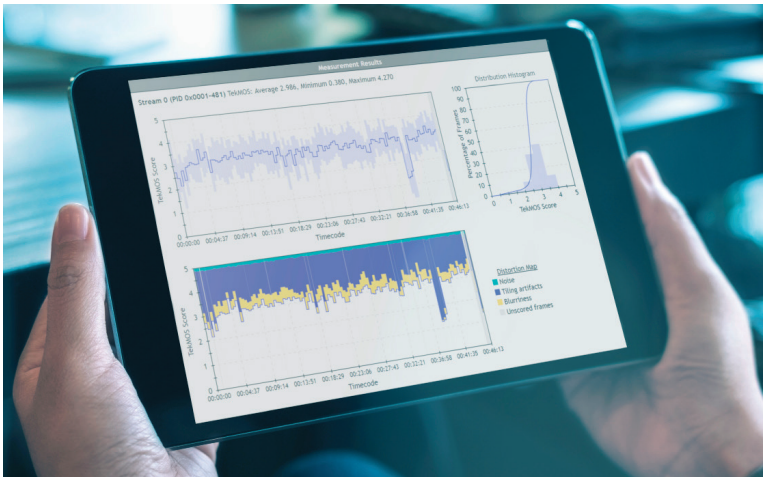

TekMOS

A general-purpose, Real-time, No-reference Video-Quality Assessment Tool

The heart of TekMOS is a machine learning algorithm. TekMOS decodes each frame to base-band and extracts from the luminance (Y') component a set of features related to perceived video quality characteristics. Artifacts such as blur, noise and over-compression tiling create different variations in these features from a high-quality scene.

Introduction

What is no-reference video-quality and is it better or worse than full-reference?

Numerous Full-Reference (FR) video quality tools/algorithms are available today. These often provide useful information about the image degradation relative to a reference. They are difference engines providing a Differential Mean Opinion Score (DMOS) correlated with human opinions of the subjective level of seeable degradation. In other words, they measure the fidelity of processing a video image sequence for distribution or storage. Both the original and processed video images are needed and must be aligned spatially and temporally for this differential comparison, limiting the utility of FR tools. More importantly, FR tools do not measure the quality of the reference image sequence itself since they are calibrated to human observer opinions, comparing only visible differences. In other words, if the reference image has a loss of detail or is of low contrast, the processed image may get a perfect rating on the DMOS scale if there is no visible difference.

The aim of any No-Reference (NR) video quality assessment tool is to somehow predict and score the video image quality, without the aid of any undistorted reference image sequence, as closely as possible to the quality score of an average human observer given that he or she also has no previous version of the video for comparison. In other words, the image quality is judged solely on the merits of that image alone with the goal of objectively determining what may be quality degradation separate from artistic intent. Clearly, this has a big utility advantage over FR tools since, without the need of any original reference image, no alignment is necessary thereby allowing quality determination after format conversion, or up/down sampling.

One method to provide a NR score of base-band video image quality is to allow observers to view images, one at a time, and rate them independently on a category basis. A useful scale for the observers to use, in absence of a reference for comparison, is the five-level Absolute Category Rating (ACR) [1] scale shown in Figure 1.

Observations from a group of observers over numerous frame captures with various ranges of image quality or distortion level is then pooled or averaged to find a Mean Opinion Score (MOS) for each image. Care is taken to present images in random order and under consistent viewing conditions such as image size, viewing distance, etc. Observers need not be experts but should be representative of the target video customer perception of video quality. However, people are different, focusing attention on different areas of a large-screen image with differing opinions of what is a distortion or simply just artistic intent. Therefore, a robust statistical method to compile ACR ratings into the overall MOS value for each image should be used to avoid certain observer biases and radical outlier opinions.

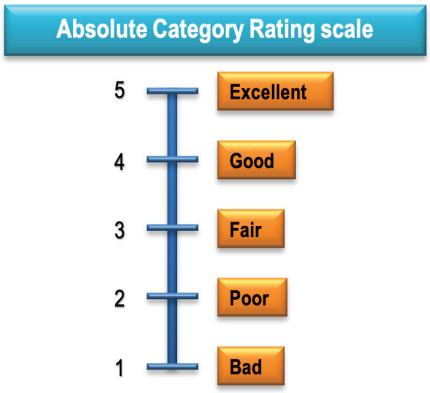

Figure 1. Absolute Category Rating scale.

TekMOS & Machine Learning

How does *machine learning* apply to video quality assessment?

With the advent of supervised Machine-Learning (ML) algorithms capable of processing large empirical data sets, it is now possible to exploit a large database of human opinion MOS scores over a range of distortion classes training a ML algorithm to learn how to form a similar opinion to those human observers.

Once properly trained, the ML algorithm provides a human educated real-time MOS score on new, never seen before images, without comparison to any previous, undistorted reference image. A key advantage of ML is the algorithms are easily retrained to different human MOS rating opinions on future distortion types or video compression artifacts. TekMOS is one of these methods targeted for both streaming and file-based video quality assessment.

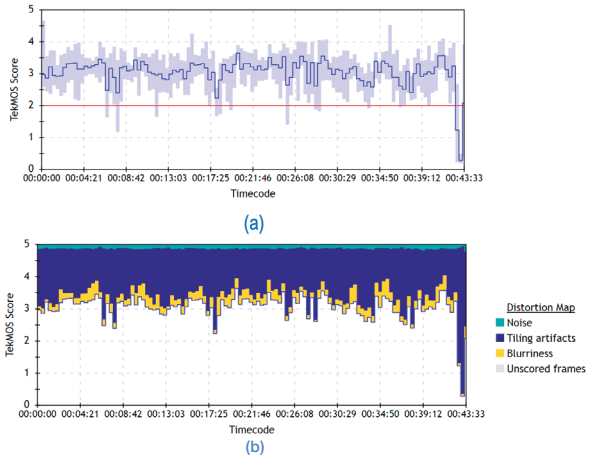

Figure 2. Example graph from file-based video quality assessment. (a) TekMOS scores for entire video stream shown in one graph and, (b) shows the probability that the resulting score is caused by one of the distortion classes.

The resulting MOS score ranks what the perceived quality the average viewer will see. However, in cases of low scores, why the quality is low may be just as important. TekMOS' s ML algorithm can answer this question too. Given a sorted database of several distortion classes such as noise, video compression artifacts and blur, the TekMOS ML algorithm is also trained as a classifier, to estimate the proportion of each distortion type (class) to aid in determining why the quality score is low. This way a graph of the composition of each distortion is available, aligned with the MOS score graph, illustrating the dominant distortion pulling down the MOS score in various regions of the image sequence (Figure 2).

Many times, much of the image area is intentionally blurred since the artistic intent is a reduced depth-of-field to focus viewer attention on a smaller segment of the image. TekMOS includes dynamic window tracking to find and analyze only that segment of the image to determine the dominant distortion and quality score.

Figure 3. Differing MOS scores and confidence of detected distortion between the two images.

The same dynamic windowing of each image frame is used during the ML model training to better mimic the human observer's training score. Figure 3 illustrates two examples of different adjustments of camera focus. TekMOS locates a region of attention (shown in rectangle) and scores only that region. Note the differing MOS scores and confidence of detected distortion between the two images.

There are several key advantages to using TekMOS over a group of human "golden-eye" observers assessing video quality.

- Human golden-eye observers can be trained to score video quality. However, this is very time-consuming and limited to only spot-checking since it is not possible to get real-time video MOS score averages. TekMOS runs in real-time and provides detailed data on each video sequence.
- Human golden-eye observers have biases that vary on different content and different times of day. Typical expert viewers will score the same image differently depending on what images preceded it or even time of day. Correlation across multiple viewers and even self-to-self correlation after waiting a day or two between scoring the same, unaltered image is seldom better than 80 to 90%. On the contrary, TekMOS scores an image exactly the same each time it sees it, if no changes have occurred.
- Often what is desired is a determination of how much a video image degrades over various downstream compression and format conversions. Human golden-eye observers can often see and agree on significant degradation but may score an image only slightly degraded or perhaps even better than the original.

  Even if you do not agree with the NR TekMOS score of a particular frame, that score almost always gets worse after added noise, compression artifacts or loss-of-detail. Therefore, downstream processing degradation can be determined by differentially comparing the TekMOS scores at various points in the distribution chain when distribution fidelity is the primary concern.

TekMOS Algorithm

How does it work and what aspect of machine learning does it use?

The heart of TekMOS is a machine learning algorithm. TekMOS decodes each frame to base-band and extracts from the luminance (Y') component a set of features related to perceived video quality characteristics. Artifacts such as blur, noise and over-compression tiling create different variations in these features from a high-quality scene. The resulting features form a vector of scalar values that are used, along with the associated subjective MOS score target and the dominant distortion class, for training a machine learning system (Figure 4). Currently, TekMOS does not analyze the chroma components, frame-rate temporal distortions or HDR/WCG aspects as part of the overall MOS score.

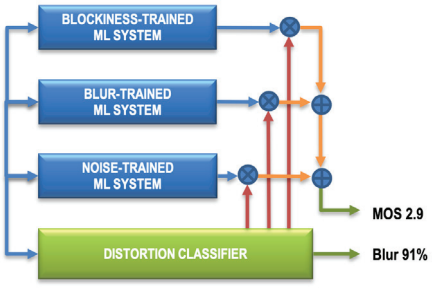

Figure 4. TekMOS ML algorithm feature vector combined into regression and classification results for pooling MOS scores and dominant distortion class.

After collecting and collating several thousand subjectively scored images of varying levels and types of distortion, the resulting database is partitioned into two sets. One is set aside as a Test set for TekMOS performance validation and not used for training. The second, larger, Training set is processed using an iterative cross-validation method to create best-fit regression (MOS score) and classification (distortion percentages) models.

After finding the best performance model fitting the training set, the performance validation test set is used to validate the model's MOS regression and distortion classification accuracy.

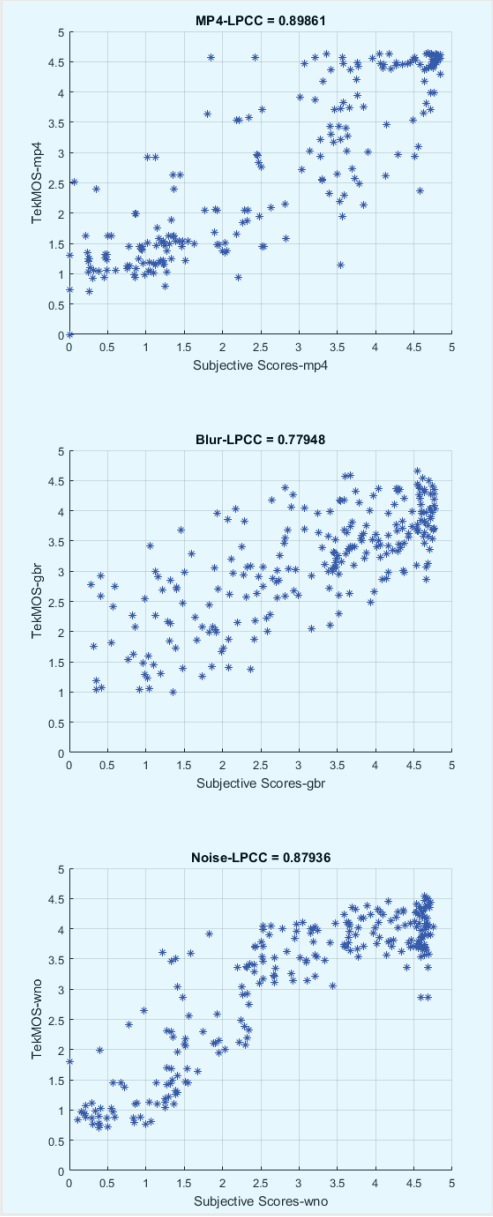

Figure 5. Scatter plots documenting TekMOS' s no-reference MOS score accuracy on a subjectively scored set

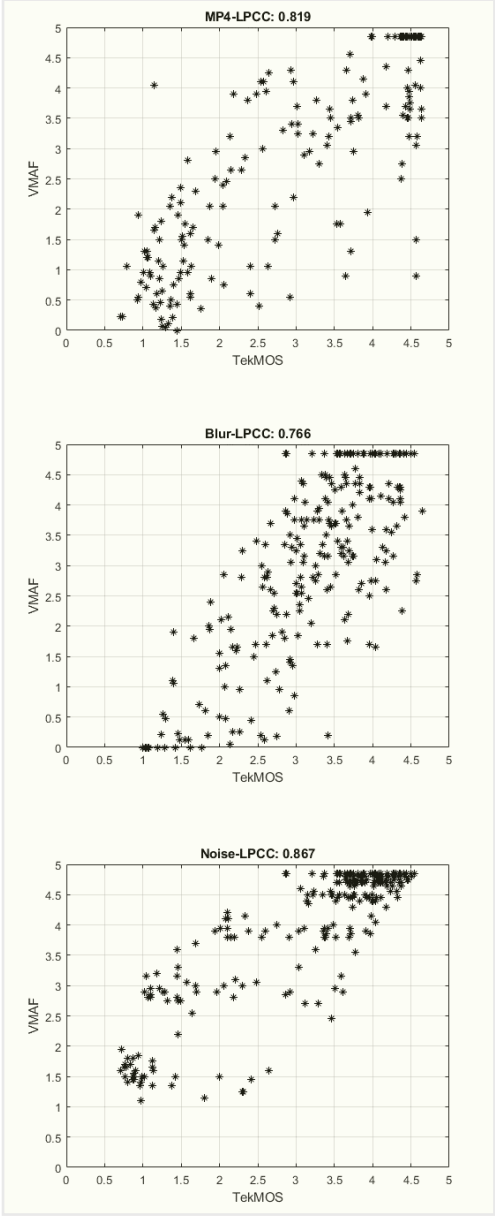

Figure 6. Netflix's full-reference VMAF scores Vs TekMOS' s no-reference scores.

Accuracy of Evaluations

How accurate is TekMOS?

Figure 5 shows scatter plots documenting TekMOS' s NR MOS score accuracy on a subjectively scored set of recently obtained 1080p test frames, each with various types and degrees of common distortions (random noise, image focus blur and H.264 compression artifacts). The horizontal axis plots the averaged subjective (human) MOS scores, using the ACR scale and viewing the images in random order each time. The vertical axis plots the scores obtained from TekMOS algorithm, trained and validated on a database of 720p and 1080p images. Recall that correlation across multiple viewers and even self-to-self correlation is seldom better than 80 to 90%. The Linear Pearson Correlation Coefficient (LPCC), computed and shown for each plot, is about 80%, clearly indicating the accuracy of NR TekMOS score.

Another possible comparison could be between FR and NR scores, since both approaches claim to measure scores closer to average human observer. However, one must refrain from drawing deeper conclusions from such a comparison because these are fundamentally different measurements.

The scores obtained from Video Multimethod Assessment Fusion (VMAF), a popular FR subjective video quality metric developed by Netflix, are compared with TekMOS' s NR scores. Figure 6 plots VMAF scores against TekMOS scores for each distortion class. The LPCC, computed and shown for each plot, is more than 75%. Notice that VMAF clips down scores of several reference images at ~4.85 on the MOS scale. It must also be noted that not all reference images are perfect and the subjective MOS scores for reference images range between ~3.5 and ~4.5.

Other NR PQ methods

Where does TekMOS fit with regard to other PQ methods?

Various NR methods exist to assess video image quality from a measure of certain, specific distortions. For example, a compressed video file has information about the level of compression, bit-rate and pixel-quantization that may be informative of the decoded image quality in some cases without any need to decode the video to baseband. However, if the quality before encoding is poor, these methods will falsely interpret a poor-quality video sequence to be of high-quality.

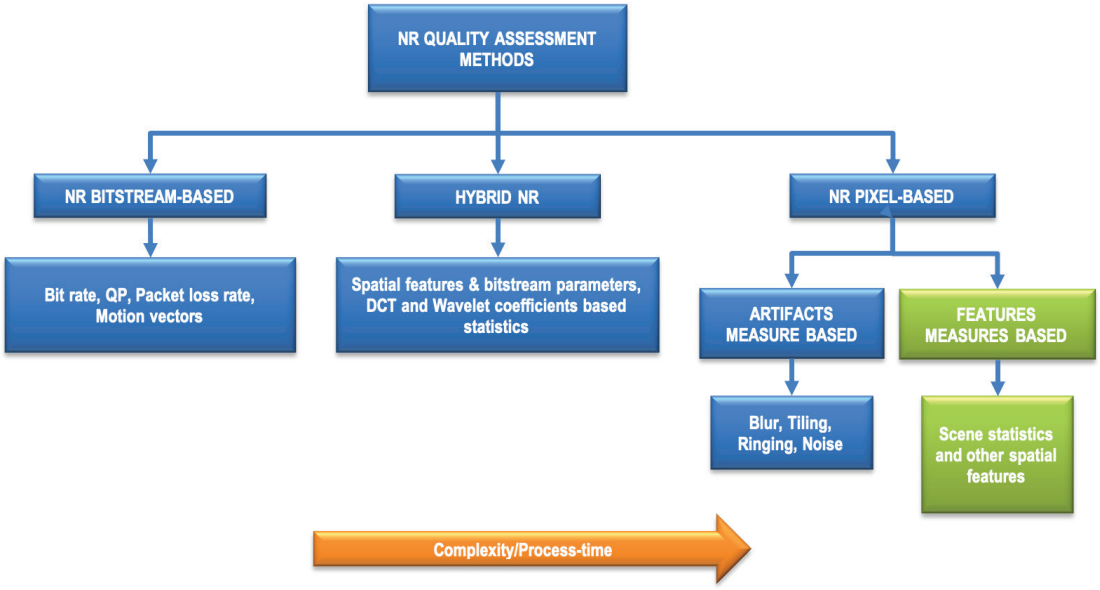

Figure 7. Hierarchy of NR methods vs. complexity and performance.

Other methods may decode the compressed video to a baseband image and process with a range of filters and other image processing to determine a level of blur, blockiness or random noise. These metrics then need to be scaled, weighted and pooled to provide an estimate of the image quality. These methods may miss distortions for which they are not designed to detect and are complicated to update to new compression artifacts and viewer expectations. Figure 7 (previous page) shows the hierarchy of NR methods vs. complexity and performance.

New features measure-based methods using ML such as TekMOS are positioned at the far right in the figure. They may, or may not, be more computationally complex and therefore slower than other NR methods shown to the left, but this is only with regard to the feature computations. Given the features computed for each frame, a trained ML MOS scoring and distortion classification algorithm is typically very fast acting much like a look-up table.

References

[1]. P.910: Subjective video quality assessment methods for multimedia applications, ITU-T Recommendation, April 2008.

www.telestream.net | info@telestream.net | tel +1 530 470 1300