The Relationship Between Performance and Spatial-Temporal Region Size for Reduced-Reference, In-Service Video Quality Monitoring Systems

Stephen Wolf and Margaret H. Pinson
Institute for Telecommunication Sciences, National Telecommunications and Information Administration
325 Broadway, Boulder, CO 80305, USA

ABSTRACT

This paper presents objective-to-subjective correlation results for a reduced-reference, in-service, video quality monitoring system. This reduced-reference system utilizes quality parameters that are computed by comparing features extracted from spatial-temporal (S-T) regions of the input video stream with identical features extracted from the output video stream. The amount of reduced-reference information that is required to compute the quality parameters is inversely related to the size of the S-T region. Smaller amounts of reference information (i.e., larger S-T regions) are desired since less transmission or storage bandwidth is required for the reference information. However, objective-to-subjective correlation drops off if the S-T region size becomes too large. In this paper we examine the tradeoffs between objective-to-subjective correlation results and S-T region size. Correlation results for S-T region sizes from 8 vertical lines x 8 horizontal pixels x 2 video frames to 128 x 128 x 24 are presented. These results utilized a total of nine subjectively rated data sets that span an extremely wide range of bit rates and compression techniques. Thus, designers of television video systems as well as Internet video streaming systems may use the results.

Keywords: Video Quality, In-Service, Reduced-Reference, Subjective Testing, Objective Testing, Compression.

1. INTRODUCTION

True end-to-end in-service quality measurements are necessary for many digital video systems since their delivered quality depends upon numerous time varying quantities such as scene content, coder bit-rate, digital transmission system characteristics (e.g., error characteristics, best effort data delivery mechanisms), and decoder error concealment strategies. One method of performing in-service quality monitoring is shown in Figure 1. This method has been referred to as the reduced-reference (RR) method by the ITU since quality measurements are based on features that are extracted from the input and output video streams [7]. Since the extracted features have much less bandwidth than the input and output video streams, they can be readily transmitted using a commonly available ancillary data channel (e.g., public switched telephone network, wireless, or Internet connection).

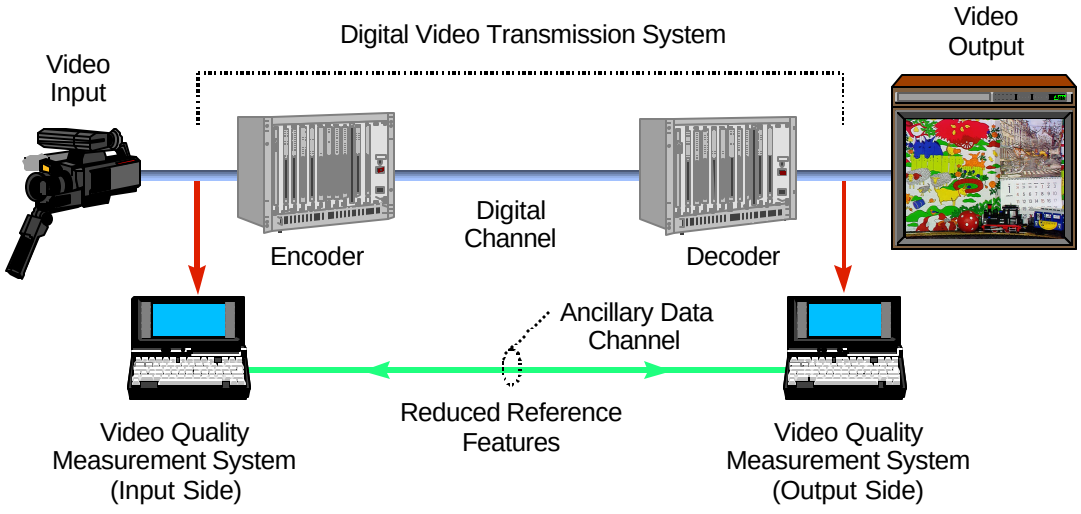

Figure 1. Reduced-reference in-service video quality measurement system.

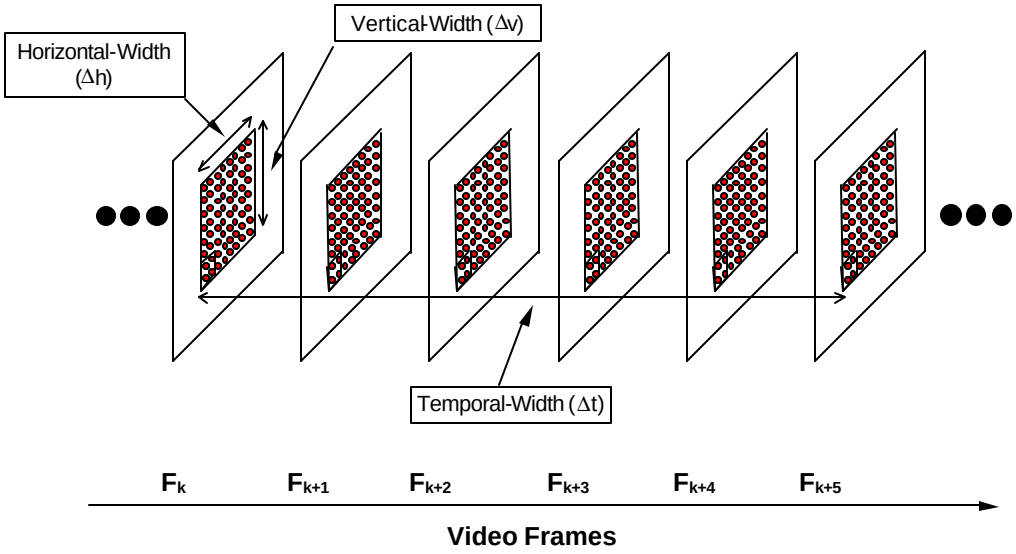

Figure 2. Illustration of a spatial-temporal (S-T) region for a video scene.

For the purposes of this paper, a feature is defined as a quantity of information (e.g., a summary statistic such as standard deviation) that is associated with a specific spatial-temporal (S-T) region of the video sequence. Figure 2 gives an illustration of an S-T region that includes 8 vertical lines x 8 horizontal pixels x 6 video frames. As the number of pixels encompassed by the S-T region increases, the required bandwidth of the ancillary data channel shown in Figure 1 decreases, since less reference information is being used to make the quality measurement. For instance, extracting one feature from an 8 x 8 x 6 S-T region yields a compression factor of 384 to 1, while extracting the same feature from a 32 x 32 x 18 S-T region yields a compression factor of 18432 to 1.

Previous papers have presented a set of features and feature comparison functions (i.e., a means of comparing input and output features to produce quality parameters) that produce excellent correlation to subjective quality assessments for a wide range of video systems [11, 12, 13, 14, 15]. These features are based on the spatial and temporal information content of the input and output video sequences.

The purpose of this paper is to present recent objective-to-subjective correlation results for these quality parameters as a function of S-T region size. To this end, we have examined S-T region sizes that have combinations of the spatial extents (in lines x pixels per line: 8 x 8, 16 x 16, 32 x 32, 64 x 64, and 128 x 128 for ITU-R Recommendation BT.601 sampled images [5]) with the temporal extents (in frames: 2, 6, 12, 18, and 24). The subjective data that was used to generate these correlation results consisted of nine experiments, conducted from 1992 to 1999. Five of the data sets were primarily digital television experiments while the other four were primarily video conferencing experiments.

Objective-to-subjective correlation results are presented for each type of video system (i.e., video conferencing and television). The correlation results are presented as a two dimensional function of the spatial and temporal extents of the S-T region illustrated in Figure 2. The designer of RR in-service quality monitoring systems (Figure 1) can use these results to optimize measurement performance (i.e., objective-to-subjective correlation) for a given ancillary data channel bandwidth. Surprisingly, even relatively large S-T regions (and hence very low ancillary data channel bandwidths) can produce good objective-to-subjective correlation results for some video quality parameters and data sets.

2. SUBJECTIVE DATA

All nine of the subjective experiments were conducted in accordance with the most recent version of ITU-R Recommendation BT.500 [4] or ITU-T Recommendation P.910 [8] that was available when the experiment was performed. All of the data sets used scenes from 9 to 10 seconds in duration. Seven of the data sets (i.e., data sets one to seven) used double stimulus testing where viewers saw both the original and processed sequences. Two of the data sets (i.e., data sets eight and nine) used single stimulus testing where viewers saw only the processed sequence. Five of the data sets (i.e., data sets one to five) contained primarily digital television systems operating at bit rates greater than 1.5 Mbits/sec. Four of the data sets (i.e., data sets six to nine) contained primarily video conferencing systems operating at bit rates from 10 kbits/sec to 1.5 Mbits/sec. In all, the nine subjective data

sets included nearly 1600 combinations of test scenes and video systems.

For brevity, only a summary of each subjective experiment is given here. The reader is directed to the accompanying references for more complete descriptions of the experiments.

2.1 Data Set One [14]

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 standard Rec. 601 test scenes. The nine MPEG-2 systems operated at bit rates from 2 Mbits/sec to 8 Mbits/sec. Naïve 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).

2.2 Data Set Two [3]

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 contained 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. The subjective test procedure was the same as data set one.

2.3 Data Set Three [14]

A panel of 32 viewers rated a total of 112 video clips that were generated by pairing two sub-groups of eight scenes each 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 standard 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 one.

2.4 Data Sets Four and Five [6]

Data sets four and five (525-line and 625-line, respectively) were each generated by pairing ten scenes with sixteen video systems to produce 160 video clips per data set. For each data set, a total of 60 to 80 naïve viewers from four different subjective testing laboratories (i.e., 15 to 20 viewers per laboratory) rated subjective quality using the double stimulus continuous quality scale (DSCQS) method defined in ITU-R Recommendation BT.500 [4]. The twenty different test scenes (ten for 525-line, ten for 625-line) included sports material, standard Rec. 601 test scenes, moving graphics, and stills. The video systems included MPEG-2 systems operating at bit rates from 2 Mbits/sec to 50 Mbits/sec, video teleconferencing systems operating at 768 kbits/sec and 1.5 Mbits/sec, some systems with digital transmission errors, multi-generation MPEG-2, and multi-generation ½-inch professional record/play cycles, where composite and/or component signal formats were used.

2.5 Data Set Six [1, 2]

Viewer panels comprising a total of 30 naïve viewers from three different subjective testing laboratories rated 600 video clips that were generated by pairing 25 test scenes with 24 video systems. The 25 test scenes 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. Viewers were shown the original version first, and then the degraded version, and were asked to rate the difference in perceived quality using the 5-point impairment scale (imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying).

2.6 Data Set Seven [11, 12]

A panel of 48 naïve 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. The subjective test procedure was the same as data set six.

2.7 Data Set Eight [10]

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 two 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 an 11-point numerical scale, with 0 being the worst quality and 10 being the best.

2.8 Data Set Nine [9]

A panel of 18 naïve viewers rated 48 video clips in a desktop video teleconferencing application. Pairing six scenes with eight different video systems generated the 48 video clips. The six test scenes were selected from ANSI T1.801.01 [2] 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).

3. OBJECTIVE DATA

Briefly, three objective parameters derived from two different features were used for the objective-to-subjective correlation results given in section 4. These features and parameters have been fully documented in a prior paper [15] and hence for brevity will only be summarized here. The first parameter, *f1_loss*, quantifies the reduction of spatial detail in the video output with respect to the video input. Because *f1* is a spatial activity feature that is only sensitive to the magnitude of the spatial gradients (i.e., edge magnitude), a loss in *f1* (i.e., *f1_loss*) measures the perceptual effects of video impairments such as blurring and smearing. Two other parameters, *f2_loss* and *f2_gain*, were derived from a second spatial activity feature *f2* that is sensitive to the angular orientation of the spatial gradients (i.e., edge orientation as in vertical, horizontal, and diagonal). The *f2* feature measures the ratio of horizontal/vertical spatial gradients to non-horizontal/vertical spatial gradients. Hence, a loss or gain in *f2* (i.e., *f2_loss* and *f2_gain*, respectively) measures the perceptual effects of video impairments such as tiling, block distortion, and line-oriented noise.

4. OBJECTIVE-TO-SUBJECTIVE CORRELATION

The three objective parameters (*f1_loss*, *f2_loss*, and *f2_gain*) were measured at each S-T region size for all the subjectively rated data sets given in section 2. The linear Pearson correlation coefficient between each subjective data set and the objective parameter was calculated. Then, an average correlation coefficient was calculated over the five television data sets (i.e., data sets one to five), and over the video conferencing data sets (i.e., data sets six to nine). This simple averaging of correlation coefficients over data sets weights each data set equally regardless of the number of clips in the data set.

The correlation results obtained from this procedure are shown in Figures 3-5. The correlation results are three dimensional (spatial extent, temporal extent, and average correlation coefficient). For each spatial extent (8 x 8, 16 x 16, 32 x 32, 64 x 64, and 128 x 128), there were five temporal extents (2, 6, 12, 18, and 24 frames). The X-axis in these plots contains both spatial and temporal extent information, with the vertical grid lines representing increasing spatial extent, while the individual points of each curve represent increasing temporal extent.

For a given spatial and temporal extent, the correlation coefficient of the television data is normally less than the video conferencing data. This is because the total range of video quality for the television data is less than that for the video conferencing data. The one exception to this observation, which is currently under investigation, is the correlation results of the *f2_gain* parameter for spatial extents of 64 x 64 and 128 x 128 (see Figure 5).

In nearly every case, the highest correlation coefficient is achieved for an S-T region size of 8 lines x 8 pixels x 6 frames. For a given spatial extent, correlation results fall off slowly when temporal extent is increased from six video frames. Surprisingly, temporal extents of up to 24 video frames (i.e., approximately 0.8 seconds) perform nearly as well as temporal extents of six video frames (i.e., 0.2 seconds). The curves suggest that increasing temporal extent by a factor of four might be better than increasing spatial extent by a factor of two in both the horizontal and vertical directions. Both options increase the compression of the features by a factor of four.

For a given spatial extent, the correlation results for a temporal extent of two video frames is almost always worse than the correlation results for a temporal extent of six video frames. This effect is most noticeable for the television data at smaller spatial extents. The reason might be that the human visual system has a reduced response to small spatial impairments of short temporal duration.

Using a temporal duration of six or more video frames has the equivalent effect of averaging through these small, brief impairments, so that their perceptual impact is reduced.

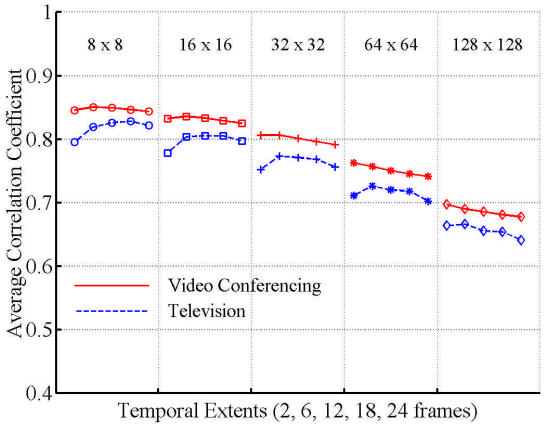

Figure 3. Correlation results for *f1_loss*.

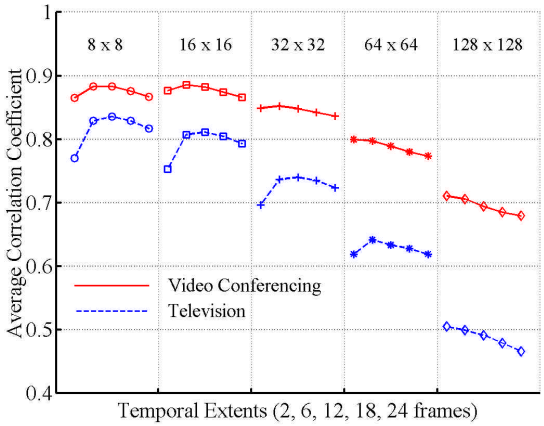

Figure 4. Correlation results for *f2_loss*.

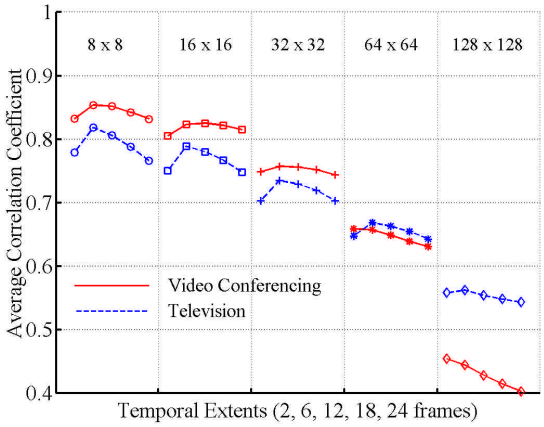

Figure 5. Correlation results for *f2_gain*.

Good correlation results (i.e., greater than 0.8) can be achieved with the *f1_loss* and *f2_loss* parameters for S-T region sizes up to 32 x 32 x 24. This corresponds to a feature compression factor of 24,576. Assuming 8 bits are used to encode each feature extracted from an ITU-R Recommendation BT.601 luminance signal (i.e., 486 lines x 720 pixels x 30 frames/sec), this compression factor would produce a feature stream with a data rate of about 3.4 kbits/sec. This feature stream is easily transmitted over many common telecommunications networks (e.g., public switched telephone network, wireless or cell phone, Internet).

We also computed the average correlation coefficient achievable with a video quality model that uses all three of the above parameters (i.e., a linear combination of *f1_loss*, *f2_loss*, and *f2_gain*). The 3-parameter model results are shown in Figure 6. The same general trends are observed as before except that the overall correlations to subjective score are higher because the 3-parameter model is able to fully utilize all of the complementary information in the parameters.

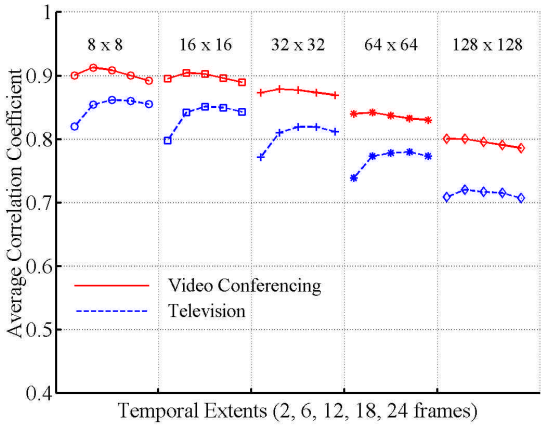

Figure 6. Correlation results for 3-parameter model.

5. CONCLUSIONS

We have investigated the relationship between performance and S-T region size for a reduced-reference video quality measurement system. While the results presented are for a specific set of reduced-reference features, we believe these results may be useful to others working in the area, even though they are using different feature sets. The optimal S-T region size for extracted features appears to be on the order of 8 lines x 8 pixels x 6 video frames; however, correlation results fall off slowly with increasing spatial and temporal extents. Reasonable correlation results can be obtained for very large S-T region sizes, making very low bandwidth reduced-reference video quality measurement systems feasible.

6. REFERENCES

1. ANSI Accredited Standards Working Group T1A1.5 contribution number T1A1.5/94-118R1, "Subjective test plan (tenth and final draft)," available from the Alliance for Telecommunications Industry Solutions (ATIS), 1200 G Street, NW, Suite 500, Washington, DC 20005, October 3, 1993.
2. 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.
3. Fenimore, C., et. al, "Perceptual effects of noise in digital video compression," *SMPTE Journal*, vol. 109, pp. 178-186, March 2000.
4. ITU-R Recommendation BT.500, "Methodology for subjective assessment of the quality of television pictures," Recommendations of the ITU, Radio-communication Sector.
5. ITU-R Recommendation BT.601, "Encoding parameters of digital television for studios," Recommendations of the ITU, Radio-communication Sector.
6. ITU-T COM 9-80-E, "Final report from the video quality experts group (VQEG) on the validation of objective models of video quality assessment," Contributions of the ITU, Study Group 9, Telecommunication Standardization Sector, June 2000.
7. ITU-T Recommendation J.143, "User requirements for objective perceptual video quality measurements in digital cable television," Recommendations of the ITU, Telecommunication Standardization Sector.
8. ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Recommendations of the ITU, Telecommunication Standardization Sector.
9. Jones, C., and Atkinson, D. J., "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. Pereira, F., "MPEG-4 video subjective test procedures and results," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No. 1, February 1997.
11. Voran, S., and Wolf, S., "The development and evaluation of an objective video quality assessment system that emulates human viewing panels," International Broadcasting Convention (IBC), July 1992.
12. Webster, A. A., et. al, "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.
13. Wolf, S., "Measuring the end-to-end performance of digital video systems," *IEEE Transactions on Broadcasting*, September 1997, Volume 43, Number 3, pages 320-328.
14. Wolf, S., and Pinson, M., "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.
15. Wolf, S., and Pinson, M., "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system," in Proc. SPIE International Symposium on Voice, Video, and Data Communications, Boston, MA, September 1999.