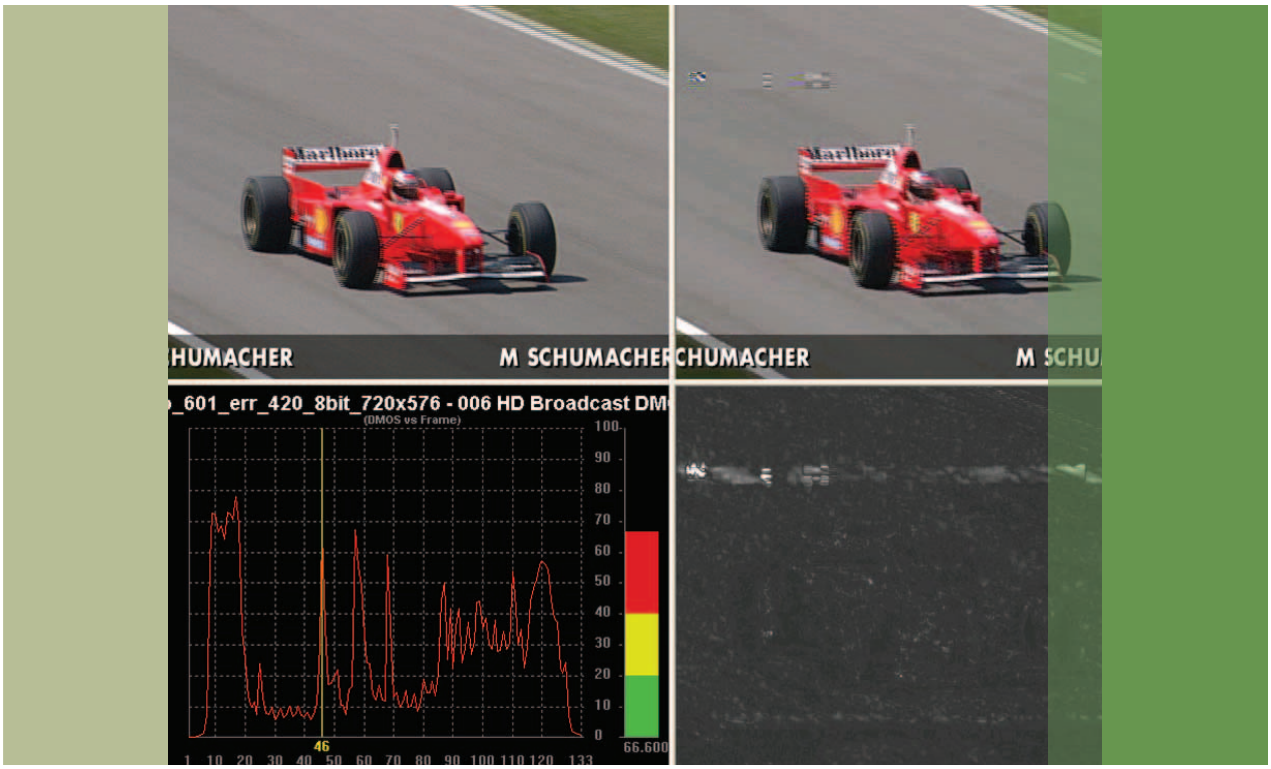


Understanding PQR, DMOS, and PSNR Measurements

Introduction

Compression systems and other video processing devices impact picture quality in various ways. Consumers' quality expectations continue to rise as analog video technology transitions to digital technology and standard definition transitions to high definition. With digital technology, video equipment manufacturers, broadcasters, network operators and content providers cannot rely solely on signal measurements and picture monitors to assess picture quality. They need other tools to verify that their devices, systems, or processes have not introduced impairments in video content that will affect perceived picture quality.

Tektronix®

Video equipment manufacturers want to minimize the impairments their products introduce in video content. Video product development and manufacturing teams need to make accurate, reliable, and repeatable picture quality assessments many times during the development process, not just once on the final product. Profitability pressures can lead to difficult tradeoffs as designers attempt to optimize performance and meet target product costs. Time-to-market pressures limit the time available for quality assurance testing.

Video broadcasters and operators of communication networks that carry video content rely on picture quality assessments when qualifying new video equipment they deploy in their networks. Once they install these products in their networks, they need to determine how various device settings and system configurations affect picture quality. In operating networks, the engineering staff benefits from picture quality evaluation that can detect system degradations before they become picture quality problems that generate viewer complaints.

Video content producers must deliver video content in an ever-increasing number of formats into a media environment that is growing more diverse. They need to effectively assess picture quality as they repurpose video content for these different applications.

Many organizations use an informal method of subjective picture quality assessment that relies on one person or a small group of people who demonstrate an ability to detect video quality impairments. These are the organization's "golden eyes." Subjective picture quality ratings by these "golden eyes" may match the end consumer's video experience. However, these discerning viewers may see artifacts that the average viewer might miss. Projects may experience delays or may be restricted to a small number of evaluations because of limited access to "golden eye" evaluators. Evaluation costs can become an issue, especially if the team uses a "golden eyes" evaluator from outside the organization. Subjective evaluations can easily take an hour or more. In these situations, evaluator error due to fatigue becomes a factor.

These factors have led organizations to consider alternative approaches to subjective picture quality evaluation. Researchers have developed several different methods of conducting formal subjective picture quality assessments. The ITU-R BT.500 recommendation describes several methods, along with requirements for selecting and configuring displays, determining reference and test video sequences, and selecting subjects for viewing audiences.

Such subjective picture quality assessments are expensive and time consuming. Testing professionals must recruit and qualify a suitable viewer audience, prepare the test facility, carefully conduct the tests and analyze the results. Some organizations could possibly afford a small number of these tests at certain points during the design and implementation of the product. However, most organizations cannot afford the expense and time to carry out repeated testing throughout the development process. They cannot afford to use this type of testing to optimize product design, tune video systems for optimal performance, or as part of their ongoing quality assurance and periodic maintenance processes.

Instead, engineering, maintenance, and quality assurance teams turn to instruments that make objective picture quality measurements for this repeated picture quality assessment. *Full-reference* measurements compare a reference video sequence and a test video sequence. In the standard case, the test video is a processed version of the reference video, where the processing has introduced differences between the reference and test videos. *No-reference* measurements operate only on test video sequences. *Reduced-reference* measurements base picture quality assessments on extracted properties of the reference and test videos rather than making a pixel-by-pixel comparison.

The Tektronix PQA500 offers full-reference objective picture quality measurements that engineering, maintenance, and quality assurance teams can use to make accurate, reliable and repeatable picture quality measurements. They can make these measurements more rapidly and cost effectively than testing with actual viewers. Over a wide range of impairments and conditions, the PQA500's Difference Mean Opinion Score (DMOS) measurements can help evaluation teams determine how much the differences introduced in test videos degrade subjective picture quality. Picture Quality Rating (PQR) measurements can help these teams determine how much viewers will notice differences between the reference and test videos, especially in the critical case of high-quality video when differences are near the visibility threshold. Finally, the PQA500 offers the traditional Peak Signal-to-Noise Ratio (PSNR) measurements as a quick, rough check for picture quality problems and for use in diagnosing these problems.

The following sections describe key concepts associated with these measurements, examine essential elements in configuring and interpreting PQR, DMOS and PSNR measurements, and discuss the most effective use of these measurements in assessing picture quality.

Figure 1.1. MSE=27.10

Figure 1.2. MSE=21.26

Figure 1. Image with Lower Mean Squared Error has Poorer Picture Quality.

Subjective Assessment and Objective Picture Quality Measurement

If people perceived all changes in video content equally, assessing picture quality would be relatively easy. A measurement instrument could simply compute the pixel-by-pixel differences between the original video content (the reference video) and the content derived from this reference video (the test video). It could then compute the Mean Squared Error (MSE) of these differences over each video frame and the entire video sequence. This is the noise introduced by the video device, system, or process.

However, people are not mechanical measuring devices that treat all differences equally. Many factors affect the viewer's ability to perceive differences between the reference and test video. Figure 1 illustrates this situation. The video frame shown in Figure 1.1 has greater MSE with respect to the original reference video than the video frame in Figure 1.2. However, the error in Figure 1.1 has high spatial frequency, while the error in Figure 1.2 consists of blocks containing much lower spatial frequencies. The human vision system has a stronger response to the lower spatial frequencies in Figure 1.2 and less response at the higher spatial frequencies in Figure 1.1. Subjectively, Figure 1.2 is worse than Figure 1.1, even though the MSE measurement would assess Figure 1.1 as the poorer image.

Clearly, human visual perception is not equivalent to simple noise detection. Objective picture quality measurements that only measure the noise difference between the reference and test video sequences, e.g. PSNR, will not accurately and consistently match viewers' subjective ratings. To match subjective assessments, objective picture quality measurements need to account for the characteristics of human visual perception.

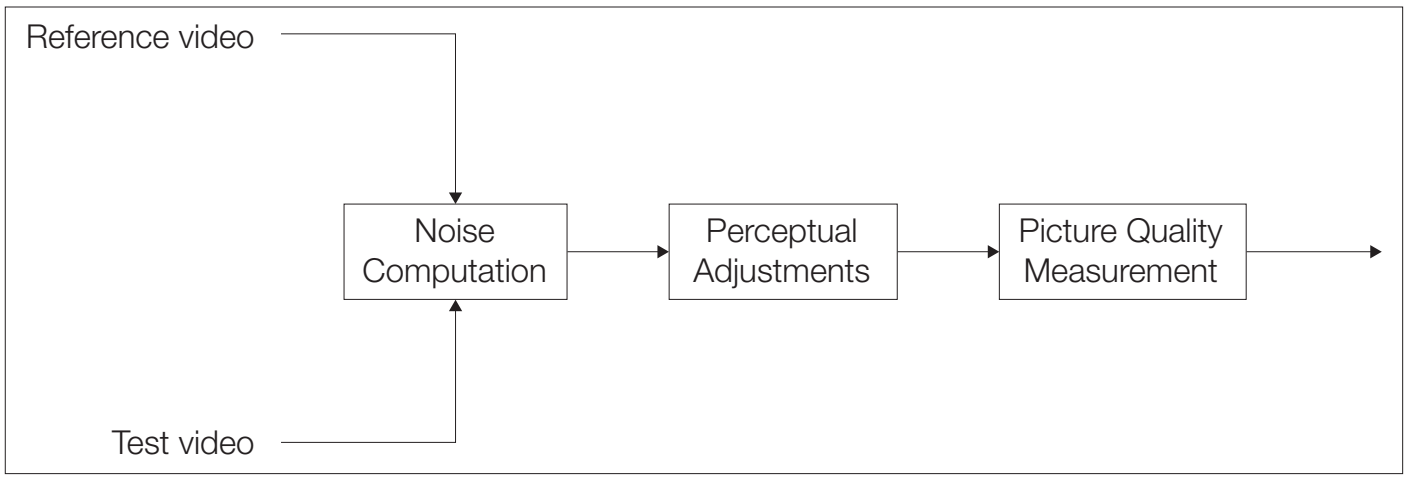


Figure 2. Noise-based Objective Picture Quality Measurements.

Figure 2 diagrams one of the two categories of full-reference objective picture quality measurements. Noise-based measurements compute the noise, or error, in the test video relative to the reference video. The PSNR measurement is a commonly-used method in this measurement category.

The PSNR measurement is especially helpful in diagnosing defects in video processing hardware and software. Changes in PSNR values also give a general indication of changes in picture quality. However, it is well-known that PSNR measurements do not consistently match viewers' subjective picture quality assessments.

Alternative versions of the PSNR measurements adjust the base measurement result to account for perceptual factors and improve the match between the measurement results and subjective evaluations. Other noised-based picture quality measurements use different methods to determine noise and make perceptual adjustments.

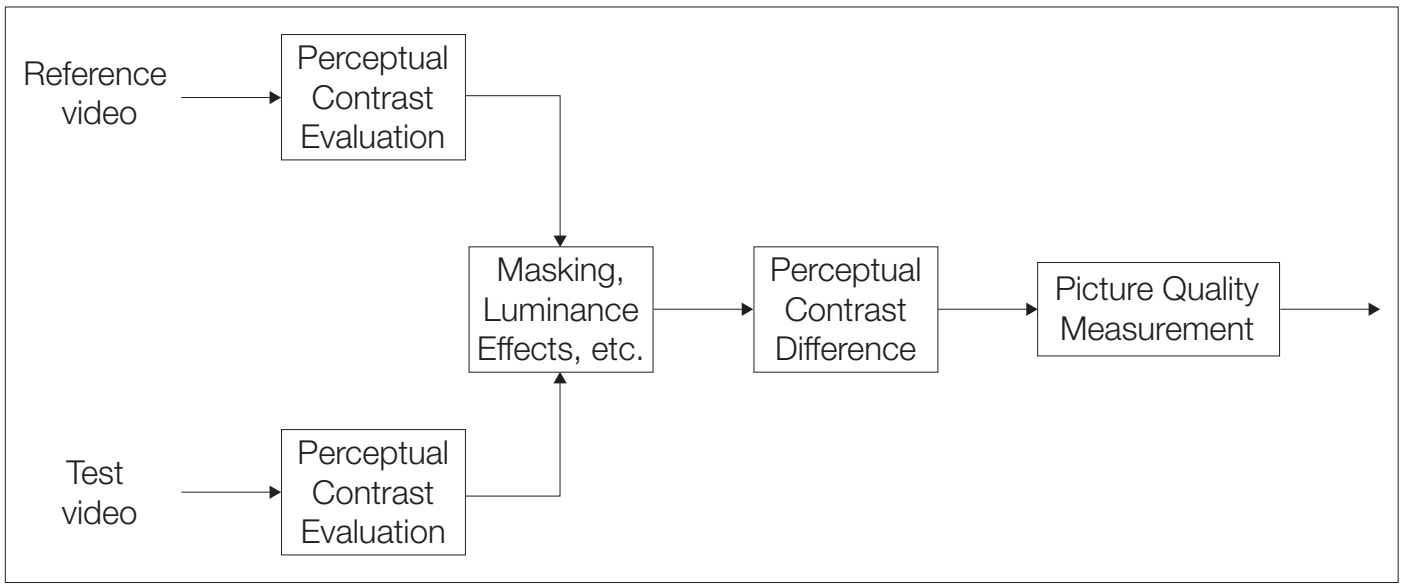


Figure 3. Perceptual-based Objective Picture Quality Measurements.

Figure 3 diagrams the second category of full-reference objective picture quality measurements. Perceptual-based measurements use human vision system models to determine the perceptual contrast of the reference and test videos. Further processing accounts for several other perceptual characteristics. These include relationships between perceptual contrast and luminance and various masking behaviors in human vision. The measurement then computes the perceptual contrast difference between the reference and test videos rather than the noise difference. The perceptual contrast difference is used directly in making perceptual-based picture quality measurements. With an accurate human vision model, picture quality measurements based on these perceptual contrast differences will match viewers' subjective evaluations.

The PQA500 offers one noise-based picture quality measurement, PSNR, and two perceptual-based picture quality measurements, the PQR and DMOS measurements. The sections below describe the configuration, interpretation, and use of the PQR and DMOS measurements, but will not describe the conceptual foundation of perceptual-based objective picture quality measurements or the human vision system model used in these measurements. The application note titled "Perceptual-based Objective Picture Quality Measurements" describes these key concepts.

Picture Quality Rating Measurements

The Picture Quality Rating measurement was introduced on the Tektronix PQA200 Picture Quality Analyzer and was offered on its successor, the PQA300. PQR measurements convert the perceptual contrast difference between the reference and test videos to a value representing viewers' ability to "notice" these differences between the videos. Perceptual sensitivity experiments measure the viewer's ability to notice differences in terms of Just Noticeable Differences (JNDs). In the PQR measurement, 1 PQR equals 1 JND.

Just Noticeable Differences

The concept of Just Noticeable Difference (JND) dates to the early 19th century and the work of E.H. Weber and Gustav Theodor Fechner on perceptual sensitivity. Most commonly, measurements of perceptual sensitivity involve repeated measurements with a single test subject.

Experiments to measure Just Noticeable Differences compare two images or videos: a reference video and a test video derived from the reference video that contains impairments. We can represent the test video as follows:

$$video_{test} = video_{reference} + k * (video_{impaired} - video_{reference})$$

where:

$video_{reference}$ is the reference video sequence,

$video_{impaired}$ is the reference video sequence with added impairments, and

k is a weighting factor $0 < k < 1$ that can be adjusted during the test.

In the test, the viewer is shown the $video_{test}$ and $video_{reference}$ pair several times for a particular value of k and is asked to identify which one of the pair has the impairments. The test is called a forced-choice pairwise comparison because the viewer must choose one of the two videos.

For low k values, when there is little difference between $video_{reference}$ and $video_{test}$, the viewer will be guessing. The percentage of correct responses will be near 50% when k is low. As k increases, the percentage of correct responses will increase. When the viewer can correctly identify the $video_{test}$ sequence on 75% of these trials, the $video_{test}$ and $video_{reference}$ sequences differ by 1 JND.

A 1 JND difference corresponds to approximately 0.1% perceptual contrast difference between the reference and test videos. With this perceptual contrast difference, most viewers can barely distinguish the test video from the reference video in the forced-choice pairwise comparison. At this, and at lower levels of perceptual contrast difference, viewers will perceive the test video as having essentially equal quality to the reference video.

There is wide agreement on the definition for 1 JND. Variations arise in definitions for larger JND values. For example, one researcher defines 2 JND as the point where viewers choose the impaired video in 93.75% of the trials [1]. Another researcher defines 87% correct responses as a 2 JND difference between the reference and test video [2]. These variations occur because of differences in applications and approaches to modeling the probability distribution of the trials.

However, researchers agree that differences become clearly noticeable, or “advertisable,” above 2 JND. They also agree that the forced-choice pairwise comparison experiment “saturates” between 2 and 3 JND. If the reference and test videos differ by 3 JND or more, viewers will always notice the impairment and choose $video_{test}$ 100% of the time.

Researchers use a technique called “stacking” to extend the JND scale. In this technique, N JND is defined by using video that has an (N-1) JND difference from the original $video_{reference}$ as the reference video in a forced-choice pairwise comparison experiment. The experiment determines a $video_{test}$ with a 1 JND difference from this new reference video. The difference between this $video_{test}$ and the original $video_{reference}$ is defined to be N JND. Repeatedly applying this stacking technique, starting with videos that have low JND values relative to the original $video_{reference}$, can build an extended JND scale to any desired amount.

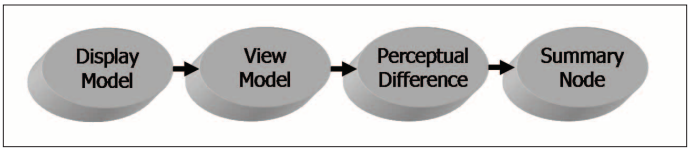


Figure 4. PQA500 Processes.

Configuring PQR Measurements

As previously described, the perceptual-based objective picture quality measurements in the PQA500 (PQR and DMOS) use a human vision system model to compute the perceptual contrast difference between the reference and test videos. Like the actual human vision system, this human vision system model operates on light. Thus, the PQA500 must convert the data in the reference and test video files into light values. This conversion process introduces several factors that influence PQR and DMOS measurements.

In a subjective picture quality evaluation, the light reaching a viewer comes from a particular type of display. The display's properties affect the spatial, temporal and luminance characteristics of the video the viewer perceives.[1]

Viewing conditions also affect differences viewers perceive in a subjective evaluation. In particular, changes in the distance between the viewer and the display screen and changes in the ambient lighting conditions can affect test results.

Since display characteristics and viewing conditions can affect subjective evaluations, objective picture quality measurements that attempt to match subjective ratings must account for these conditions. Figure 4 shows the PQA500 processes that deal with these aspects.

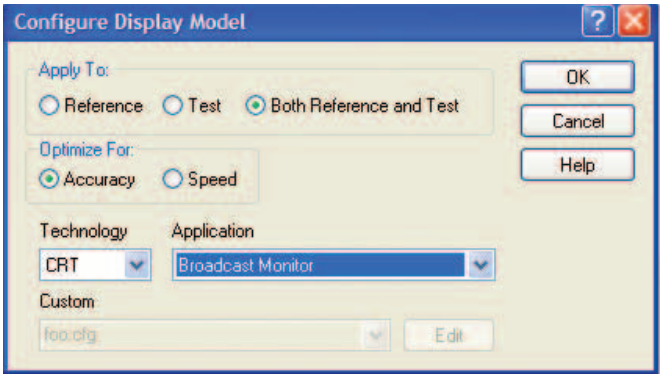
- The Display Model converts the luminance information contained in the reference and test video data files into light values based on display characteristics.
- The View Model adjusts the light values generated by the Display Model to determine the light values that would reach the viewer's eyes based on viewing distance and ambient lighting conditions.

The Perceptual Difference process creates the perceptual contrast difference map used in determining the perceptual-based PQR and DMOS measurements. See the application note titled "Perceptual-based Objective Picture Quality Measurements" for more information on this topic. The Summary Node controls the computation and display of PQA500 measurements. See the PQA500 User Manual and PQA500 Technical Reference for more information on this process.

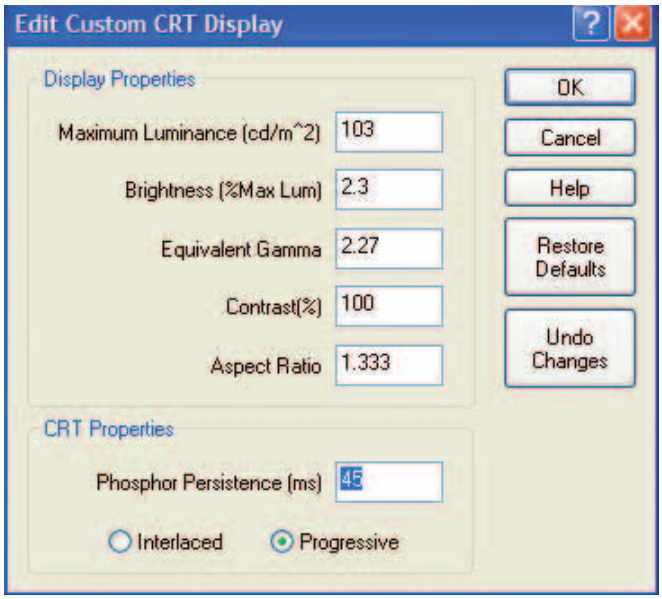
Evaluation teams conducting subjective evaluation can select the display technology and viewing conditions. The PQA500 offers evaluation teams this same capability with objective picture quality measurements. The PQA500 provides a set of pre-configured measurements with pre-determined choices for display characteristics and viewing conditions. These can be used for evaluations or as starting points for creating custom measurements with different choices for display technologies or viewing conditions. Custom measurements are created by "editing" the processes shown in Figure 4.[2]

[1] See the application note titled "Perceptual-based Objective Picture Quality Measurements" for more information on the inter-relationship of spatial, temporal, and luminance characteristics in viewer perception of video quality.

[2] See the PQA500 User Manual and PQA500 Technical Reference for more information on creating custom measurements.



(a)



(b)

Figure 5. Display Model Configuration.

In configuring custom PQR (or DMOS) measurements, different Display Models correspond to different display technologies. The PQA500 has several built-in Display Models covering a range of CRT, LCD, and DLP technologies and includes the ability to create custom display models. Figure 5a shows the configuration screen used to select a Display Model. Figure 5b shows the parameters available for creating custom Display Models.

The list of pre-configured measurements on the PQA500 includes four PQR measurements that use different Display Models. The “SD Broadcast PQR” measurement uses a Display Model which converts the video file data into light in a manner consistent with the behavior of an interlaced scanned CRT display appropriate for monitoring video in a broadcast center. The Display Model in the “HD Broadcast PQR” measurement corresponds to a similar broadcast quality CRT display, but with a progressive scan. The Display Models in the “CIF and QVGA PQR” and “D-Cinema PQR” pre-configured measurements correspond to PDA/Mobile phone-style LCD and DLP display technologies, respectively.

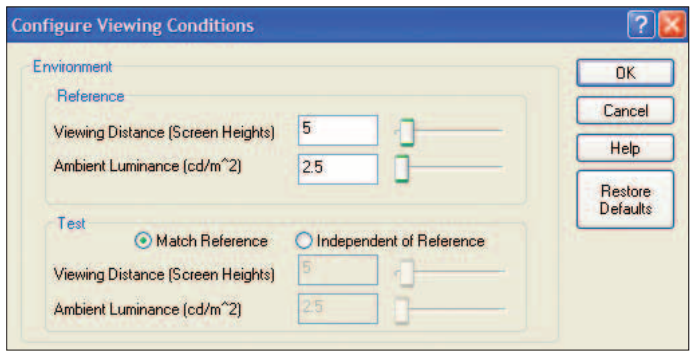


Figure 6. View Model Configuration.

Figure 6 shows the configuration screen used to set viewing conditions for PQR (and DMOS) measurements. Viewing distance is specified in screen heights and ambient luminance is specified in candela/meter^2.

Appropriate viewing conditions are set for each pre-configured PQR measurement listed above. For example, in the SD Broadcast PQR and HD Broadcast PQR measurements, the viewing distance is set at the conventional 5 screen heights. The CIF and QVGA Broadcast PQR measurement uses a viewing distance of 7 screen heights and increases the ambient luminance. This lower resolution display technology is often used in personal video devices. People tend to use these devices in brighter light conditions and the smaller screen size means the typical viewing distance spans more screen heights. Conversely, in digital cinema applications, viewers watch video on very large screens in very low light. Thus, the D-Cinema PQR measurement uses lower values for viewing distance and ambient luminance.

Because the PQA500’s human vision system model operates on light values, every PQR and DMOS measurement has a Display Model and a View Model. As described, these elements determine the display technology and viewing conditions needed to convert data values into light values and compute perceptual contrast differences.

Figure 7. Interlaced Scan Effects.

However, in many applications the effects of display technology or viewing conditions may not play a significant role in picture quality assessment. Frequently, teams evaluating picture quality may not know, may not control, or may not care about the displays that will eventually show the video content or about the final viewing conditions. For example, an engineering team may want to compare picture quality from several different encoders and is completely "agnostic" about display technologies and viewing conditions.

In these applications, teams can use the pre-configured PQR and DMOS measurements available on the PQA500 without modification to the Display Model or viewing condition parameters. They can simply choose the measurement whose configuration best fits the application. One-time adjustments can tailor the measurement to the application if needed, but there is no need to make multiple measurements with different display technologies or viewing conditions.

Of course, applications that do care about the impact of display technologies and viewing conditions on picture quality can configure the PQA500's PQR and DMOS measurements to thoroughly examine these effects. Generally, as the differences between reference and test videos decrease, it becomes more important to measure video quality with different display technologies and viewing conditions to ensure the content reaching the end consumer has acceptable quality over the range of viewing environments.

For PQR and DMOS measurements configured to use interlaced scan display technology, interlaced scanning effects can impact the measurement results. Two sets of measurement conditions affect the results. In the first set of conditions, (1) the data in the reference and test video files are organized in the same scanning format, and (2) the measurement is configured so the reference and test videos use the same Display Model. In this case, the perceptual contrast difference map[3] may show some evidence of the interlaced scan in bright regions of the test video if there are differences between the reference and test videos.

In the second set of conditions, one or both of the items listed do not apply. For example, the reference video might be stored in a progressive scan format while the test video is stored in an interlaced scan format. In another case, the video processing that created the test video from the reference video might have scaled and shifted the video. When the PQA500 spatially aligns the test and reference videos, the resulting interlaced scans will not align.

Any of these situations could create perceptual contrast differences between the reference and test videos. These appear as horizontal lines on the perceptual contrast difference map, as expected from an interlaced scan effect (Figure 7). This additional perceptual contrast difference will increase the PQR or DMOS measurement result.

Viewers in subjective evaluations that used interlaced scan displays and the same reference and test video sequences would also see these effects. However, they would not necessarily find them to be quality problems.

[3] See the application note titled "Perceptual-Based Objective Picture Quality Measurements" for more information on perceptual contrast difference maps.

When using interlaced scan display technologies in a PQR or DMOS measurement, any differences in interlaced scan format between the reference and test videos will affect the measurement result. If these differences are not important for the application, reconfiguring the PQR or DMOS measurement to use a progressive scan display technology will reduce this effect.

In most cases, however, if the reference and test video sequences were created using an interlaced scan, evaluators will minimize interlaced scan effects by using interlaced scan display technology in the measurement. For example, reference and test videos acquired in broadcast studios often meet the first set of conditions. In these situations, using a progressive scan display technology will mix the video fields. Differences between the reference and test videos in regions of motion will show interlaced scan effects. Viewers would also see these effects in subjective evaluations that used a progressive scan monitor.

Evaluators using interlaced scan video should consider changing to progressive scan display technology in a PQR or DMOS measurement only if the first set of measurement conditions does not apply and the resulting interlaced scan effects are not important in their evaluation.

Interpreting PQR Measurements

The PQR scale introduced in the Tektronix PQA200 and carried forward in the PQA300 was developed in collaboration with Sarnoff Laboratories and was based on their work in modeling Just Noticeable Difference experiments. When Tektronix introduced an improved human vision system model, DMOS measurements, and support for HD formats on the PQA500, the PQR scale was calibrated to ensure results agreed with the PQA300 measurements on SD video formats. In both the PQA300 and PQA500, measurements were carefully calibrated using data from perceptual sensitivity experiments to ensure that 1 PQR corresponded to 1 JND and that measurements around this visibility threshold matched the perceptual sensitivity data.

The following scale offers some guidance in interpreting PQR measurement results.

- **0:** The reference and test image are identical. The perceptual contrast difference map is completely black.
- **<1:** The perceptual contrast difference between the reference and test videos is less than 0.1% or less than 1 JND. Viewers cannot distinguish differences between videos. Video products or systems have some amount of video quality “headroom.” Viewers cannot distinguish subtle differences introduced by additional video processing, or by changes in display technology or viewing conditions. The amount of headroom, i.e. the level of difference viewers will not notice, decreases as the PQR value approaches 1.
- **1:** The perceptual contrast difference between the reference and test videos equals approximately 0.1% or 1 JND. Viewers can barely distinguish differences between the videos. Video products or systems have no amount of video quality headroom. Viewers are likely to notice even slight differences introduced by additional video processing, or by changes in display technology or viewing conditions.
- **2-4:** Viewers can distinguish differences between the reference and test videos. These are typical PQR values for high bandwidth, high quality MPEG encoders used in broadcast applications. Generally recognized as excellent to good quality video.
- **5-9:** Viewers can easily distinguish differences between the reference and test videos. These are typical PQR values for lower bandwidth MPEG encoders used in consumer-grade video devices. Generally recognized as good to fair quality video.
- **>10:** Obvious differences between reference and test videos. Generally recognized as poor to bad quality video.

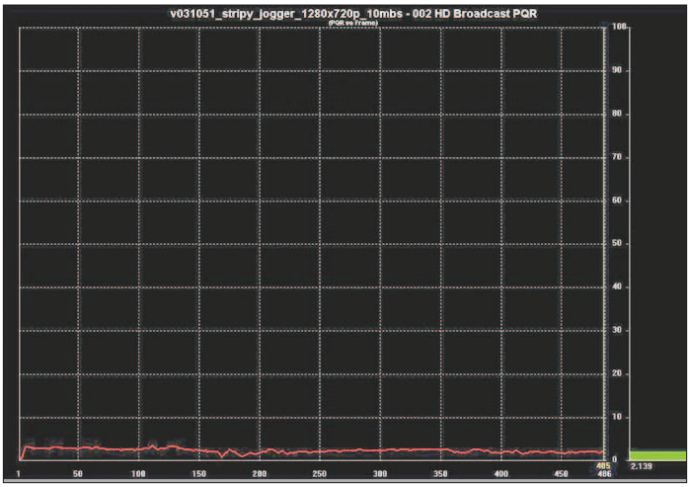


Figure 8. PQR Measurement.

Figure 8 shows the results of a typical PQR measurement.

Perceptual contrast differences near 1 JND (~ 0.1%) are called *threshold* conditions. Contrast differences at these levels just cross internal thresholds in the viewer's visual system. *Supra-threshold* conditions occur at perceptual contrast difference levels >> 0.1%. There is no definite value of perceptual contrast that marks the boundary of the supra-threshold region. PQR measurements with values below 2 are near the visibility threshold. PQR measurements with values above 6 are well into the supra-threshold region.

As noted above, researchers use a "stacking" technique to establish JND levels for supra-threshold conditions. In essence, this technique determines values in supra-threshold conditions by repeating an experiment performed at threshold conditions.

Researchers who examine perceptual contrast report differences in how the visual system responds in the supra-threshold region compared to the threshold region ([3], [4], [5]). For example, area and spatial frequency have a much larger effect in the threshold region than in the supra-threshold region. This suggests that JND levels constructed by "stacking" may not precisely model viewers' perception in the supra-threshold region.

Recognizing the implications of threshold and supra-threshold conditions in establishing JND levels adds some insight into interpreting the PQR measurements based on this concept, but it does not fundamentally compromise the PQR measurement. In particular, the PQR measurement is especially helpful in applications dealing with high-quality video.

In these applications, engineering or quality assurance teams typically want to determine if products or systems have introduced any amount of noticeable differences in the test video. In other words, these teams are assessing video content at threshold conditions. None of the supra-threshold concerns apply in this case. The connection between perceptual contrast differences is well known and well understood. The PQR measurement is calibrated using data gathered from subjective evaluations and can give results well matched to perceptual sensitivity experiments.

Applications involving reference and test videos with perceptual contrast differences in the supra-threshold region can also effectively use PQR measurements. The extended PQR scale based on the stacked JND concept conforms to standard industry and academic practices. PQR measurements in supra-threshold regions can provide useful comparisons of picture quality between video products and systems, and helpful supplementary data to subjective evaluations.

However, the interpretation of PQR measurements in supra-threshold regions is somewhat ambiguous. The forced-choice pairwise comparison used in setting JND levels saturates around 3 JNDs. Research in supra-threshold perceptual contrast raises questions about using the stacked JND method to extend the scale. As a result, in the supra-threshold region, the relationship between perceptual contrast differences and JND levels, and thus PQR levels, is not completely clear.

The DMOS measurement described in the next section spans threshold and supra-threshold regions without these concerns. It can assess picture quality over a broad range of impairment levels and perceptual conditions. Combining DMOS and PQR measurements give engineering, verification, and quality assurance teams unique capabilities to efficiently and effectively assess picture quality.

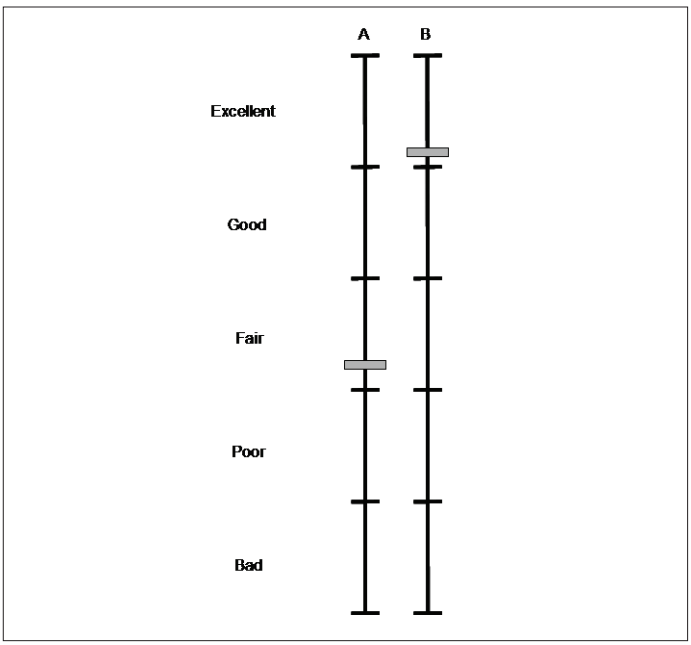


Figure 9. Quality Scale.

Difference Mean Opinion Score Measurements

The perceptual contrast difference map produced by the PQA500's human vision system model contains information on differences viewers will perceive between reference and test videos. As a result, the PQA500 can predict how viewers would score the test videos if they evaluated the video content using methods described in ITU-R BT.500. In particular, the PQA500 can produce predicted Difference Mean Opinion Score (DMOS) values for test videos. Unlike testing with people, the PQA500 can produce a DMOS result for each frame in the test video sequence as well as the overall sequence.

Subjective Picture Quality Evaluation Methods in ITU-R BT.500

Recommendation ITU-R BT.500-11 describes several methods for the subjective assessment of television picture quality. They differ in the manner and order of presenting reference and test videos. They share characteristics for scoring video and analyzing results.

In methods that compare both reference and test videos, viewers grade the videos separately. They use the grading scale shown in Figure 9. The scale is divided into equal lengths using the ITU five-point quality scale. For each video in a reference/test pair, A and B, viewers place a mark at any location on the scale (continuous quality scale).

The marks on the grading scale are converted to a numeric value representing viewers' opinion scores for the videos they evaluate in the test. In this conversion, marks in the "Excellent" region result in values between 0 and 20 while marks in the "Bad" region result in values between 80 and 100.

Opinion scores are collected from each viewer participating in the test. Subjective evaluations typically involve groups of around two dozen viewers. These scores are averaged to create the Mean Opinion Score or MOS for the evaluated videos. The MOS for the reference video sequences is subtracted from the MOS for the test video sequences. This generates a Difference Mean Opinion Score or DMOS for each test sequence. The DMOS value for a particular test video sequence represents the subjective picture quality of the test video relative to the reference video used in the evaluation.

Before viewers evaluate any video, they are shown training video sequences that demonstrate the range and types of impairments they will assess in the test. ITU-R BT.500 recommends that these video sequences should be different than the video sequences used in the test, but of comparable sensitivity. In other words, the training video sequences cover the range from the "best case" to the "worst case" videos the viewers will see in the test.

Without the training session, viewers' assessments would vary widely and change during the test as they saw different quality videos. The training session ensures coherent opinion scores. However, this means the DMOS results for test sequences depend on the video content shown in the training session.

Suppose a test audience was trained using video sequences covering a very narrow range of quality. During the test, this audience views Video Clip A and gives it a DMOS of 45. A different test audience is trained using video sequences that cover a wider quality range. The worst case video in this training is lower quality than the worst case video shown to the first test audience. When the second test audience sees Video Clip A, they will assess the video clip as having higher quality than the first test audience. The DMOS result for Video Clip A will be less than 45.

Thus, DMOS scores have a relative character. Their values depend on the range between “best case” and “worst case” videos used in the training sequence. If this range changes, the DMOS value viewers give a test video will also change. The relative character of DMOS scores reflects the relative quality scales of particular applications. For example, we would expect more visible differences in a mobile video application than in a digital cinema application. The video quality range in the mobile video application would differ from the range of video quality seen in the digital cinema application. The videos used in the training sequence, in particular the “best case” and “worst case” videos, are used to capture these differences and normalize the evaluation scale to each application's quality dynamic range.

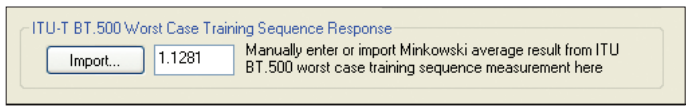


Figure 10. Worst Case Training Sequence Response.

Configuring Predicted DMOS Measurements

The considerations about display technologies and viewing conditions discussed in the section titled “Configuring PQR Measurements” also apply to configuring DMOS measurements. In particular, using interlaced scan display technologies can significantly impact DMOS measurement results when the reference and test videos have different interlaced scan formats. See the earlier section for more information on this topic. In addition to these configuration concerns, DMOS measurements also have a configuration parameter related to the training sessions described in the previous section.

As explained above, the training session held before the actual subjective evaluations ensures consistent scoring by aligning viewers on the “best case” and “worst case” video quality they will see. In effect, the training session establishes the range of perceptual contrast differences viewers will see in the evaluation. The worst case training sequence response configuration parameter performs the same function in a DMOS measurement. This parameter specifies the perceptual contrast difference between the “best case” and “worst case” videos for a particular DMOS measurement.

Figure 10 shows the configuration screen used to set the worst case training sequence response parameter. This parameter is a generalized mean of the perceptual contrast differences between the best case and worst case training video sequences associated with the DMOS measurement. This generalized mean, called the Minkowski metric or k-Minkowksi metric,[4] was calculated by performing a perceptual-based picture quality measurement, either PQR or DMOS, using the best case video sequence as the reference video and the worst case video sequence as the test video in the measurement.

[4] See the application note “Perceptual-based Objective Picture Quality Measurements” for more information on the Minkowski metric.

The PQA500 has several pre-configured DMOS measurements. These DMOS measurements contain different values for the worst case training sequence response parameter, determined by using video sequences appropriate for the measurement. For example, the worst case training sequence response parameter for the SD Broadcast DMOS measurement was determined from standard definition video with marginal quality for broadcast applications. Similarly, a high definition video with marginal quality for broadcast applications was used to set this parameter for the HD Broadcast DMOS measurement. As much as possible, appropriate video sequences were used in configuring other measurements, e.g., a marginal quality sports video was used in configuring the SD Sports Broadcast ADMOS and HD Sports Broadcast ADMOS measurements.[5]

The PQA500's pre-configured measurements provide "starting points" for picture quality evaluation. They serve as templates for creating custom measurements that more precisely address a specific application's characteristics and requirements for picture quality evaluation. In particular, the worst case training sequence responses used in DMOS measurements can easily be changed. The video sequences used for the pre-configured DMOS measurements were selected from a set of available video content. As discussed in the next section, many engineering and quality assurance teams may find it useful to establish their own definition of "worst case" in performing DMOS measurements.

Modifying the worst case training sequence response for a DMOS measurement consists of the following steps:

1. Choose a video sequence that represents the "best case" video for the evaluation. Choose a second video sequence that represents the "worst case" video for the evaluation. The video sequences do not need to be long (10-20 seconds) but should contain the impairments of interest in the test.
2. Perform a perceptual-based picture quality measurement (PQR or DMOS) using the "best case" video as the reference video and the "worst case" video as the test video. The perceptual-based measurement selected should use the same display technology and viewing conditions that will be used in the custom measurement. It does not matter whether a PQR or DMOS measurement is selected. Both measurements use the same Minkowski metric derived from the perceptual contrast difference map.
3. Create a new measurement. Edit the Summary Node in this measurement.[6] In the configuration screen (Figure 10), press the "Import" button. This will open a file browser. Locate and select the results (.csv file) for the measurement performed in step #2. Opening this .csv file will insert the overall Minkowski metric from the test video as the worst case training sequence response for the new measurement.

Interpreting DMOS Measurements

The PQA500's DMOS measurements predict the DMOS values viewers would give the reference and test videos used in the measurement if they evaluated these videos in a subjective evaluation conducted according to procedures defined in ITU-R BT.500. These ITU procedures consist of rating videos on a quality scale. When rating video quality, or any property, on a scale, people do not readily rate items at the extreme ends of the scale. They are not sure if the next item they see will be better or worse than the item they are rating.

[5] The ADMOS measurement is an "Attention-weighted" DMOS measurement. See the PQA500 User Manual and PQA500 Technical Reference for information on the Attention Model and attention-weighted measurements.
[6] See the PQA500 User Manual and PQA500 Technical Reference for more information on creating custom measurements.

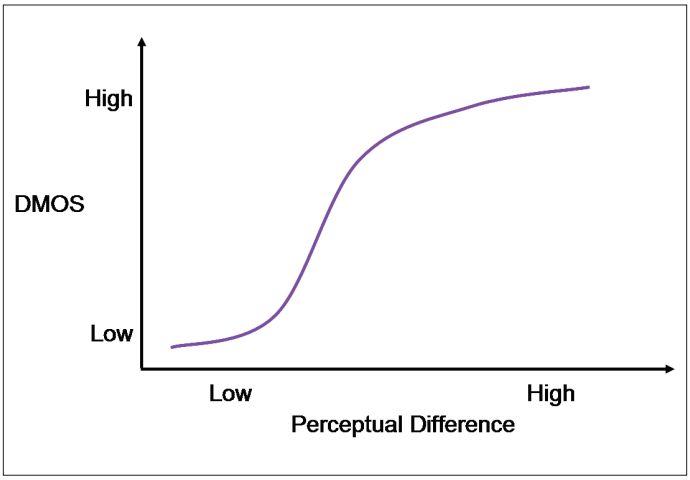


Figure 11. Compression in Subjective Evaluation.

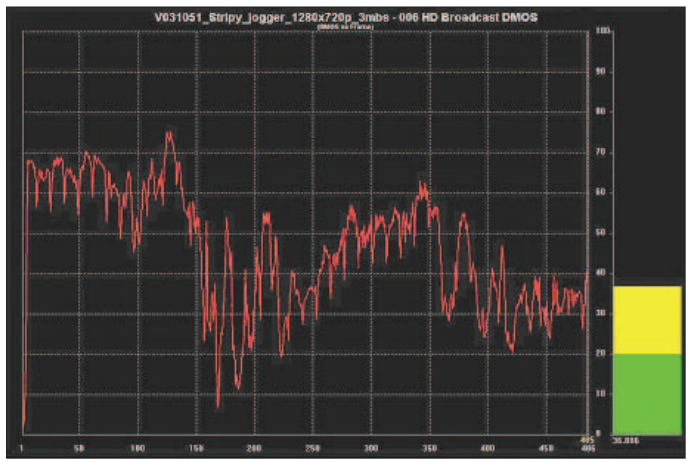


Figure 12. DMOS Measurement.

This behavior is called *compression*. Due to compression, results from subjective evaluations appear qualitatively similar to the S-shaped curve shown in Figure 11. Compression has a significant impact on DMOS values for videos whose quality equals the "worst case" video shown in the training sequence. If viewers used the extreme ends of the quality scale in their ratings, test videos whose quality matched the "worst case" should have DMOS values at the top end of the DMOS scale (near 100). However, due to compression, viewers consistently give test videos with "worst case" quality a DMOS value around 65.

In the PQA500, the procedure used to predict DMOS values from perceptual contrast differences accounts for this compression. Using data from subjective testing, the procedure has been calibrated to track the S-curve response. If the perceptual contrast difference between the reference and test video equals the worst case training sequence response, the DMOS value equals 65.

Figure 12 shows a typical DMOS measurement. In the pre-configured DMOS measurements, values in the 0-20 range indicate test video that viewers would rate as Excellent to Good relative to the reference video. Results in the 21-40 range correspond to viewers' subjective ratings of Fair to Poor quality video. DMOS values above 40 indicate the test video has Poor to Bad quality relative to the reference video. These threshold values can be changed to adjust to application-specific requirements.

The PQA500's PQR measurements predict the results of perceptual sensitivity experiments. The PQA500's DMOS measurements differ because they predict the results of a subjective picture quality rating procedure. Issues around threshold and supra-threshold conditions do not arise in calibrating DMOS measurements. Ample subjective evaluation data exists to calibrate the PQA500's human vision system model in both regions. Independent calibration parameters ensure the model operates appropriately in both threshold and supra-threshold regions. The conversion function used to calculate the predicted DMOS values from the perceptual contrast differences uses a separate calibration and validation procedure, performed after the calibration of the human vision system model. Engineering and quality assurance teams can use the PQA500's DMOS measurement to assess quality over a wide range of impairments and evaluation conditions with confidence that these measurements match subjective ratings.

The perceptual sensitivity experiments and JND concept associated with the PQR measurement does not involve quality scales or training sessions. As explained in "Configuring PQR Measurements" above, different choices for display technologies or viewing conditions can affect PQR measurement results, but there is no concept of "best case" or "worst case" video for these measurements. If two PQR measurements are configured with the same display technology and viewing conditions they will produce the same results.

DMOS measurements behave differently. As described in "Subjective Picture Quality Evaluation Methods in ITU-R BT.500," the same test videos can receive different DMOS values from different viewer audiences. It depends on the video sequences used to train the viewers. Similarly, DMOS measurements configured with the same display technology and viewing conditions can produce different results if they are also configured with different worst case training sequence responses.

In this sense, the DMOS measurement is a relative scale. The DMOS value depends on the worst case training sequence response used to configure the measurement, just as the results of the associated ITU-R BT.500 subjective evaluation depend on the video sequences used to train the viewing audience. When comparing DMOS measurement results, evaluators need to verify that the measurements use the same display technologies, viewing conditions *and* worst case training sequence response parameters.

As noted in the preceding section, picture quality evaluation teams can alter the worst case training sequence response parameter in a DMOS measurement. Reasons for making this configuration change include:

- The evaluation team may have specific video sequences they feel represent "worst case" video for their application. They may want to use the perceptual contrast differences associated with these video sequences to configure DMOS measurements rather than the worst case training sequence responses used in the pre-configured DMOS measurements.
- An application may involve very high quality video that produces low DMOS values. In such a case, the DMOS plots for different video sequences typically lie close to each other at the bottom of the graph shown in Figure 12. To separate these measurement plots, the evaluation team can create a custom DMOS measurement that uses the perceptual contrast difference from one of the test video sequences as the worst case training sequence response. This new measurement will "expand" the DMOS scale and separate the results for the different test videos for easier analysis. This is equivalent to repeating a subjective evaluation with a new viewer audience and training this audience using videos that have a smaller difference in quality between the "best case" and the "worst case" videos.
- An engineering team may be modifying a product or system and want to ensure changes do not degrade picture quality. They can use the PQA500 to measure the picture quality of the current system and use the results of this measurement to set the worst case training response in a custom DMOS measurement. Engineers use this DMOS measurement to assess the picture quality of the product or system as they make modifications. As long as the DMOS result remains lower than 65, they know the picture quality of the modified product or system is as good as or better than the old product or system. The DMOS measurement can also tell the team how much, if any, their modifications have improved video quality compared to the old product.

This ability to "alter the scale" of DMOS measurements by setting the worst case training sequence response enhances their utility in picture quality evaluation. DMOS measurements perform equally well at perceptual contrast levels near the visibility threshold and in supra-threshold conditions. Subjective evaluation data available across this range of conditions helps ensure the predicted DMOS values match subjective assessments.

This combination of factors makes the DMOS measurement an excellent choice for picture quality evaluation teams needing to understand and quantify how differences between a reference and test video degrade subjective video quality. The PQR measurement complements the DMOS measurement by helping these teams determine if viewers can notice this difference, especially near the visibility threshold. The third measurement offered on the PQA500, the PSNR measurement, lets evaluation teams determine the level of difference between the reference and test videos, regardless of viewers' ability to perceive the difference.

$$PSNR(f_n) = 20 \log_{10} \left[\frac{235}{\sqrt{\frac{1}{N_v N_h} \sum_{j=0}^{N_v-1} \sum_{i=0}^{N_h-1} [\, Y_{ref}(i, j, f_n) - Y_{test}(i, j, f_n) \,]^2}} \right]$$

(a)

$$PSNR_{seq} = 20 \log_{10} \left[\frac{235}{\sqrt{\frac{1}{M N_v N_h} \sum_{n=0}^{M-1} \sum_{j=0}^{N_v-1} \sum_{i=0}^{N_h-1} [\, Y_{ref}(i, j, f_n) - Y_{test}(i, j, f_n) \,]^2}} \right]$$

(b)

Figure 13. PSNR Measurement Formulas (dB Units).

Peak Signal-to-Noise Ratio Measurements

The PQA500 calculates a standard Peak-Signal-to-Noise Ratio (PSNR) measurement. It does not make any perceptual adjustments to the measurement results (see discussion in "Subjective Assessment and Objective Picture Quality Measurement").

To calculate the PSNR value, the PQA500 computes the root mean squared (RMS) difference between the reference and test video and divides this into the peak value. It computes the PSNR value for every frame in the test video and for the entire video sequence.

Figure 13a shows the formula for computing the PSNR value for a frame in the test video. Figure13b is the formula for computing the PSNR value for the entire test video sequence. In these formulas, N_h is the number of pixels in the video line, N_v is the number of lines in the video frame, and M is the number of frames in the video sequence. Following convention, the PQA500's pre-configured PSNR measurement reports the results in decibels (dB).

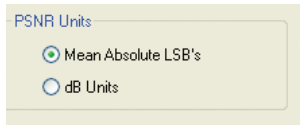
The PQA500 supports 8-bit video formats. In these formats, the largest value for the luminance (Y) component equals 255. The formulas above use a peak value of 235 because the PQA500 makes PSNR measurements in conformance to the T1.TR.74-2001 recommendation titled "Objective Video Quality Measurement Using a Peak-Signal-to-Noise Ratio

$$PSNR(f_n) = \frac{1}{N_v N_h} \sum_{j=0}^{N_v-1} \sum_{i=0}^{N_h-1} \left[\, | \, Y_{ref}(i, j, f_n) - Y_{test}(i, j, f_n) \, | \, \right]$$

(a)

$$PSNR_{seq} = \frac{1}{M N_v N_h} \sum_{n=0}^{M-1} \sum_{j=0}^{N_v-1} \sum_{i=0}^{N_h-1} \left[\, | \, Y_{ref}(i, j, f_n) - Y_{test}(i, j, f_n) \, | \, \right]$$

(b)



(c)

Figure 14. PSNR Measurement Formulas (Mean Absolute LSB Units).

(PSNR) Full Reference Technique" issued by the Video Quality Expert Group (VQEG). This recommendation specifies that the peak value in the PSNR measurement should equal the peak white luminance level of 235.

On occasion, design engineers may want to see a less common measure that calculates the difference between the reference and test videos as Mean Absolute LSBs (least significant bits). The PQA500 offers this measurement as an alternative configuration to the PSNR measurement. Figures 14a and 14b show the formulas for computing this measurement for a frame of the test video and for the test video sequence, respectively. In these formulas, N_h, N_v, and M have the same values noted above.

As the formulas show, this computation is not actually a ratio. Rather, it is an average of the differences across the frame and over the entire sequence. In some applications, knowing the actual noise levels in addition to the ratio of noise to peak can help diagnose picture quality problems more effectively and efficiently. Figure 14c shows the configuration screen in the Summary Node that selects between the two alternative versions of the PSNR measurements.

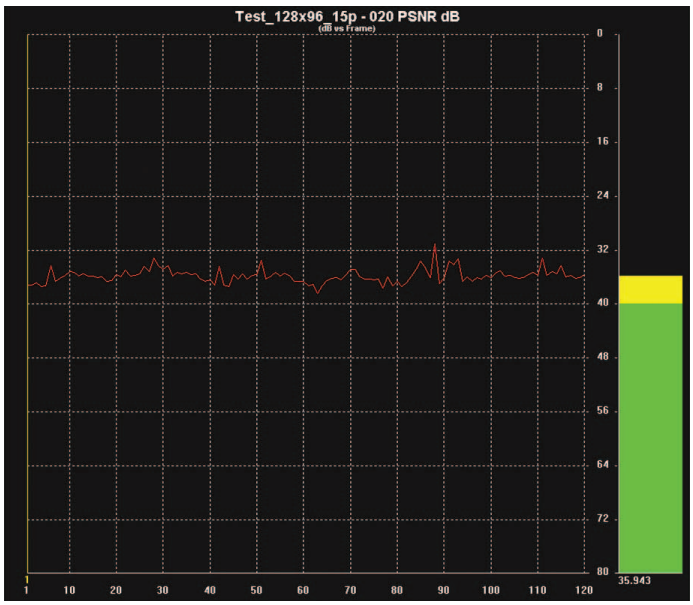


Figure 15. PSNR Measurement.

Figure 15 shows a typical PSNR measurement. In PSNR measurements, as the difference between the reference and test video increases, the PSNR measurement result decreases. On the PQA500, if the reference and test videos are identical, the PSNR measurement result equals 80 dB. If high quality video is used as the reference video in the PSNR measurement, a PSNR value above 40 dB indicates that the test video is also high quality. PSNR values below 30 dB indicate lower quality test video.

Combining PSNR measurements with the perceptual-based measurements on the PQA500 offers unique insight into the impact of differences between the reference and test videos. Figure 16 shows a comparison of a PSNR measurement in Mean Absolute LSBs units (solid blue line) and a DMOS measurement (dotted magenta line). The PSNR measurement shows when differences occur between the two video sequences. The DMOS measurement shows the perceptual impact of these differences.

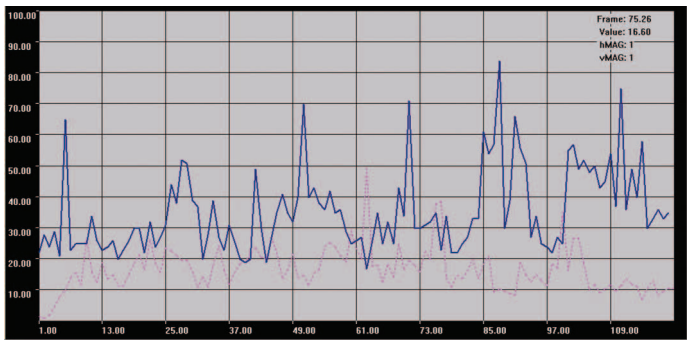


Figure 16. Comparison of PSNR and DMOS Measurements.

In these comparison graphs, evaluation teams can see the how differences do, or do not, impact perceived quality. They can see how adaptation in the visual system affects viewers' perception. For example, a large transition in average luminance during a scene change can mask differences. Comparing the difference map created in the PSNR measurement and the perceptual contrast difference map created in a PQR or DMOS measurement can reveal problem regions within the video field or frame. These comparisons can help engineers more easily map visual problems to hardware or software faults.

The preceding sections have described key concepts in configuring and interpreting the PQR, DMOS and PSNR measurements available on the PQA500. The PQA500 offers additional measurements that complement these primary picture quality measurements. These include measurements that detect video artifacts, e.g., lost edges (blurring), added edges (ringing, mosquito noise), or blockiness. Other measurements weight the results of DMOS, PQR or PSNR with the results from these artifact detectors or from the PQA500's Attention Model. Using the results of these measurements to weight the basic picture quality measurements, evaluators can account for viewers' focus-of-attention or tolerance for different types of artifacts in assessing picture quality. The PQA500 User Guide, the PQA500 Technical Reference, and the application note titled "Picture Quality Analysis for Video Applications" have more information on these PQA500 capabilities and how these capabilities address requirements for picture quality evaluation in various video applications.

Conclusion

Engineering and quality assurance teams need to perform frequent, repeated, and accurate picture quality assessments to diagnose picture quality problems, optimize product designs, qualify video equipment, optimize video system performance, and produce, distribute and re-purpose high quality video content. They cannot afford the time and expense associated with recruiting viewers, configuring tests, and conducting subjective viewer assessments. They need objective picture quality measurements that can make these assessments more quickly than subjective evaluation and at a lower cost. However, these objective measurements should match subjective evaluations as closely as possible.

The PQA500's full-reference objective picture quality measurements uniquely address these requirements. The perceptual-based DMOS and PQR measurements offer results well matched to subjective evaluations. Over a wide range of impairments and conditions, DMOS measurements can help evaluation teams determine how differences between reference and test videos can affect subjective quality ratings. PQR measurements can help these teams determine to what extent viewers will notice these differences, especially for applications that place a premium on high quality video. Finally, the PSNR measurements on the PQA500 give evaluation teams an industry standard aid for diagnosing picture quality problems.

The PQA500's human vision system model used by the PQR and DMOS measurements has the adaptive filtering needed to fully model the temporal and spatial characteristics of contrast sensitivity and accounts for key factors affecting viewers' ability to perceive contrast differences. Configuration parameters let evaluation teams study the impact of different display technologies and viewing conditions on picture quality. Used singly or in combination, the PQA500's objective picture quality measurements offer the accuracy, reliability, and repeatability needed to diagnose picture quality problems, optimize video products and systems, and verify video content quality.

References

[1] Jeffrey Lubin, Michael H. Brill, Roger L. Crane; "Vision Model-Based Assessment of Distortion Magnitudes in Digital Video," David Sarnoff Research Center, Princeton, NJ.

[2] Reinhold Thiel, Paul Clark, Richard B. Wheeler, Paul W. Jones, Marcel Riveccie, Jean-Fabien Dupont.; "Assessment of Image Quality in Digital Cinema Using the Motion Quality Ruler Method," SMPTE Motion Imaging Journal, Vol. 116, No. 2&3, February/March 2007, pp. 61 – 73.

[3] M. Cannon; "A Multiple Spatial Filter Model for Suprathreshold Contrast Perception," in Vision Models for Target Detection and Recognition, ed. Eli Peli (World Scientific Publishing, River Edge, NJ, 1995), pp. 88-117.

[4] Eli Peli, Jian Yang, Robert Goldstein, Adam Reeves; "Effect of luminance on suprathreshold contrast perception," J. Opt. Soc. Am., August 1991, Vol. 8, No. 8, pp. 1352-1359.

[5] Christopher C. Taylor, Zygmunt Pizlo, Jan P. Allebach; "Image Quality Assessment with a Gabor Pyramid Model of the Human Vision System," Proceedings of the 1997 IS&T/SPIE International Symposium on Electronic Imaging Science and Technology, San Jose, CA, 8-14 February 1997, vol. 3016, pp. 58-69.

Contact Tektronix:

ASEAN / Australasia (65) 6356 3900
Austria +41 52 675 3777
Balkans, Israel, South Africa and other ISE Countries +41 52 675 3777
Belgium 07 81 60166
Brazil +55 (11) 3759 7600
Canada 1 (800) 661-5625
Central East Europe, Ukraine and the Baltics +41 52 675 3777
Central Europe & Greece +41 52 675 3777
Denmark +45 80 88 1401
Finland +41 52 675 3777
France +33 (0) 1 69 86 81 81
Germany +49 (221) 94 77 400
Hong Kong (852) 2585-6688
India (91) 80-22275577
Italy +39 (02) 25086 1
Japan 81 (3) 6714-3010
Luxembourg +44 (0) 1344 392400
Mexico, Central, South America and Caribbean 52 (55) 54247900
Middle East, Asia and North Africa +41 52 675 3777
The Netherlands 090 02 021797
Norway 800 16098
People's Republic of China 86 (10) 6235 1230
Poland +41 52 675 3777
Portugal 80 08 12370
Republic of Korea 82 (2) 6917-5000
Russia & CIS +7 (495) 7484900
South Africa +27 11 206 8360
Spain (+34) 901 988 054
Sweden 020 08 80371
Switzerland +41 52 675 3777
Taiwan 886 (2) 2722-9622
United Kingdom & Eire +44 (0) 1344 392400
USA 1 (800) 426-2200
For other areas contact Tektronix, Inc. at: 1 (503) 627-7111

For Further Information

Tektronix maintains a comprehensive, constantly expanding collection of application notes, technical briefs and other resources to help engineers working on the cutting edge of technology. Please visit **www.tektronix.com**

12/08 EA/ 28W-21224-0

Tektronix®