LOW BANDWIDTH REDUCED REFERENCE VIDEO QUALITY MONITORING SYSTEM

Stephen Wolf and Margaret H. Pinson

Institute for Telecommunication Sciences (ITS)
National Telecommunications and Information Administration (NTIA)
325 Broadway, Boulder, CO 80305

ABSTRACT

This paper presents a new reduced reference (RR) video quality monitoring system that utilizes less than 10 kbits/s of reference information from the source video stream. This new video quality monitoring system utilizes feature extraction techniques similar to those found in the NTIA General Video Quality Model (VQM) that was recently standardized by the American National Standards Institute (ANSI) and the International Telecommunication Union (ITU). Objective to subjective correlation results are presented for 18 subjectively rated data sets that include more than 2500 video clips from a wide range of video scenes and systems. The method is being implemented in a new end-to-end video quality monitoring tool that utilizes the Internet to communicate the low bandwidth features between the source and destination ends.

1. INTRODUCTION

To be accurate, digital video quality measurements must measure perceived "picture quality" of the actual video being sent by the end-user (i.e., in-service measurement). This is because the perceived quality of a digital video system is variable and depends upon dynamic characteristics of both the input video scene and the digital transmission channel. A full reference quality measurement system (i.e., a system that has full access to the original source video stream), cannot be used to perform in-service monitoring since the original source video is not available at the destination end. However, a reduced reference (RR) quality measurement system can provide an effective method for performing perception-based in-service measurements. RR systems operate by extracting low bandwidth features from the source video and transmitting these source features to the destination location, where they are used in conjunction with the destination video stream to perform a perception based quality measurement.

This paper presents a new low bandwidth RR video quality monitoring system that utilizes techniques similar to those of the NTIA General Video Quality Model (VQM) [1, 2]. The NTIA General VQM was one of the top performing video quality measurement systems in the recent Video Quality Experts Group (VQEG) Full Reference Television (FRTV) phase 2 tests [3] and as a result has been standardized by both ANSI [4] and the ITU [5, 6]. While the NTIA General VQM was submitted to the VQEG FRTV tests, this VQM is in fact a high bandwidth RR system. NTIA chose to submit a RR system to the full reference VQEG tests, since research with the best NTIA video quality metrics demonstrated that there was little to be gained by using more than several Mbits/s of reference information [7], which is the approximate bit-rate of the NTIA General VQM.

This paper presents an overview of the new RR system that utilizes less than 10 kbits/s of reference information while still achieving high correlation to subjective quality. Results are presented for 18 subjectively rated data sets that include more than 2500 video clips from a wide range of video scenes and systems. The method is being implemented in a new end-to-end video quality monitoring tool that utilizes the Internet to communicate the low bandwidth features between the source and destination ends.

2. OVERVIEW OF OBJECTIVE MODEL

This section will present an overview of the RR model, including (1) the low bandwidth features that are extracted from the source and destination video streams, (2) the parameters that result from comparing like source and destination feature streams, and (3) the VQM calculation that combines the various parameters, each of which measures a different aspect of video quality. Due to the brevity of this paper, extensive references will be made to prior publications for technical details.

2.1. Features

The 10 kbits/s RR model uses the same f_{SI13}, f_{HV13}, and f_{COHER_COLOR} features that are used by the NTIA General VQM. These features are described in detail in sections

4.2.2 and 4.3 of [1]. Each feature is extracted from a spatial-temporal (S-T) region size of 32 vertical lines by 32 horizontal pixels by 1 second of time (i.e., 32 x 32 x 1s) whereas the NTIA General VQM used S-T region sizes of 8 x 8 x 0.2s for the f_{SI13} and f_{HV13} features and 8 x 8 x 1 frame for the f_{COHER_COLOR} feature. The f_{SI13} and f_{HV13} features measure the amount and angular distribution of spatial gradients in S-T sub-regions of the luminance (Y) image while the f_{COHER_COLOR} feature provides a 2-dimensional vector measurement of the amount of blue and red chrominance information (C_B, C_R) in each S-T region. For video at 30 frames per second (fps), these features achieve a compression ratio of more than 30,000 to 1.

Quantization to 9 bits of accuracy is sufficient for these features, provided one uses a non-linear quantizer design where the quantizer error is proportional to the magnitude of the signal being quantized. Very low values may be uniformly quantized to some cutoff value, below which there is no useful quality assessment information. Such a quantizer design minimizes the error in the corresponding parameter calculation, which is normally based on an error ratio or log ratio of the destination and source feature streams (see section 2.2 below). Figure 1 provides a plot of the 9-bit non-linear quantizer used for the f_{SI13} source feature.

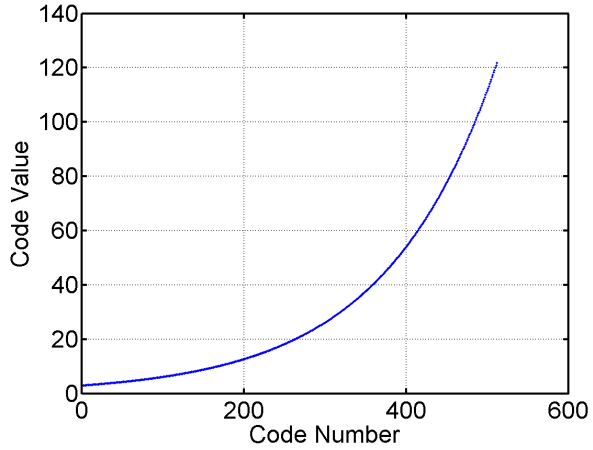

Figure 1. Non-linear 9-bit quantizer for the f_{SI13} feature.

Powerful estimates of perceived video quality can be obtained from the above feature set. However, since the S-T regions from which the above feature statistics are extracted span many video frames (1 second of video frames), they tend to be insensitive to brief temporal disturbances in the picture. Such disturbances can result from noise or digital transmission errors; and, while brief in nature, they can have a significant impact on the perceived picture quality. Thus, a temporal-based RR feature was developed to quantify the perceptual effects of temporal disturbances. This feature measures the absolute temporal information (ATI), or motion, in all three image planes (Y, C_B, C_R), and is computed as:

$$f_{ATI} = rms\left\{ YC_BC_R(t) - YC_BC_R(t-0.2s) \right\}.$$

The entire three dimensional image at time t-0.2s is subtracted from the three dimensional image at time t and the root mean square error (*rms*) of the result is used as a measure of ATI. This feature is sensitive to temporal disturbances in all three image planes: the luminance image (Y), and the blue and red color difference images (C_B and C_R, respectively). For 30 fps video, 0.2s is 6 video frames while for 25 fps video, 0.2s is 5 video frames. Subtracting images 0.2s apart makes the feature insensitive to real time 30 fps and 25 fps video systems that have frame update rates of at least 5 fps. The quality aspects of these low frame rate video systems, common in multimedia applications, are sufficiently captured by the f_{SI13}, f_{HV13}, and f_{COHER_COLOR} features. The 0.2s spacing is also more closely matched to the peak temporal response of the human visual system than differencing two images that are one frame apart in time.

Figure 2 provides an example plot of the f_{ATI} feature for a source (solid blue) and destination (dashed red) video scene from a digital video system with transient burst errors in the digital transmission channel. Transient errors in the destination picture create spikes in the f_{ATI} feature. The bandwidth required to transmit the f_{ATI} feature is extremely low (even using 16 bits/sample) since it requires only 30 samples per second for 30 fps video. The feature can also be used to perform time alignment of the source and destination video streams. Other types of additive noise in the destination video, such as might be generated by an analog video system, will appear as a positive DC shift in the time history of the destination feature stream with respect to the source feature stream. Video coding systems that eliminate noise will cause a negative DC shift.

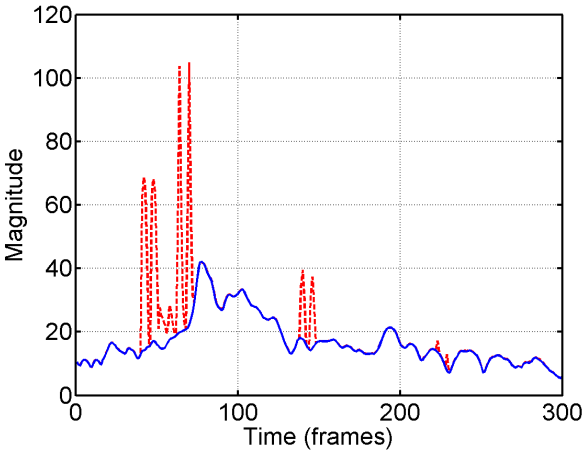

Figure 2. Example time history of f_{ATI} feature.

2.2. Parameters

Several steps are involved in the calculation of parameters that track the various perceptual aspects of video quality. The steps may involve (1) applying a perceptual threshold to the extracted features from each S-T sub-region, (2) calculating an error function between destination features and corresponding source features, and (3) pooling the resultant error over space and time. The reader is directed to section 5 of [1] for a detailed description of these techniques and their accompanying mathematical notation.

This paper will concentrate on new methods in this area that have been found to improve the objective to subjective correlation beyond what is achievable from the methods found in [1]. It is worth noting that no improvements have been found for the error functions in step 2 (given in section 5.2.1 of [1]). The two error functions that consistently produce the best results are a logarithmic ratio [log10 (*f*_destination / *f*_source)] and an error ratio [(f_destination – f_source) / f_source]. As described in section 5.2 of [1], these errors must be separated into gains and losses, since humans respond differently to additive (e.g., blocking) and subtractive (e.g., blurring) impairments. Applying a lower perceptual threshold to the features (step 1) before application of these two error functions prevents division by zero.

One new error pooling method is called macro-block (MB) error pooling. MB error pooling groups a contiguous number of S-T sub-regions and applies an error pooling function to this set. For instance, the function denoted as "MB(3,3,2)max" will perform a max function over parameter values from each group of 18 S-T sub-regions that are stacked 3 vertical by 3 horizontal by 2 temporal. For the 32 x 32 x 1s S-T regions of the f_{SI13}, f_{HV13}, and f_{COHER_COLOR} features described above, each MB(3,3,2) region would encompass a portion of the video stream that spans 96 vertical lines by 96 horizontal pixels by 2 seconds of time. MB error pooling has been found to be useful in tracking the perceptual impact of impairments that are localized in space and time. Such localized impairments often dominate the quality decision process.

A second error pooling method is a generalized *Minkowski*(*P*,*R*) summation, defined as:

$$Minkowski(P,R) = \sqrt[R]{\frac{1}{N}\sum_{i=1}^{N}|v_i|^P}\ .$$

Here v_i represents the parameter values that are included in the summation. This summation might, for instance, include all parameter values at a given instance in time (spatial pooling), or may be applied to the macro-blocks described above. The Minkowski summation where the power *P* is equal to the root *R* has been used by many developers of video quality metrics for error pooling. The generalized Minkowski summation, where $P \neq R$, provides additional flexibility for linearizing the response of individual parameters to changes in perceived quality. This is a necessary step before combining multiple parameters into a single estimate of perceived video quality (see section 2.3 below).

Before extracting a transient error parameter from the f_{ATI} feature streams shown in Figure 2, it is advantageous to increase the width of the motion spikes (red spikes in Figure 2). The reason is that short motion spikes from transient errors do not adequately represent the perceptual impact of these types of errors. One method for increasing the width of the motion spikes is to apply a maximum filter to both the source and destination feature streams before calculation of the error function between the two waveforms. We used a 7-point wide maximum filter that produces an output sample at each frame that is the maximum of itself and the 3 nearest neighbors on each side (i.e., earlier and later time samples).

2.3. 10 kbits/s VQM Calculation

Similar to the NTIA General VQM, the 10 kbits/s VQM calculation linearly combines 2 parameters from the f_{HV13} feature (loss and gain), 2 parameters from the f_{SI13} feature (loss and gain), and 2 parameters from the f_{COHER_COLOR} feature. The one noise parameter in the NTIA General model has been replaced with 2 parameters based on the low bandwidth f_{ATI} feature described in this paper; one parameter measures added noise and the other parameter measures temporal disturbances in the destination picture.

For 30 fps video in the 525-line format, a 384-line x 672-pixel sub-region centered in the ITU-R Recommendation BT.601 video frame (i.e., 486 line x 720 pixel) produces a VQM bit rate before any coding (e.g., Huffman) that is less than 10 kbits/s. Since Internet connections are ubiquitously available at this bit rate, the new 10 kbits/s VQM can be used to monitor the end-to-end quality of video transmission between nearly any source and destination location.

3. OBJECTIVE TO SUBJECTIVE RESULTS

The techniques presented in [8] were used together with the NTIA General VQM parameters to map 18 subjective data sets onto a (0, 1) common subjective quality scale, where "0" represents no perceived impairment and "1" represents maximum impairment. With the subjective mapping procedure used, occasional excursions less than 0 (quality improvements) and more than 1 are allowed. The 18 subjectively rated video data sets contained 2651 video clips that spanned an extremely wide range of scenes and

video systems. The resulting subjective data set was used to determine the optimal linear combination of the 8 video quality parameters in the 10 kbits/s VQM (section 2.3). Figure 3 gives the scatter plot for the subjective data versus the 10 kbits/s VQM where each data set is shown in a different color.

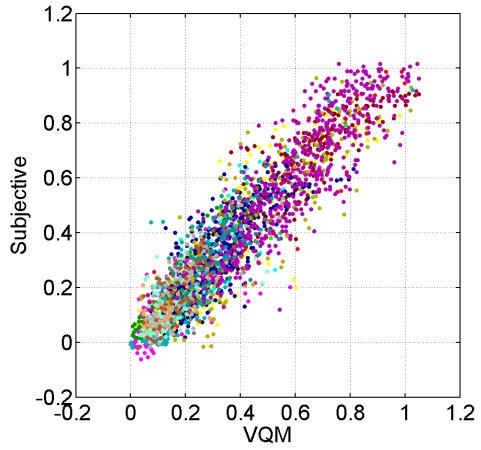

Figure 3. Subjective data versus 10 kbits/s VQM.

4. MONITORING SYSTEM AND CONCLUSIONS

The NTIA General VQM, as well as the new 10 kbits/s VQM, have been implemented in a new PC-based software system that has been specifically designed to perform continuous in-service monitoring of video quality. Figure 4 gives a screen snapshot of the running system.

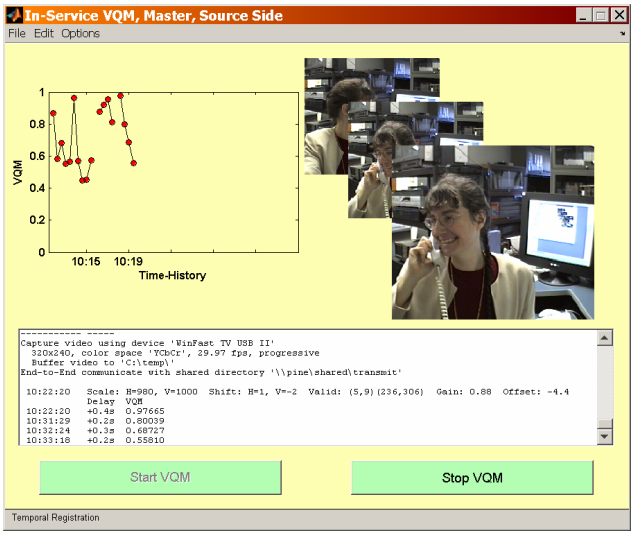

Figure 4. New in-service video quality monitoring system.

The video quality monitoring system runs on two PCs and communicates the RR features via an Internet connection. The software supports frame capture devices, including newer USB 2.0 frame capture devices that attach to laptops. The duty cycle of the continuous quality monitoring (i.e., percent of video stream from which video quality measurements are performed) depends upon the CPU speed of the host machine. Calibration of the system (e.g., gain, level offset, spatial scaling/registration, and temporal registration) can be performed at user-defined time intervals.

The new 10 kbits/s VQM algorithm, combined with the new in-service monitoring system, gives end-users and industry a powerful tool for assessing video quality.

5. REFERENCES

[1] S. Wolf and M. Pinson, "Video Quality Measurement Techniques," NTIA Report 02-392, June, 2002. Available at http://www.its.bldrdoc.gov/n3/video/documents.htm.

[2] M. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting*, v. 50, n. 3, pp. 312-322, September, 2004. Available at http://www.its.bldrdoc.gov/n3/video/documents.htm.

[3] "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II," Video Quality Experts Group, August, 2003. Available at ftp://ftp.its.bldrdoc.gov/dist/ituvidq/frtv2_final_report/.

[4] ANSI T1.801-2003, "Digital Transport of One-Way Video Signals – Parameters for Objective Performance Assessment," American National Standards Institute, approved September, 2003.

[5] ITU-T J.144R, "Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference," Telecommunication Standardization Sector, approved March, 2004.

[6] ITU-R BT.1683, "Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference," Radiocommunication Sector, approved June, 2004.

[7] S. Wolf and M. H. Pinson, "The Relationship Between Performance and Spatial-Temporal Region Size for Reduced-Reference, In-Service Video Quality Monitoring Systems," SCI / ISAS 2001 (Systematics, Cybernetics, and Informatics / Information Systems Analysis and Synthesis), July, 2001. Available at http://www.its.bldrdoc.gov/n3/video/documents.htm.

[8] M. Pinson and S. Wolf, "An Objective Method for Combining Multiple Subjective Data Sets," SPIE Video Communications and Image Processing Conference, Lugano, Switzerland, July, 2003. Available at http://www.its.bldrdoc.gov/n3/video/documents.htm.