Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system

Stephen Wolf, Margaret H. Pinson

Institute for Telecommunication Sciences

National Telecommunications and Information Administration

Boulder, CO 80303

ABSTRACT

Many organizations have focused on developing digital video quality metrics which produce results that accurately emulate subjective responses. However, to be widely applicable a metric must also work over a wide range of quality, and be useful for in-service quality monitoring. The Institute for Telecommunication Sciences (ITS) has developed spatial-temporal distortion metrics that meet all of these requirements. These objective metrics are described in detail and have a number of interesting properties, including utilization of 1) spatial activity filters which emphasize long edges on the order of 1/5 degree while simultaneously performing large amounts of noise suppression, 2) the angular direction of the spatial gradient, 3) spatial-temporal compression factors of at least 384:1 (spatial compression of at least 64:1 and temporal compression of at least 6:1, and 4) simple perceptibility thresholds and spatial-temporal masking functions. Results are presented that compare the objective metric values with mean opinion scores from a wide range of subjective data bases spanning many different scenes, systems, bit-rates, and applications.

Keywords: video quality metrics, subjective, objective, correlation, spatial, temporal, in-service, compression

1. INTRODUCTION

To be widely applicable, a digital video quality metric must:

1. Produce results that accurately emulate subjective responses.
2. Work over the full range of quality, from very low bit rate video teleconferencing systems to very high bit rate studio and broadcast systems.
3. Be computationally efficient, so that it may be implemented on common PC platforms.
4. Be bandwidth efficient, compressing quality information into the smallest possible bandwidth, thereby making the metric useful for end-to-end in-service quality monitoring.

While many organizations have focused exclusively on (1) for a narrowly defined application or video compression technology, we at the Institute for Telecommunication Sciences (ITS) have focused our research efforts on developing video quality metrics with all of the above attributes. We have recently obtained several new subjective data sets and the means to process extensive amounts of digital video. This has allowed us to examine a total of seven independent subjective data sets that spanned an extremely wide range of digital video systems and test scenes. In this paper, we present detailed descriptions of computationally efficient spatial-temporal distortion metrics that have a high degree of correlation with subjective ratings from these seven independent subjective experiments.

Item (4) above is often neglected in the development of objective digital video quality metrics. This is unfortunate since the video quality of a modern digital video system is variable and depends upon the dynamic characteristics of the input video (e.g., spatial detail, motion) and the digital transmission system (e.g., bit-rate, error-rate). Thus, out-of-service testing made at different times or using different scenes than what is actually used in-service cannot quantify video quality as it is truly perceived by the end-user. This perceived video quality must be measured in-service using the actual video being sent by the users of the digital video system. We have shown that accurate perception-based in-service video quality measurements can be made using the technique shown in Figure 1.[1-10] Rather than relying on input and output pixel comparisons (which require full bandwidth reference information for in-service measurements, or *a priori* knowledge of the test scenes for out-of-service

measurements), the technique shown in the figure uses reduced reference information in the form of *features* that are extracted from processed spatial-temporal (S-T) regions of the input and output video streams. A feature is defined here as a quantity of information that is associated with a specific S-T region of the video sequence. Examples of features are summary statistics (e.g., mean, standard deviation) calculated using all the image pixels within a processed S-T region. The reduced reference information is compressed by many orders of magnitude versus the ITU-R Recommendation BT.601[11] video stream (referred to as Rec. 601 later in this paper) and thus can be continuously transmitted using a readily available low-bandwidth ancillary data channel (e.g., modem, Internet, in-service data channel). Being low bandwidth, these features can also be easily archived and used as a historical record of video performance.

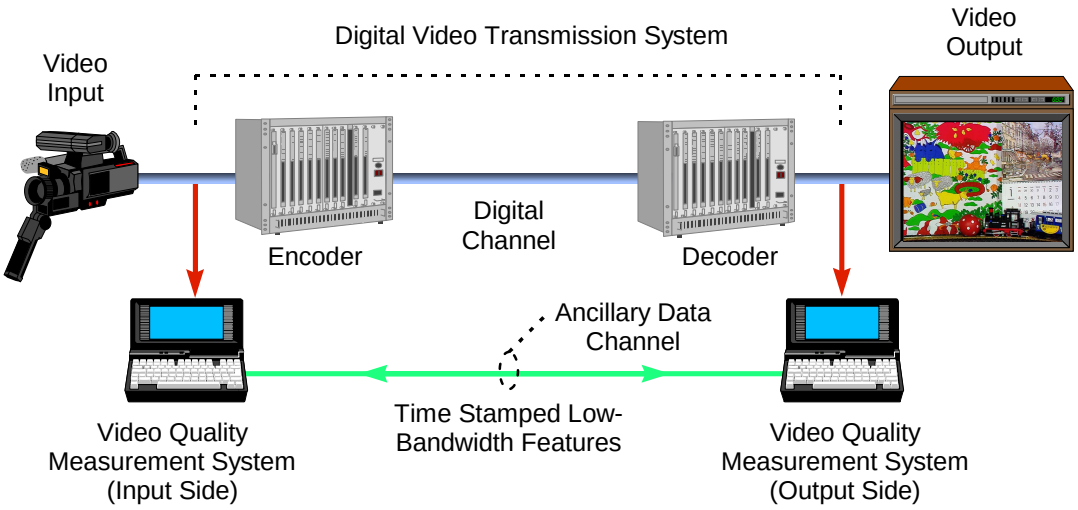

Figure 1. In-service video quality measurement system.

The focus of this paper is perceptual measurement of spatial distortions over time. Hence, we have categorized the metrics as spatial-temporal, since some temporal aspects have been included. Similar techniques to the ones presented here have also been applied with success to measuring "pure" temporal and chroma distortions.[10] However, for reasons of brevity, we have chosen to focus on perceptual measurement of spatial distortions over time since these types of distortions are the principal contributors to video quality (this will become apparent in sections 4.2 and 4.3).

2. DESCRIPTION OF METRICS

The goal of this section is to describe the spatial-temporal distortion metrics in sufficient detail so that they may be implemented by researchers. An overview of the algorithm to extract the metrics is given in Figure 2. The luminance component of the Rec. 601 input and output video streams (i.e., the Y signal in Rec. 601) are processed using horizontal and vertical edge enhancement filters. Next, these processed video streams are divided into S-T regions from which features, or summary statistics, are extracted that quantify the spatial activity as a function of angular orientation. Then, these features are clipped at the lower end to emulate perceptibility thresholds. Next, distortions in video quality due to gains and losses in the feature values are calculated for each S-T region by comparing their input and output values using functional relationships that emulate visual masking. These distortions are then pooled across space (spatial collapsing) and time (temporal collapsing) to produce quality metric values for a video clip which is nominally 5 to 10 seconds in duration.

The edge enhancement filters and the size of the S-T regions can be optimized based on their correlation with perceptual distortions. At viewer distances from 4 to 6 picture heights, optimal S-T region sizes achieve compression factors of at least 384:1 (spatial compression of at least 64:1 and temporal compression of at least 6:1) versus the uncompressed Rec. 601 video stream.

The sampled input and output video streams are assumed to have been calibrated before the processes described herein are performed. This calibration includes compensation for system gain and level offset, as well as spatial and temporal registration of the images. Fortunately, this calibration can also be performed using low bandwidth features extracted from the input and output video streams.[10]

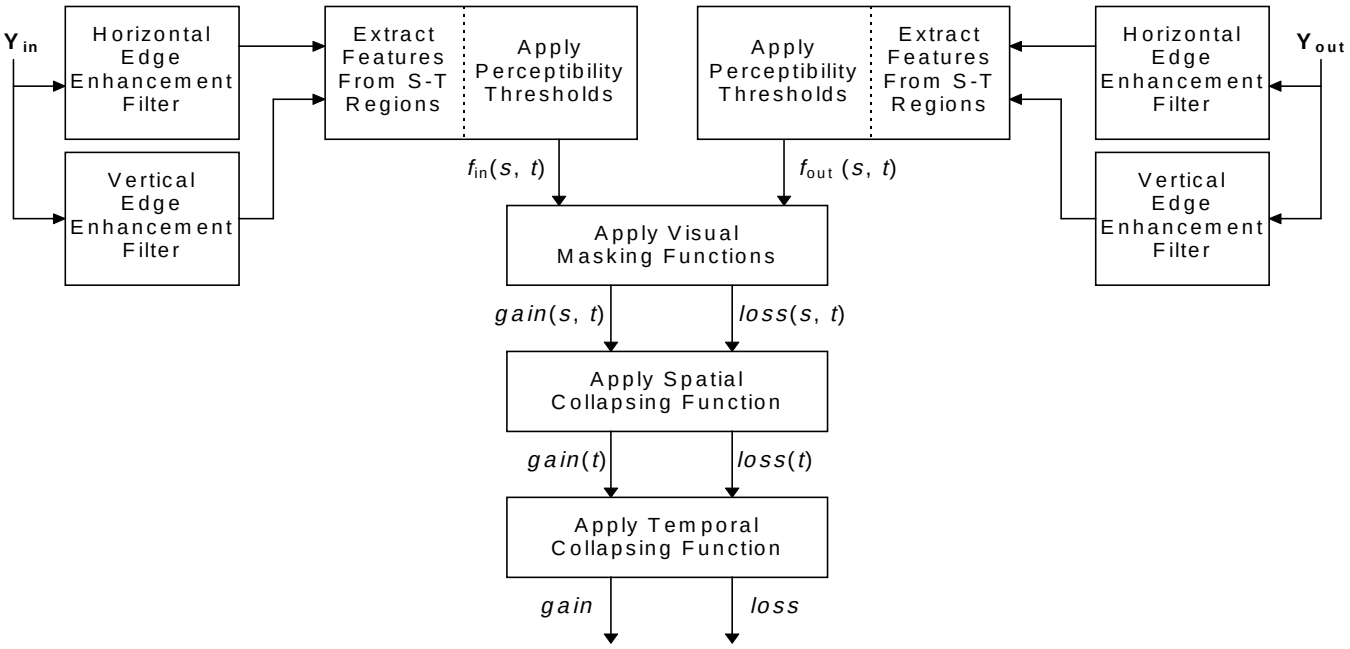

Figure 2. Overview of algorithm to extract video quality metrics.

2.1 Edge Enhancement Filter Size

The input and output video frames are first processed with horizontal and vertical edge enhancement filters that enhance edges while reducing noise. Prior papers have shown that the Sobel filters shown in Figure 3 work well for this step.[1-10] These two Sobel filters are applied separately, one to enhance horizontal pixel differences while smoothing vertically (left filter), and the other to enhance vertical pixel differences while smoothing horizontally (right filter). We once again examined the Sobel filter pair but also examined a number of other filter pairs that perform more edge enhancement and noise suppression. Figure 4 shows the general form for one such family of filter pairs. Only the horizontal bandpass/vertical lowpass filter is shown in Figure 4 (the vertical bandpass/horizontal lowpass filter can be generated by taking the transpose). These large edge enhancement filters were examined to see if higher amounts of edge enhancement and noise suppression could produce better spatial distortion metrics than the Sobel filter.

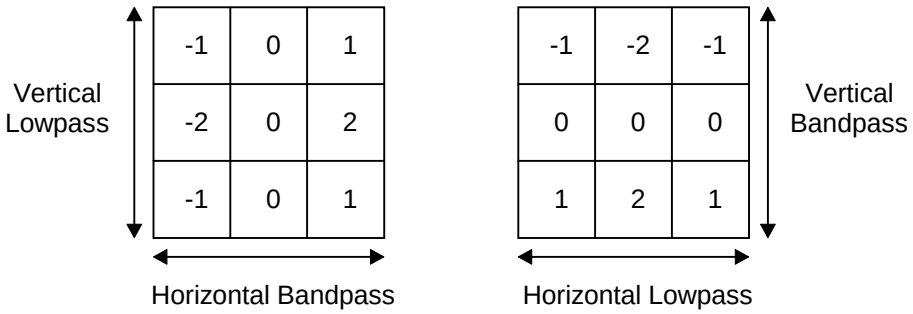

Figure 3. Sobel edge enhancement filters.

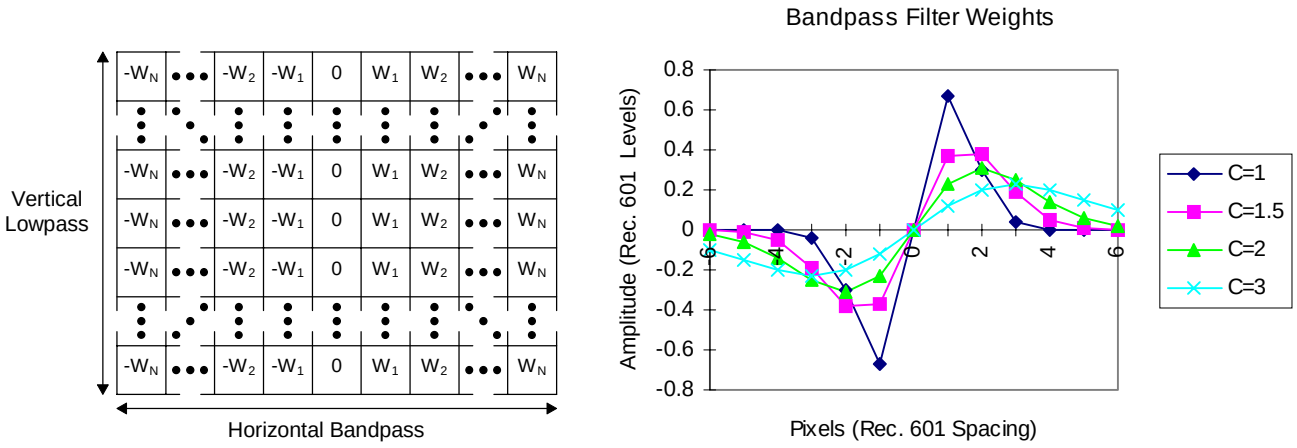

Figure 4. Large edge enhancement filters.

The weights for the bandpass filters shown in Figure 4 are given by

$$w_x = k * \left(\frac{x}{c}\right) * \exp\left\{-\left(\frac{1}{2}\right)\left(\frac{x}{c}\right)^2\right\},$$

where x is the horizontal pixel displacement from the center of the filter (0, 1, 2, …, N), c is a constant that sets the width of the bandpass filter (the bandpass filter shapes for c = 1, 1.5, 2, and 3 are plotted on the right hand side in Figure 4), and k is a normalization constant. For our tests, the filter was always square with an odd number of columns and rows so it could be centered over the pixel of interest. The normalization constant k was selected such that each filter would produce the same gain on a vertical edge as the left hand Sobel filter shown in Figure 3. Notice that the left hand Sobel filter shown in Figure 3 has a vertical amplitude taper (falling from 2 to 1 as one moves vertically off center) while the large edge enhancement filters do not have a vertical taper. We have found that non-tapered filters can be beneficial for quality assessment and they have the added advantage of being computationally efficient (i.e., one merely has to sum the pixels in a column and multiply once by the weight of that column).

2.2 S-T Region Size

The horizontal and vertical edge enhanced input and output video streams are each divided into localized S-T regions. Figure 5 gives an illustration of a S-T region that includes 8 horizontal pixels x 8 vertical lines x 6 video frames. Features are extracted from each S-T region by calculating summary statistics over the S-T region. As the number of pixels encompassed by the S-T region increases, the compression factor increases and hence the required ancillary data channel bandwidth shown in Figure 1 decreases.

The objective video quality metrics used in American National Standards Institute (ANSI) T1.801.03-1996[5] use extremely low bandwidths for the reference information since the S-T region size includes all of the valid pixels of a single video frame. For MPEG-2 video systems, we have found that a S-T region size of 8 horizontal pixels x 8 vertical lines x 1 video frame is near optimal for making fine-grain spatial distortion measurements.[10] However, as we will discuss in section 4.1, temporal widths on the order of 6 video frames appear to be more optimal for general purpose spatial distortion metrics that work well for *any* digital video system, from low bit-rate video teleconferencing systems to high bit-rate MPEG-2 systems.

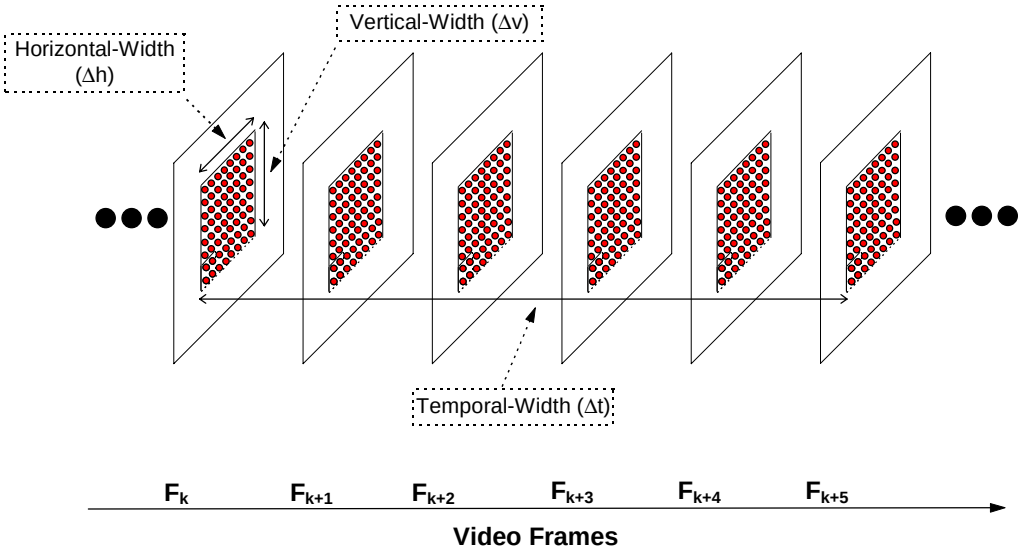

Figure 5. Illustration of a spatial-temporal (S-T) region for a video scene.

We have performed preliminary experiments which demonstrate that the compression factors for the features can be further increased by synchronized random sub-sampling in space and time. In this case, a subset of randomly selected S-T regions are selected for feature extraction. The additional feature compression resulting from synchronized random sub-sampling can be quite dramatic and is basically limited only by the desired repeatability of the quality measurement. Even stringent requirements on repeatability (such as 0.2%) make random sub-sampling a viable option for further increasing the compression factors of the extracted features, and thereby lowering the required ancillary data channel bandwidth shown in Figure 1.

2.3 Description of Features

This section describes the extraction of two spatial activity features from S-T regions of the edge enhanced input and output video streams from section 2.1. The filters shown in Figure 3 (left) or Figure 4 enhance spatial gradients in the horizontal (H) direction while the transposes of these filters enhance spatial gradients in the vertical (V) direction. The response at each pixel from the H and V filters can be plotted on a two dimensional diagram such as the one shown in Figure 6 with the H filter response forming the abscissa value and the V filter response forming the ordinate value. For a given image pixel located at row i, column j, and time t, the H and V filter responses will be denoted as $H(i, j, t)$ and $V(i, j, t)$, respectively. These responses can be converted into polar coordinates (R, θ) using the relationships

$$R(i,j,t) = \sqrt{H(i,j,t)^2 + V(i,j,t)^2} \text{ , and}$$

$$\theta(i,j,t) = \tan^{-1}\left[\frac{V(i,j,t)}{H(i,j,t)}\right].$$

The first feature, f_1, is computed simply as standard deviation (*stdev*) over the S-T region of the $R(i, j, t)$ samples, and then clipped at the perceptibility threshold of P (i.e., if the results of the *stdev* calculation falls below P, f_1 is set equal to P), namely

$$f_1 = \left\{ stdev\left[R(i,j,t)\right] \right\}\Big|_P : \; i,j,t \in \{\text{S} - \text{T Region}\}.$$

This feature is sensitive to changes in the overall amount of spatial activity within a given S-T region. For instance, localized blurring produces a reduction in the amount of spatial activity whereas noise produces an increase. For the results presented in section 4, the perceptibility threshold P for this feature was set equal to 12.

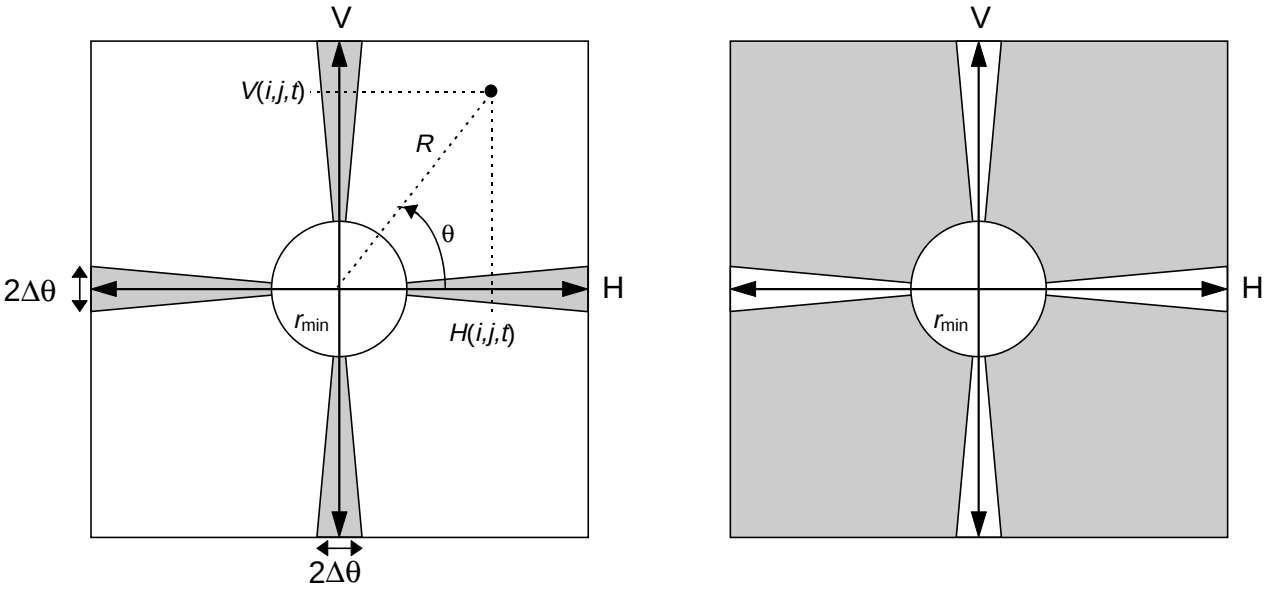

Figure 6. Division of horizontal (H) and vertical (V) spatial activity into *HV* (left) and $\overline{HV}$ (right) distributions.

The second feature, f_2, is sensitive to changes in the angular distribution, or orientation, of spatial activity. Complementary images are computed with the shaded spatial gradient distributions shown in Figure 6. The image with horizontal and vertical gradients, denoted as *HV*, contains the $R(i, j, t)$ pixels that are horizontal or vertical edges (pixels that are diagonal edges are zeroed). The image with the diagonal gradients, denoted as $\overline{HV}$ **,** contains the $R(i, j, t)$ pixels that are diagonal edges (pixels that are horizontal or vertical edges are zeroed). Gradient magnitudes $R(i, j, t)$ less than $r_{\min}$ are zeroed in both images to assure accurate θ computations. Pixels in *HV* and $\overline{HV}$ can be represented mathematically as

$$HV(i,j,t) = \begin{cases} R(i,j,t) \quad if \quad R(i,j,t) \geq r_{\min} \quad and \quad m\frac{\pi}{2} - \Delta\theta \; < \; \theta(i,j,t) \; < \; m\frac{\pi}{2} + \Delta\theta \quad (m = 0,1,2,3) \\ 0 \quad otherwise \end{cases}, \text{ and}$$

$$\overline{HV}(i,j,t) = \begin{cases} R(i,j,t) \quad if \quad R(i,j,t) \geq r_{\min} \quad and \quad m\frac{\pi}{2} + \Delta\theta \; \leq \; \theta(i,j,t) \; \leq \; (m+1)\frac{\pi}{2} - \Delta\theta \quad (m = 0,1,2,3) \\ 0 \quad otherwise \end{cases}, \text{ where}$$

$$i, j, t \in \{\text{S} - \text{T Region}\}.$$

Following the recommendations of ANSI T1.801.03, we used $r_{\min} = 20$ and $\Delta\theta = 0.05236$ radians for the computation of *HV* and $\overline{HV}$.[5] Feature f_2 for one S-T region is then given by the ratio of the mean of *HV* to the mean of $\overline{HV}$, where these resultant means are clipped at their perceptibility thresholds *P*, namely

$$f_2 = \left\{ mean\left[HV(i,j,t) \right] \right\}\Big|_P \Big/ \left\{ mean\left[\overline{HV}(i,j,t) \right] \right\}\Big|_P .$$

For the results presented in section 4, the perceptibility threshold *P* for the mean of HV and $\overline{HV}$ was set equal to 3. The f_2 feature is sensitive to changes in the angular distribution of spatial activity within a given S-T region. For example, if horizontal and vertical edges suffer more blurring than diagonal edges, f_2 of the output will be less than f_2 of the input. On the other hand, if erroneous horizontal or vertical edges are introduced, say in the form of blocking or tiling distortions, then f_2 of

the output will be greater than f_2 of the input. The f_2 feature thus provides a simple means to include variations in the sensitivity of the human visual system with respect to angular orientation.

2.4 Feature Comparison Functions (Quality Metrics)

The following provides a generic description of how distortions are calculated from the input and output feature streams given in section 2.3. For this discussion, an input feature stream will be denoted as $f_{in}(s, t)$ and the corresponding output feature stream will be denoted as $f_{out}(s, t)$, where s and t are indices that denote the spatial and temporal positions, respectively, of the S-T region within the calibrated input and output video streams. First, the perceptual impairment at each S-T region is calculated using a function that models visual masking. Next, impairments from S-T regions with the same time index are pooled using a spatial collapsing function. Finally, the results from the spatial collapsing function are pooled using a temporal collapsing function to produce an objective metric for the video clip, which is nominally 5 to 10 seconds in length.

2.4.1 Impairment Masking

Gain and loss must be examined separately, since they produce fundamentally different effects on quality perception (e.g., loss of spatial activity due to blurring and gain of spatial activity due to noise or blocking). Of the many comparison functions that we have evaluated, two have consistently produced the best correlation to subjective ratings. These comparison functions model the perceptibility of spatial or temporal impairments. For a given S-T region, gain and loss distortions are computed using:

$$gain(s,t) = pp\left\{\log_{10}\left[\frac{f_{out}(s,t)}{f_{in}(s,t)}\right]\right\}, \text{ and}$$

$$loss(s,t) = np\left\{\frac{f_{out}(s,t) - f_{in}(s,t)}{f_{in}(s,t)}\right\},$$

where *pp* is the positive part operator (i.e., negative values are replaced with zero), and *np* is the negative part operator (i.e., positive values are replaced with zero). These visual masking functions imply that impairment perception is inversely proportional to the amount of localized spatial or temporal activity in the input scene. In other words, spatial impairments become less visible as the spatial activity in the input scene is increased (i.e., spatial masking), and temporal impairments become less visible as the temporal activity in the input scene is increased (i.e., temporal masking). While the logarithmic and ratio comparison functions behave very similarly, we have found that the logarithmic function tends to be slightly more advantageous for gains while the ratio function tends to be slightly more advantageous for losses.

2.4.2 Spatial Collapsing

Extensive investigation has revealed that optimal spatial collapsing functions normally involve some form of worst case processing. This is because localized impairments tend to draw the focus of the viewer, making the worst part of the picture the predominant factor in the subjective quality decision. The spatial collapsing function we used is computed for each temporal index t as the average of the worst 5% of the measured distortions over the spatial index s. This produces a time history of the gain and loss samples, namely $gain(t)$ and $loss(t)$, which must then be temporally collapsed.

2.4.3 Temporal Collapsing

Viewers seem to use several temporal collapsing functions when subjectively rating video clips that are from 9 to 10 seconds in length. One of these functions is indicative of the average or best quality that is observed during the time period, while the other function is indicative of the worst transient quality that is observed (e.g., digital transmission errors may cause a 1 to 2 second disturbance in the output video). We have found that the long time mean (i.e., mean over 9 to 10 seconds) and the short time mean (i.e., mean over 1 to 2 seconds) capture most of the perceptual impact. We suspect that the short time mean by itself would capture most of the perceptual impact for continuously sampled subjective data (e.g., viewers produce continuously updated subjective scores by means of a slider that may be moved at will). However, since all of our subjective data consisted of a single score for a 9 to 10 second clip of video, it was not possible to test this hypothesis. For simplicity, the temporal collapsing function we used is computed as the mean of the $gain(t)$ and $loss(t)$ time samples over the entire 9 to 10 second time period. This temporal collapsing function may be sub-optimal for video clips with large variations in picture quality during this time period.

3. DESCRIPTION OF SUBJECTIVE DATA SETS

The seven subjective experiments were performed from 1992 to 1998. Data sets one to six were conducted in accordance with the most recent version of ITU-R Recommendation BT.500[12] that was available when the experiment was performed. Data set seven, being a personal computer (PC) video teleconferencing application, was conducted in accordance with ITU-T Recommendation P.910.[13] All of the data sets used scenes from 9 to 10 seconds in duration. For brevity, only a summary of each subjective experiment is given. The reader is directed to the accompanying references for more complete descriptions.

3.1 Data Set One:[1, 2]

A panel of 48 viewers rated a total of 132 video clips that were generated by random and deterministic pairing of 36 test scenes with 27 video systems. The 36 test scenes contained widely varying amounts of spatial and temporal information. The 27 video systems included digital video compression systems operating at bit-rates from 56 kbits/sec to 45 Mbits/sec with controlled error rates, NTSC encode/decode cycles, VHS and S-VHS record/play cycles, and VHF transmission. Viewers were shown the original version first, then the degraded version, and asked to rate the difference in perceived quality using the 5-point impairment scale (imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying).

3.2 Data Set Two:[14]

Viewer panels comprising a total of 30 viewers from three different laboratories rated 600 video clips that were generated by pairing 25 test scenes with 24 video systems. The 25 test scenes were standardized by ANSI T1.801.01-1995[15] and included scenes from 5 categories: (1) one person, mainly head and shoulders, (2) one person with graphics and/or more detail, (3) more than one person, (4) graphics with pointing, and (5) high object and/or camera motion. The 24 video systems included proprietary and standardized video teleconferencing systems operating at bit rates from 56 kbits/sec to 1.5 Mbits/sec with controlled error rates, one 45 Mbits/sec codec, and VHS record/play cycle. The subjective test procedure was the same as data set one.

3.3 Data Set Three:[16]

This data set was a subjective test evaluation of proponent MPEG-4 systems that utilized a panel of 15 expert viewers. We selected a subset of 164 video clips from the main data set. The subset was selected to span the full range of quality and included eight common intermediate format (CIF) resolution test scenes and 41 video systems from the basic compression tests. The eight video scenes included scenes from 2 categories: (1) low spatial detail and low amount of movement, and (2) medium spatial detail and low amount of movement or vice versa. The 41 video systems operated at bit rates from 10 kbits/sec to 112 kbits/sec. Viewers were shown only the degraded version and asked to rate the quality on a 11-point numerical scale, with 0 being the worst quality and 10 being the best.

3.4 Data Set Four:[17]

A panel of 32 viewers rated the difference in quality between input scenes with controlled amounts of added noise and the resultant MPEG-2 compression-processed output. The data set contains a total of 105 video clips that were generated by pairing seven test scenes at three different noise levels with five MPEG-2 video systems. The seven test scenes were chosen to span a range of spatial detail, motion, brightness, and contrast. The five MPEG-2 video systems operated at bit rates from 1.8 Mbits/sec to 13.9 Mbits/sec. Viewers were shown the input and processed output in randomized A/B ordering and asked to rate the quality of B using A as a reference. The experiment utilized a seven-point comparison scale (B much worse than A, B worse than A, B slightly worse than A, B the same as A, B slightly better than A, B better than A, B much better than A).

3.5 Data Set Five:[10]

A panel of 32 viewers rated a total of 112 video clips that were generated by pairing sub-groups of eight scenes each (total number of scenes in the test was 16) with 14 different video systems. The 16 test scenes spanned a wide range of spatial detail, motion, brightness, and contrast and included scene material from movies, sports, nature, and classical Rec. 601 test scenes. The 14 video systems included MPEG-2 systems operated at bit rates from 2 Mbits/sec to 36 Mbits/sec with controlled error rates, multi-generation MPEG-2, multi-generation ½ inch professional record/play cycles, VHS, and video teleconferencing systems operating at bit rates from 768 kbits/sec to 1.5 Mbits/sec. The subjective test procedure was the same as data set four.

3.6 Data Set Six:[10]

A panel of 32 viewers rated a total of 42 video clips that were generated by pairing sub-groups of six scenes each (total number of scenes in the test was 12) with seven different MPEG-2 systems. The 12 test scenes included sports material and classical Rec. 601 test scenes. The nine MPEG-2 systems operated at bit rates from 2 Mbits/sec to 8 Mbits/sec. The subjective test procedure was the same as data set four.

3.7 Data Set Seven:[9]

A panel of 18 viewers rated 48 video clips in a desktop video teleconferencing application. The 48 video clips were generated by pairing six scenes with eight different video systems. The six test scenes were selected from ANSI T1.801.01[15] and were the scenes *5row1, filter, smity2, vtc1nw, washdc,* and one scene that included portions of both *vtc2zm* and *vtc2mp*. The eight video systems included seven desktop video teleconferencing systems operating at bit rates from 128 kbits/sec to 1.5 Mbits/sec and one NTSC encode/decode cycle. Viewers were shown only the degraded version and asked to rate the quality on the absolute category rating scale (excellent, good, fair, poor, bad).

4. RESULTS

The metrics presented in this paper were evaluated on a subset of data set two described in section 3.2. The knowledge gained by examining this subset was used to develop edge enhancement filters and S-T region sizes that were then tested on all seven subjective data sets. In this section, we present objective to subjective correlation results for the individual spatial-temporal distortion metrics described in section 2 for one edge enhancement filter and one S-T region size. A combined spatial-temporal distortion metric is then proposed that is sensitive to both added and missing spatial activity.

4.1 Preliminary Training on a Subset of Data Set Two

A subset of 181 video clips from data set 2 was selected for preliminary training since it was felt that this data set contained the widest variation in perceived quality. These selected clips were chosen to produce the most challenging quality assessment problem (i.e., the selected clips contained impairments that were not easily quantified by simpler metrics[1-7]). The results from this preliminary test were then used to limit the number of edge enhancement filters and S-T region sizes for the other data sets.

4.1.1 Edge Enhancement Filter Size

The 181-clip training subset revealed that the large filters shown in Figure 4 outperformed the Sobel filters shown in Figure 3. For the 181-clip subset, the optimal amount of vertical lowpass filtering (i.e., the vertical size, or number of rows in the Figure 4 filter) was found to be about 13 lines. With this filter size (13 x 13), the optimal amount of horizontal bandpass filtering (i.e., c = 1, 1.5, 2, or 3) for the 181-clip subset was found to be given by the c = 2 filter. This bandpass filter has a peak response at about 4.5 cycles/degree for Rec. 601 video viewed at 6 times picture height.

Comparing the Sobel filter (Figure 3) with the 13 x 13 filter specified by c = 2 (Figure 4), gain and loss metrics derived from the f_1 feature in section 2.3 had modestly better correlation results but metrics derived from the f_2 feature produced substantial improvements. It appeared that the Sobel filter did not perform sufficient averaging to obtain robust estimates of the angular orientation of the spatial gradient energy. Therefore, we decided to use the 13 x 13 filter specified by c = 2 for preprocessing prior to extracting both the f_1 and f_2 features described in section 2.3.

4.1.2 S-T Region Size

We discovered that since most lower bit-rate systems do not preserve frame integrity (i.e., many of these systems transmit fewer than 30 frames per second and repeat prior frames to fill in for the missing frames), the optimal temporal-width of the S-T region (see Figure 5) must be increased from the 1-frame temporal width previously obtained for MPEG-2 video systems.[10] Larger temporal extents were found to accommodate temporal misalignments for the lower bit-rate video teleconferencing systems while still preserving the high correlation results for the MPEG-2 video systems. The optimal S-T region size that achieved the maximum correlation with subjective ratings for the 181-clip subset was on the order of 8 horizontal pixels x 8 vertical lines x 6 video frames as shown in Figure 5. It should be noted, however, that the correlation was found to worsen *slowly* as one moves away from the optimum point (i.e., either reducing the S-T granularity or increasing the S-T granularity). This result agreed with what was previously found for MPEG-2 systems.[10] Horizontal and vertical

widths up to 32 pixels or lines, and temporal widths up to 30 frames, can be used with satisfactory results, giving the objective measurement system designer considerable flexibility in adapting the techniques presented here to lower ancillary data channel bandwidths (see Figure 1). For the results in this paper, we decided to extract the f_1 and f_2 features from a S-T region size of 8 horizontal pixels x 8 vertical lines x 6 video frames.

4.2 Testing on All Seven Data Sets

Figure 7 presents the Pearson linear correlation coefficients between the two spatial activity *loss* metrics described in section 2 and the seven subjective data sets described in section 3. The plots include the preliminary training subset from data set two that was used to develop the metrics. As previously mentioned in sections 4.1.1 and 4.1.2, these results are for the large 13 x 13 edge enhancement filter specified by $c = 2$ (see Figure 4) and a S-T region size of 8 horizontal pixels x 8 vertical lines x 6 video frames (see Figure 5). In Figure 7, the correlation results for the f_1*_loss* parameter (i.e., the loss of the f_1 feature measured using the loss equation in section 2.4) is on the left while the correlation results for the f_2*_loss* parameter (i.e., the loss of the f_2 feature measured using the loss equation in section 2.4) is on the right. With the exception of data set three, both loss parameters yield consistently high correlation results across all data sets, demonstrating that both the magnitude and angular orientation of the spatial activity contains meaningful quality assessment information. Data set three spans the lowest range of quality and contains temporal and chroma impairments that are not well quantified by spatial gradient metrics derived from the luminance signal.

Figure 8 presents the Pearson linear correlation coefficients between the f_2*_gain* metric in section 2 (i.e., the gain of the f_2 feature measured using the gain equation in section 2.4) and the subjective data sets in section 3. The f_1*_gain* metric is not displayed since it did not produce consistent correlation results across all data sets (correlations ranged from 0.22 to 0.91 for the f_1*_gain* metric). Spatial activity gain in digital video systems quite often takes the form of tiling or block distortion, which is picked up by the f_2*_gain* metric. One would expect the gain metrics to complement the loss metrics since in general both types of impairments can be present in the same digital video clip. The complementary nature of the loss and gain parameters will be investigated further in the next section.

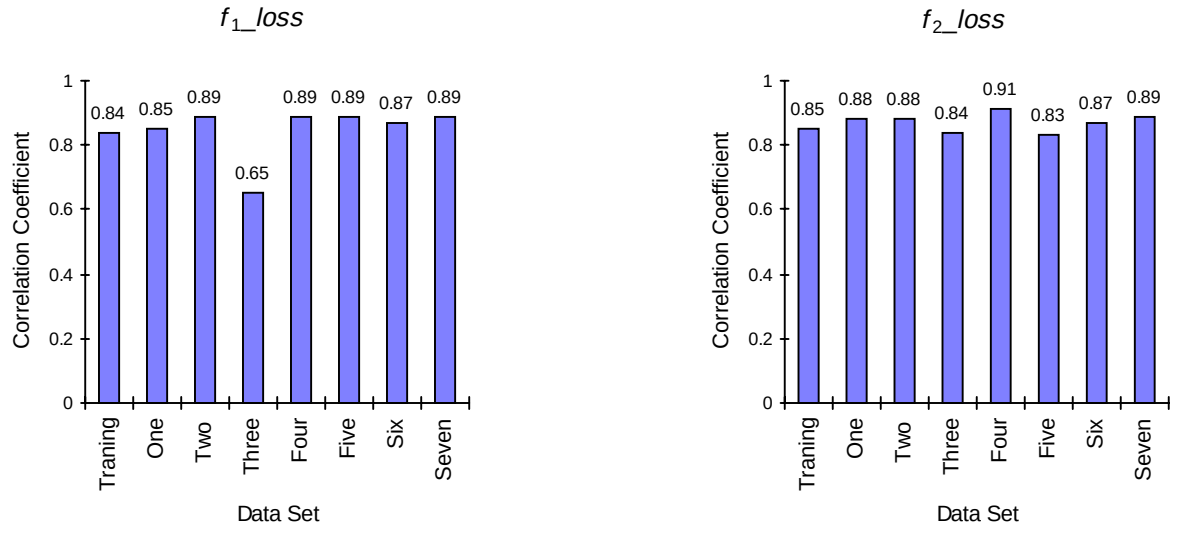

Figure 7. Pearson linear correlation coefficients for the f_1*_loss* and f_2*_loss* parameters.

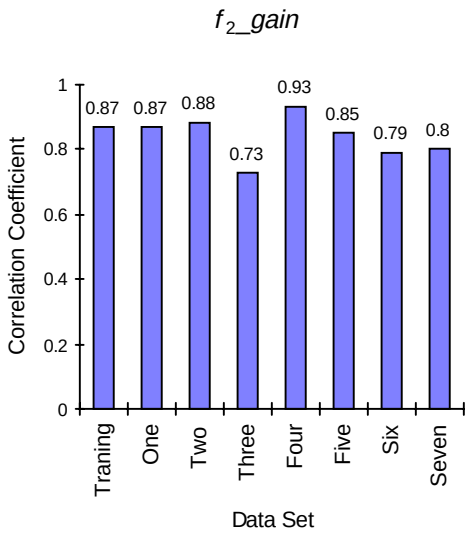

Figure 8. Pearson linear correlation coefficient for the f_2_gain parameter.

4.3 Combined Spatial-Temporal Distortion Metric

This section proposes a combined spatial-temporal distortion metric that is sensitive to losses in the magnitude of spatial activity (i.e., f_1_loss), losses in the HV to $\overline{HV}$ ratio of spatial activity (i.e., f_2_loss), and gains in the HV to $\overline{HV}$ ratio of spatial activity (i.e., f_2_gain). This combined parameter (*join*) was computed by determining the optimal proportions of each parameter for each of the seven data sets and then averaging these proportions across all seven data sets to produce

$$join = 0.38 * f_1\_loss + 0.39 * f_2\_loss - 0.23 * f_2\_gain .$$

This process assures that each data set is treated equally regardless of the number of clips in the data set. In the above equation, f_1_loss and f_2_loss have positive weights since these parameters range between zero and negative one. On the other hand, f_2_gain has a negative weight since it is always greater than or equal to zero. Across the seven data sets, the combined metric *join* ranged from zero (no impairment) to approximately negative one (really poor quality). Testing the combined metric *join* on each of the data sets produced the Pearson linear correlation coefficients shown in Figure 9. The combined metric *join* achieved an average correlation coefficient of 0.88 across the seven data sets. In view of the breath of the subjective data in these seven experiments, this result is quite significant. The combined metric *join* thus explained about 77% (0.88^2) of the variance in the subjective data sets. Typically, 10% of the subjective variance is random noise due to finite viewer populations. This leaves only 13% of the subjective variance to be explained by other measures (e.g., temporal transients, chroma distortion).

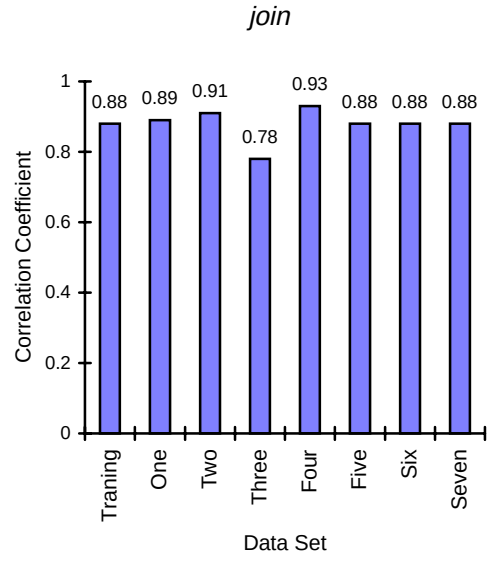

Figure 9. Pearson linear correlation coefficient for the *join* parameter.

5. CONCLUSIONS

We have presented video quality metrics based on features extracted from S-T regions that quantify both the magnitude and direction of the spatial gradient. These metrics explain a large percentage of the variance in seven subjective data sets that span an extremely wide range of test scenes and digital video systems. We therefore feel that these metrics are indicative of the perceived quality that results from any digital video system, regardless of operating bit-rate. The horizontal and vertical edge enhancement filters that are utilized for estimation of the spatial gradients emphasize long edges (on the order of 1/5 degree) while simultaneously performing large amounts of noise suppression. Two separate visual masking functions have been presented that emulate human perception; one for gains in feature values and another for losses.

In addition to being highly correlated to subjective ratings, the metrics may also be used for continuous in-service quality monitoring. This is important since digital video quality depends upon dynamic characteristics of the input video (e.g., spatial detail, motion) and the digital transmission system (e.g., bit-rate, error-rate) and thus accurate perception-based measurements must be made in-service. The S-T region size from which the reduced reference information features are extracted can be adjusted to match the bandwidth

of the in-service data channel used to communicate the features between the input and output ends. While we have presented results for a S-T region size of 8 horizontal pixels x 8 vertical lines x 6 video frames, for a feature compression factor of 384, larger S-T region sizes may also be used with only a minor drop in the ability of the metrics to track perceptual video quality.

REFERENCES

1. Stephen Voran and Stephen Wolf, "The Development and evaluation of an objective video quality assessment system that emulates human viewing panels," International Broadcasting Convention (IBC), July, 1992.
2. Arthur A. Webster, Coleen T. Jones, Margaret H. Pinson, Stephen D. Voran, and Stephen Wolf, "An objective video quality assessment system based on human perception," Human Vision, Visual Processing, and Digital Display IV, Proceedings of the SPIE, Volume 1913, February, 1993.
3. Neal Seitz, Stephen Wolf, Stephen Voran, and Randy Bloomfield, "User-oriented measures of telecommunication quality," *IEEE Communications Magazine*, January, 1994.
4. United States Patent 5,446,492, "Perception-Based Video Quality Measurement System," awarded August 29, 1995.
5. ANSI T1.801.03-1996, "American National Standard for Telecommunications - Digital Transport of One-Way Video Telephony Signals - Parameters for Objective Performance Assessment," American National Standards Institute.
6. United States Patent 5,596,364, "Perception-Based Audio-Visual Synchronization Measurement System," awarded January 21, 1997.
7. S. Wolf, "Measuring the end-to-end performance of digital video systems," *IEEE Transactions on Broadcasting*, September 1997, Volume 43, Number 3, pages 320-328.
8. G.W. Cermak, S. Wolf, E. P. Tweedy, M.H. Pinson, and A. A. Webster, "Validating objective measures of MPEG video quality," *SMPTE Journal*, April 1998, Volume 107, Number 4, pages 226-235.
9. Coleen Jones and D. J. Atkinson, "Development of opinion-based audiovisual quality models for desktop video-teleconferencing," 6th IEEE International Workshop on Quality of Service, Napa, California, May 18-20, 1998.
10. S. Wolf and M. Pinson, "In-service performance metrics for MPEG-2 video systems," Made to Measure 98 - Measurement Techniques of the Digital Age Technical Seminar," technical conference jointly sponsored by the International Academy of Broadcasting (IAB), the ITU, and the Technical University of Braunschweig (TUB), Montreux, Switzerland, November 12-13, 1998.
11. ITU-R Recommendation BT.601, "Encoding parameters of digital television for studios," Recommendations of the ITU, Radiocommunication Sector.
12. ITU-R Recommendation BT.500, "Methodology for subjective assessment of the quality of television pictures," Recommendations of the ITU, Radiocommunication Sector.
13. ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Recommendations of the ITU, Telecommunication Standardization Sector.
14. ANSI Accredited Standards Working Group T1A1 contribution number T1A1.5/94-118R1, "Subjective test plan (tenth and final draft)," Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC, October 3, 1993.
15. ANSI T1.801.01-1995, "American National Standard for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Video Test Scenes for Subjective and Objective Performance Assessment," American National Standards Institute.
16. Fernando Pereira, "MPEG-4 video subjective test procedures and results," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No. 1, February 1997.
17. Charles Fenimore, et. al., "Perceptual effects of noise in digital video compression," 140th SMPTE Technical Conference, Pasadena, CA, October 28-31, 1998.