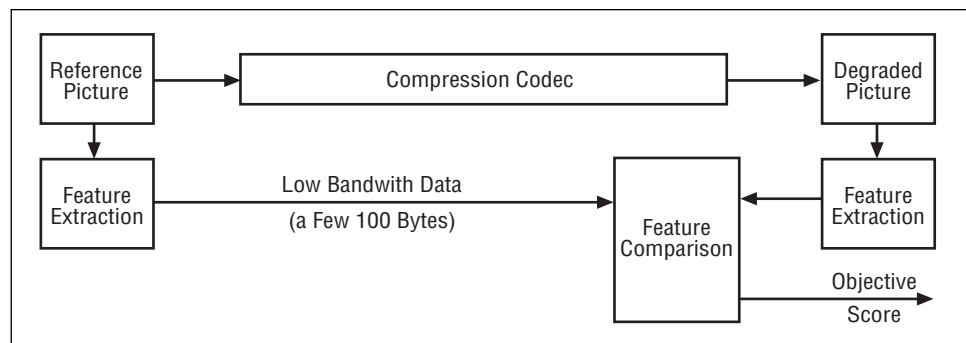


Figure 8a. Feature Extraction

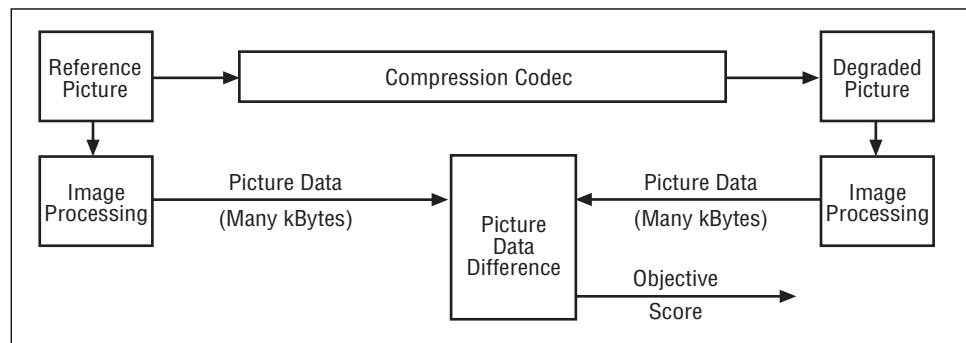


Figure 8b. Picture Differencing

¹¹ANSI Standard T1.801.03-1996 "Digital Transport of One-way Video Signals, Parameters for Objective Performance Assessment".

¹²ANSI T1A1.5/96-121 "Objective and Subjective Measures of MPEG Video Quality".

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 PSNR which computes the log of the ratio of the square of the peak signal (255hex 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, 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. As an example, consider the comparison of two types of degradation; one is the addition of a small amount of random noise, say five quantizing levels, and the second is the 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 9. 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 picture differencing methods based on feature extraction parameter calculations has been shown to improve on the basic ANSI approach, the result is not application or technology independent. 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.



Figure 9. MSE measurement examples.

Application of the Human Visual System

System. Researchers at the David Sarnoff Research Center (Sarnoff Labs) have devoted significant resources, over a number of years, to studying the human visual system and applying the knowledge gained to television display and picture quality evaluation. Based on this work, the Just Noticeable Difference (JND) image quality metric has been developed for automatically and accurately assessing the perceptual magnitude of differences between a test and reference sequence.¹³ Figure 10 shows an overview of the JND model 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 11. 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.

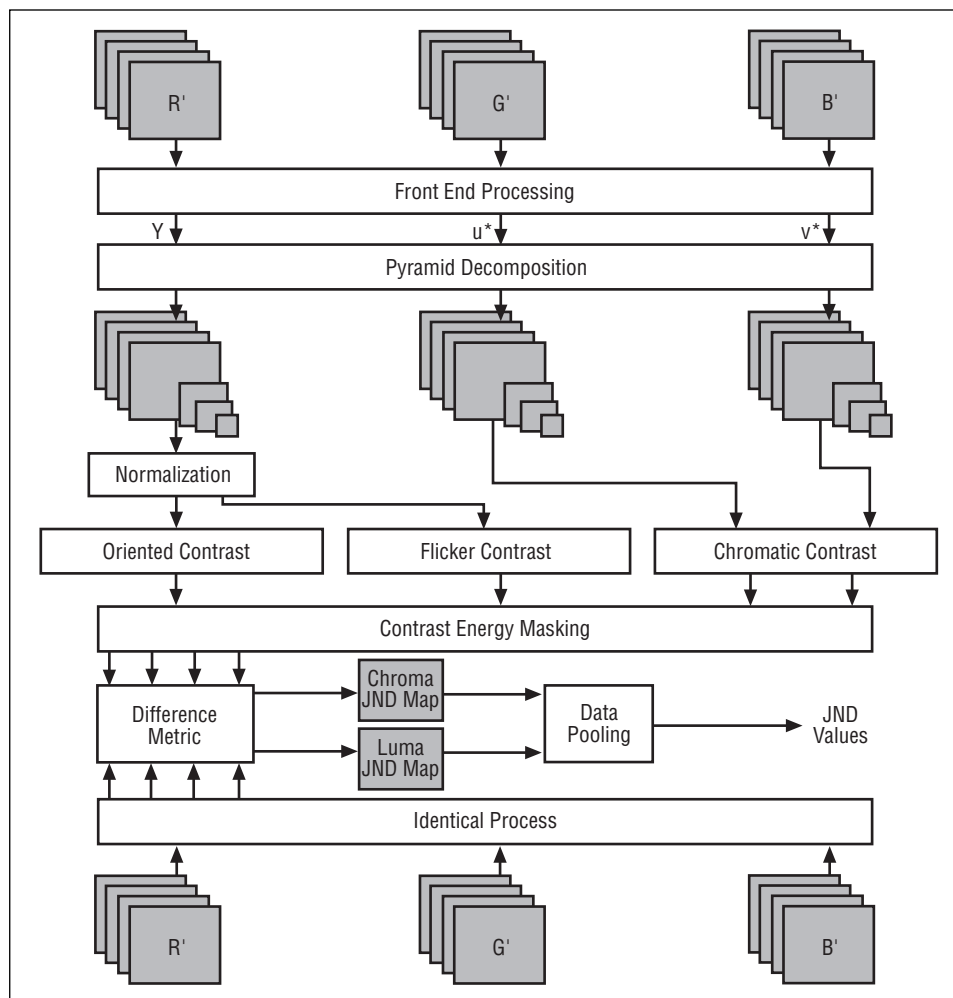


Figure 10. JND image quality metric architecture.



Figure 11. JND map example.

¹³Material for this section of the guide is excerpted from the paper “Vision Model-based Assessment of Distortion Magnitudes in Digital Video” by J. Lubin, M. Brill and R. Crane, presented at the Made to Measure '96 symposium, Montreux, Switzerland, November 1996.

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 the compression process and resulting artifacts. This is particularly important in concatenated television systems which are expected to involve several different compression methods.¹⁴ Objective measurement methods that rely on a model of the compression codec or evaluate specific types of artifacts will have very limited application in such systems. In addition to being appropriate for overall system measurement, it is expected that combining the results of the JND image quality metric for separate parts of a concatenated system will provide a useful indication of overall performance.

¹⁴An example of up to ten different compression methods in a complete television system is described in the paper "Why is Objective Evaluation Needed for Compressed Digital Video", by C. Dalton, presented at the Made to Measure '96 Symposium, Montreux, Switzerland, November, 1996.

System Approach to Objective Testing.

Objective testing requires a valid algorithm, such as the JND image quality metric, as its foundation. However, implementation of a real-world measurement system must include a number of other aspects such as: reference scene motion sequences, a physical source for the reference scenes, format conversions, scene changes due to processing in the non-compressed parts of the system, and accurate alignment of pictures as an input to the measurement algorithm. An overall block diagram of the measurement system is shown in Figure 12 for application of the JND image quality metric. A reference sequence is supplied to the system under test from a source such as a video recorder or other picture generating equipment (providing a defined video quality). Objective measurements of picture quality

including temporal aspects of the human visual system should be possible with about two seconds of video sequence. However, subjective assessment by an expert viewer may also be desired so the test sequence source should provide five or more seconds of continuous video which may be repeated or palindromed for longer viewing. 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, differencing and data pooling is accomplished with extensive compute power and the results made available by an appropriate human interface or computer data connection.

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 in either direct or 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

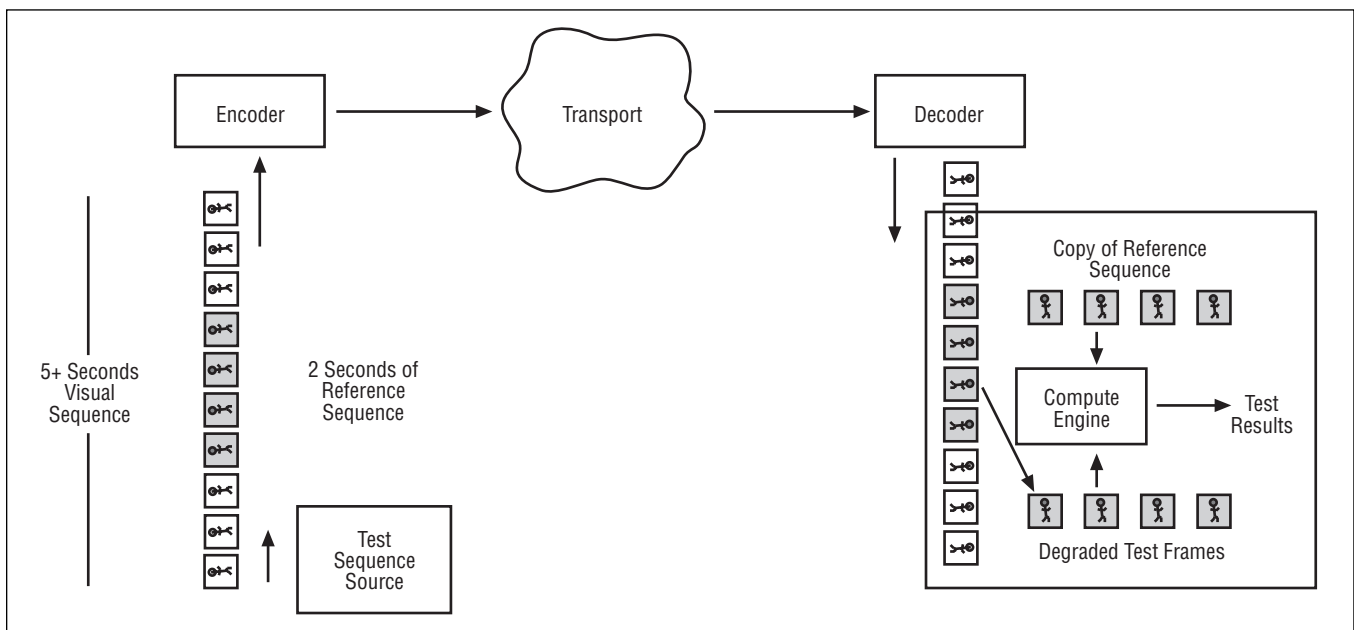


Figure 12. Objective measurement system.

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 variety of

standardized motion sequences for objective measurement of picture quality. This will allow development of very useful, if not perfect, picture quality measurement equipment. An ANSI standard defines a number of scenes for testing of video conferencing systems¹⁵ and there are a number of standards organizations working on definition of a wider variety of test scenes. It will be very important to have a set of standardized test scenes so measurement data will correlate between different manufacturer's test equipment and all systems designed for similar applications.

In order to make objective picture quality measurements, it is necessary to insure the two video sequences are presented to the image quality metric calculation in much the same manner as required for subjective tests. That is, gain and dc level of both the luminance and chrominance must be closely matched. In addition, there must be temporal alignment and very accurate spatial alignment. These latter two requirements are due to the need to do a type of a differencing process between video frames as done with PSNR and the JND image quality metric model. Spatial alignment to an accuracy of one-twentieth of a pixel is required. Format con-

version may be required as part of the matching process. Many compression systems have analog composite NTSC or PAL inputs and/or outputs. 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 coder) there are two further requirements for the picture quality measurement instrument: an excellent quality composite decoder and a reference sequence that includes the composite artifacts. Experiments conducted at Tektronix have indicated that picture quality testing where composite encode and decode processes are included will tend to mask measurements of compression systems with small amounts of degradation, such as, MPEG-2 main profile @ main level with 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 perfect NTSC/PAL or very 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)

¹⁵ANSI T1A1.801.01-1995, "Digital Transport of Video, Teleconferencing/Video Telephony Signals, Video Test Scenes for Subjective and Objective Performance Assessment".

based on the JND image quality metric.

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. Beyond that is an operational parameter that may not be satisfactory with general program material. Time to make the measurement is an important

feature in test equipment. If the picture matching; gain, spatial alignment, etc., is to be done on program material, a large amount of compute time will be required to make correlation calculations. This is in addition to the time required to just make the measurement after the two video scenes are correctly matched. Therefore, it is proposed that some known alignment signals, or calibration stripes, be added to the video sequences for rapid picture matching as shown in Figure 13. 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.

Picture Quality Measurement Instruments

The need for objective measurement of picture quality (degra-

ation with respect to a reference) is well established and immediate. Formal and informal subjective picture quality assessment has been used to develop, test, install and operate today's compressed television systems. In this guide, we have emphasized the continuing need for traditional test methods and described the new methods being proposed for objective measurement of picture quality. Tektronix and Sarnoff Labs are cooperating on the development of a picture quality measurement product based on the JND image quality metric and the signal processing required for operation within the complete modern television system.

This is an exciting new measurement paradigm for the television and telecommunications industries. Please contact Tektronix to express your interest in this technology so you can be informed as more technical and product information becomes available. Further theoretical information and experiment data will be disseminated by revisions of this guide, papers presented at conferences, informational seminars and publication of articles in journals and magazines.

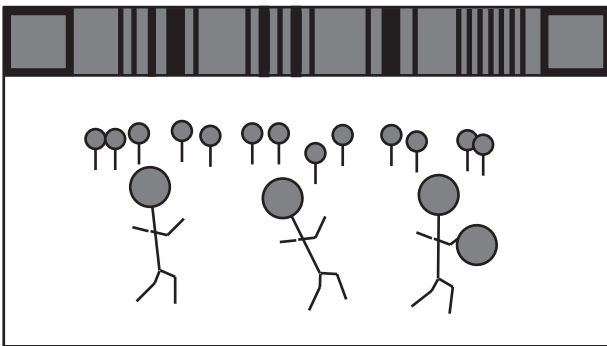


Figure 13. Calibration stripes (top of picture).

For further information, contact Tektronix:

World Wide Web: <http://www.tek.com>; **ASEAN Countries** (65) 356-3900; **Australia & New Zealand** 61 (2) 888-7066; **Austria, Eastern Europe, & Middle East** 43 (1) 7 0177-261; **Belgium** 32 (2) 725-96-10; **Brazil and South America** 55 (11) 3741 8360; **Canada** 1 (800) 661-5625; **Denmark** 445 (44) 850700; **Finland** 358 (9) 4783 400; **France & North Africa** 33 (1) 69 86 81 08; **Germany** 49 (221) 94 77-400; **Hong Kong** (852) 2585-6688; **India** 91 (80) 2275577; **Italy** 39 (2) 250861; **Japan** (Sony/Tektronix Corporation) 81 (3) 3448-4611; **Mexico, Central America, & Caribbean** 52 (5) 666-6333; **The Netherlands** 31 23 56 95555; **Norway** 47 (22) 070700; **People's Republic of China** (86) 10-62351230; **Republic of Korea** 82 (2) 528-5299; **Spain & Portugal** 34 (1) 372 6000; **Sweden** 46 (8) 629 6500; **Switzerland** 41 (41) 7119192; **Taiwan** 886 (2) 765-6362; **United Kingdom & Eire** 44 (1628) 403300; **USA** 1 (800) 426-2200

From other areas, contact: Tektronix, Inc. Export Sales, P.O. Box 500, M/S 50-255, Beaverton, Oregon 97077-0001, USA (503) 627-1916



Copyright © 1997, Tektronix, Inc. All rights reserved. Tektronix products are covered by U.S. and foreign patents, issued and pending. Information in this publication supersedes that in all previously published material. Specification and price change privileges reserved. TEKTRONIX and TEK are registered trademarks.