How to Build an Objective Model for Packet Loss Effect on High Definition Content Based On SSIM and Subjective Experiments

Piotr Romaniak[1] and Lucjan Janowski[1]

AGH University of Science and Technology, Department of Telecommunications,
{romaniak, janowski}@kt.agh.edu.pl,

Abstract. In this paper the authors present a methodology for building a model for packet loss effect on High Definition video content. The goal is achieved using the SSIM video quality metric, temporal pooling techniques and content characteristics. Subjective tests were performed in order to verify proposed models. An influence of several network loss patterns on diverse video content is analyzed. The paper deals also with encountered difficulties and presents intermediate steps to give a better understanding of the final result. The research aims at the perceived evaluation of a network performance for IPTV and video surveillance systems. The final model is generic and shows high correlation with the subjective results....

1 Introduction

Packet networks became one of a critical technology for video streaming services. Important example of such service is IPTV implemented by many network operators. Recent premiere of High Definition IPTV brought new requirements in terms of bit-rate and quality of service assurance. Problem of network losses is still / again vivid and affects mainly the "last mile" of the delivery path. Another example of a service using packet network for video streaming are surveillance systems of urban areas. In such systems, availability of a dedicated uplink is quite rare and the quality assurance problems of the "first mile" play a crucial role. Competition on both markets is strict and service providers desperately seek video quality monitoring and assurance solutions in order to satisfy more and more quality aware customers. The impact of network losses on the perceived video quality is still challenging task because (among others) "not all packets are equal" as claimed in [1].

Evaluation of packet loss effect on video content was extensively analyzed over recent years. Several models were proposed for low bit-rate videos (dedicated for mobile scenarios) and Standard Definition (SD) resolution. The majority of proposed solutions are so-called parametric models, operating on the network and transport layers. Verscheure in [2] explains the problem of quality prediction and control of a MPEG-2 video stream transmitted through the lossy network. The MPEG-2 video standard is analyzed and the impact on the visual quality

of packet loss is discussed. In [3] the authors presented a quality metric based on two network parameters related to packet loss. An application of the customer oriented measurements for H.264/AVC video is presented in [4]. Another model accounting for effect of burst losses and correlation between error frames was detailed in [5]. It is dedicated to low bit-rate H.264/AVC video. In contrast to the presented parametric approaches a simple model for network impairments based on image analysis was proposed by Dosselmann in [6].

There are many other interesting models aiming at low bit-rate and SD resolution content. In contrast, hardly few can be found on High Definition content. One of the first substantial publications on this particular topic describes performance of the NTIA General Video Quality Metric (VQM) [7] in the task of High Definition TV (HDTV) video quality assessment [8]. It is devoted mainly to compression artifacts for 5 different video encoders. Network losses are considered additionally, with a lower stress put on it. Another research published recently in [9] is dedicated exclusively to network losses. Correlation of three existing quality metrics was verified upon the subjective results, namely PSNR, SSIM [10] and VQM [7]. However, only one network loss pattern with variable number of occurrences per video sequence was considered. What is more, subjective and objective scores were averaged over 2 minutes of video material consisting of 12 video sequences. This simplifies the quality assessment task because an important factor influencing perceived quality is omitted this way. This factor is related to diverse content characteristics and may significantly influence perceived quality of different contents affected with the same (in terms of quantity) impairments [11], [12]. In result, the authors claim that even PSNR metric can achieve extremely high correlation with perceived quality, what is obviously wrong. Recent discussion on performance of mean squared error metrics is presented by Wang and Bovik in [13].

In our research an influence of diverse network loss patterns on the perceived video quality is investigated. We focus on Full HD videos, being diverse in terms of content characteristics. Our model is based on SSIM metric so it represents image analysis approach. Moreover, we show that average SSIM calculated over all video frames is not sufficient. We propose more advanced strategies for temporal pooling aided with content characteristics calculation. A process of building the final model is presented step by step with description of all the problems met along the way. Final result proves correctness of applied methodology and high quality prediction accuracy of the model.

2 Preparation of the Experiment

This section details necessary steps we had to perform prior to the subjective experiment. It includes video pool selection, simulation environment setup and selection of the quality metric.

2.1 Selection of Video Sequences

Our original video pool consists of eight VQEG video sequences [14], [15] in Full HD resolution (1920 × 1080), 29.97 FPS (frames per second), around 10 seconds long (exactly 300 frames). The selected pool represents wide variety of content and video characteristics. We calculated two characteristics: 1) spatial activity *SA* and 2) temporal activity *TA*, as proposed in [16]. Both measures are combined into one describing video complexity *O* (the difficulty of encoding a scene), detailed in [17]. Table 1 orders video sequences according to the scene complexity *O*. It is interesting to notice that sequence number 8 has high *SA* and *TA* at the same time, what is quite rare case in natural videos.

The sequences were encoded using H.264 codec (Apple QuickTime implementation) main-profile (Level 40) and the average bit-rate of 20 Mbit/s. The selected video container was QuickTime *MOV*.

Table 1. Characteristics of the selected video sequences and MOS values for the "perfect" streaming scenario

Name	**TA**	**SA**	**O**	**MOS**	**No.**
SnowMnt	2.49	168.80	6.04	4.04	1
ControlledBurn	4.37	138.37	6.24	3.87	2
SpeedBag	17.82	53.11	6.49	4.58	3
TouchdownPass	13.16	68.42	6.64	4.17	4
Aspen	14.22	85.83	6.95	4.50	5
RedKayak	24.45	82.88	7.44	3.67	6
RushFieldCuts	12.74	158.70	7.61	4.25	7
WestWindEasy	23.07	180.80	8.34	4.33	8

2.2 Simulation Environment

Our simulation environment was simple video server–network emulator–client architecture. Both server and client were *VideoLAN* — free streaming solution, running under *Ubuntu* Linux. For streaming we used Transport Stream container and *RTP* protocol. The video client is capable of producing stream dumps, necessary for further video sequences processing (quality assessment). The network emulator can produce 15 different scenarios, where one is perfect and other 14 are different in terms of packet loss ration and loss pattern. From our perspective the only important aspect of the emulator is the impact on the perceived video quality. Resulting video degradation ranges from imperceptible to very annoying. Streaming scenarios and network emulator itself is a topic for another paper.

2.3 Selection of Video Quality Metric

We decided to select some well-known quality metric, operating in full-reference mode. The choice was the SSIM (Structural Similarity Index Metric) [10]. The motivation for our choice is availability of the SSIM, simplicity and good correlation with the human perception, proved in the VQEG FR-TV Phase I report [18]. The SSIM was originally designed for still images quality assessment, however, an extension for video applications was presented in [19]. As presented by Wang in [20], [21] the human visual system (HVS) is very sensitive to the structural information provided on an image in the viewing field. Based on this assumption, the SSIM can have good correlation with the perceptual quality in our case, since artifacts caused by packet loss introduce structural changes to the video frames.

3 Subjective Experiment

In this section we describe in detail the subjective experiment. Applied methodology strictly follows the VQEG HDTV Test Plan methodology [22]. The ultimate goal of the experiment was to obtain MOS (Mean Opinion Score) for video sequences transmitted by our network architecture. Based on the subjective scores we hope to build an objective model for packet loss effect on Full HD content, using the SSIM metric.

3.1 Test set

Our test set consists of 58 video sequences. It was obtained by streaming the original video pool through the network architecture described in section 2.2. The following runs were selected: 1) all 8 sequences were streamed using perfect scenario (only in theory perfect, as it has been shown in the remaining part of the paper), 2) sequences 2 and 7 streamed using scenarios 1-5, 3) sequences 2, 4, 6, 8 streamed using 5 scenarios 6-10, and 4) sequences 1, 3, 5, 7 streamed using scenarios 11-15. Afterwards, all 58 sequences were transcoded using *FFmpeg* into *MPG* container. Selected video codec was *MP2* and one important setting during encoding was maximum peak bit-rate limited to 40 Mbit/s. This configuration was enforced by our Blu-Ray player, capable of handling *MP2* video stream up to 40 Mbit/s. Although the transcoding, no further distortions were introduced to the content (bit-rate of 20 Mbit/s were up-scaled to < 40 Mbit/s). It ensures that the quality degradation is exclusively due to network losses introduced during the streaming. Prepared test set was recorded at Blu-Ray discs.

3.2 Methodology

The methodology we used is called Absolute Category Rating with Hidden Reference (ACR-HR) and is described in recommendation ITU-T P.910 [23]. It represents a Single-Stimulus (SS) approach, i.e. all video sequences contained

in a test set are presented one by one without a possibility to compare with the reference video. The reference sequences are also included in a test set and rated according to the same procedure. Video sequences were rated according to the five-grade MOS quality scale [24]. An instruction was given and read to the subjects before the test.

3.3 Environment

Subjective tests was performed at AGH University of Science and Technology laboratory using calibrated LG 42LH7000 42" LCD displays, 1920×1080 resolution. The viewing distance was fixed to 1 minute of arc, which means $3H$, where $H = PhysicalPictureHeight$. Each subject was seated in front of his/her own display with eyes aligned vertically and horizontally with the display center. The test room lightning conformed to ITU-R Rec. BT.500-11 [25] requirements. The test sequences were played from Blu-Ray disc, using Blu-Ray LG BD-370 player, connected to the display using HDMI interface. Such set-up assures flawless playback of the prepared video material. The time of viewing was around 20 minutes (10 seconds for a single sequence and 10 seconds for voting).

3.4 Subjects

We selected a group of 30 subjects in order to fulfill the requirement of minimum 24 valid ones at the end. The group was diverse in terms of age, gender and education background. The criterion for subjects validation was the linear Pearson correlation coefficient calculated per sequence for the subject vs. all subjects (as defined in [22]) higher than 0.75.

4 Computation of the SSIM metric

In this section we describe necessary steps towards the SSIM metric calculation for our streamed sequences. What really matters for full reference metrics is synchronization in order to ensure that the corresponding frames from both the reference and the distorted sequences are compared. In our case, calculation of the SSIM metrics caused also necessity of down-scaling video frames by factor 4, as proposed here [26]. Sections 4.1 and 4.2 describe also potential problems one may face while streaming Full HD content using non-professional software and hardware, as we did.

4.1 Inspection of the Sequences

We decided to make visual and technical inspection of the streamed sequences first, to make sure that the streaming performed as expected. We discovered two alarming things. First, the nominal playback speed (FPS rate) on PC computer of the streamed sequences was not the same as for the original ones (despite that roughly no frames were lost). In order to eliminate this problem we ensured that

the subjects saw the original FPS rate during the subjective tests (see section 3.1).

$$DMOS = MOS(distorted) - MOS(original) + 5 \quad (1)$$

The second problem suggests that the flawless streaming of 20 Mbit/s video content using our setup is problematic. The "perfect" streaming scenario introduced slight distortions caused by a single packet loss for 2 out of 8 sequences (number 2 and 6 from table 1). For the purpose of our analysis it is not a problem since we decided to use DMOS (Differential MOS, given by equation 1) instead of MOS for the purpose of further analysis. It is much more suitable for full reference metrics, which are not capable of the absolute quality assessment. DMOS eliminates also known problem "Is the reference really a reference?".

4.2 Synchronization of the Sequences

Another task to be fulfilled prior to the SSIM calculation is synchronization of the reference and distorted sequences. By the "reference" we understand original sequences streamed over the "perfect" scenario, and by "distorted" we understand original sequences streamed over other (lossy) scenarios. In result, SSIM = 1 for the corresponding frames not affected with packet loss should be obtained. In our case synchronization was limited to temporal alignment only.

The first encountered problem was related to missing frames (skipped during streaming by the server or client). We have classified this type of loss as another problem related to our streaming architecture. A single missing frame within 10 seconds long sequence (300 frames) is imperceptible, but for a full reference metric operating on frames level it means lost of synchronization and has to be detected.

Another challenge was the synchronization recovery after long freeze caused by an extensive packet loss. Such freeze may spread over dozens of frames (in our case the longest freeze was around 25 frames). In order to solve this problem our first thought was to increase search depth to cover the longest freeze, however, another problem arose. It was especially visible for still scenes affected with few-frames-long packet loss artifact. For such scenes, the synchronization caught the first frame from the distorted sequence not effected with packet loss. Even if the distorted frame was around 5 frames ahead from the reference one, the scene change resulted in smaller difference in terms of the PSNR than the difference introduces by a packet loss on the corresponding frame. To conclude the above consideration, we need low search depth for accurate detection of missing frames, but at the same time high depth to recover synchronization after long freeze. We managed to satisfy both contrary requirements by setting the initial search depth to 3 with possibility to temporally extend over the whole freeze. After the freeze past away and the synchronization is recovered (it is indicated by high PSNR value or simply *inf* for identical frames), search depth toggles to the initial value.

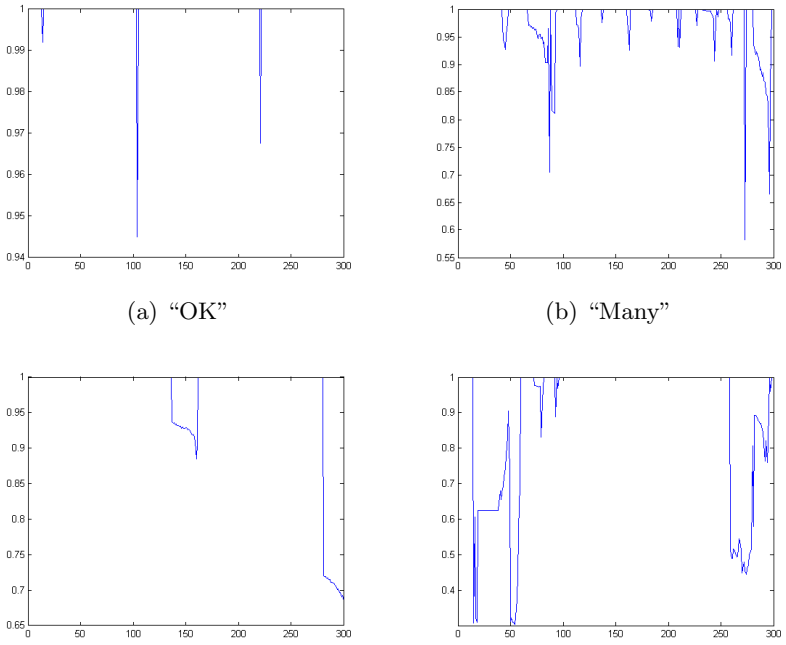
Additionally, a threshold of 2 dB was set for the PSNR values comparison to avoid false positives during detection of missing frames. This solved a problem of synchronization on frames affected with packet loss artifact. For such frames it happened very often that the missing frames were mis-detected based on a very slight differences in the PSNR value.

5 Building a Model

All the video sequences from our test set were verified, synchronized and downscaled. SSIM metric was calculated for all of them against the reference ones (streamed using the "perfect" scenario). Subjective tests with the same sequences were performed resulting in MOS values. The only remaining part to achieve our ultimate goal was to build an objective model for packet loss effect using the SSIM metric. For model building we use *Statistica* software and exact methodology is given in [27].

In the first attempt the average SSIM value calculated over all 300 frames for each sequence was considered. Figure 2(a) shows its correlation with DMOS while table 2 shows the model parameters. It obviously fails in the task of perceived quality assessment. Sequences (almost) perfect according to the average SSIM (values between 0.98 and 1) spread over almost the whole range of DMOS values. In order to propose better fitted model we decided to make a visual inspection of the SSIM plots first. Figure 1 represents the SSIM plots obtained for the streamed sequences. It shows how diverse in terms of packet loss ratio and loss pattern were our streaming scenarios. Each plot contains the SSIM values calculated for 300 video frames. Visual inspection of the plots and corresponding sequences allows to distinguish the following loss patterns: 1) Few very short (e.g. lasting for 1 frame only) artifacts affecting only small part of a video frame (relatively high SSIM values for affected frames) for sequence in figure 1(a), 2) Many artifacts a bit longer and stronger in 1(b), 3) Long artifacts (many frames affected) in 1(c), and 4) Strong and long artifact, frame freeze (very low SSIM values) in 1(d).

Based on the analysis of the SSIM plots we deduce: 1) the average SSIM is not good enough, 2) number of separate losses matters, 3) one loss but long in time also matters, and 4) strength of a loss matters. Hence, in the second attempt we proposed a model including: 1) average SSIM — AvgSSIM, 2) average SSIM calculated over 2 worst seconds — Worst(2s), 3) number of separate losses — NoLoss, and 4) count of frames with the SSIM value below 0.9 — NoF(0.9). Justification for the average SSIM is that it cannot act as a single parameters but may introduce some improvement while combined with others. The average calculated over the worst 2 seconds is intended to catch long artifacts and frame freezes. Another parameter, number of losses, is quite obvious. The more times we see an artifact the worst quality we experience. The last one corresponds to the number of frames affected with a strong artifact. Figure 2(b) shows correlation with DMOS of the proposed model while table 2 shows the model parameters.



(a) "OK" (b) "Many"

(c) "Long" (d) "Freeze"

Fig. 1. Different loss patterns represented using the SSIM values calculated for 300 video frames of streamed sequences

Among the selected parameters two show the highest statistical significance (i.e. p-value equal to 0 in table 2), namely the average SSIM calculated over the 2 worst seconds and the number of separate artifacts. In further analysis it turned that the average calculated over 1 seconds performs even better. We also noticed that the most important are changes of the SSIM value around 1 (e.g. change from 1 to 0.95 i much more significant in terms of the quality than from 0.9 to 0.8). In order to account for it an improved average SSIM over one second was calculated, as depicted in equation 2, where Worst(1s) is the SSIM average from the worst second.

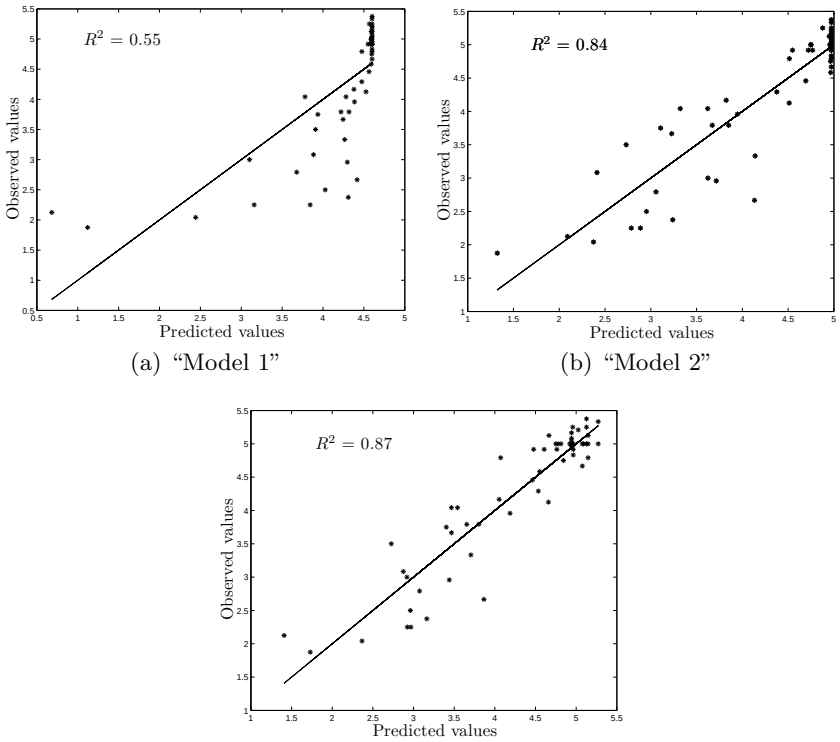
$$WorstSq(1s) = \sqrt{1 - Worst(1s)} \tag{2}$$

Next we decided to simplify our model in order to make it more generic and eliminate possible over-fitting to the data. The third model (the final one) consist of two parameters. In order to eliminate an influence of diverse video content we used spatial and temporal characteristics calculated previously (see table 1 for results). In statistical analysis temporal activity was removed as being insignificant (p-value higher than 0.05, according to [28]). Therefore, our final model includes spatial activity also SA. The model is given by equation 3, where D_p is DMOS value predicted by the model. Figure **??** shows its correlation with DMOS. The ability to estimate the perceived quality is very high and by a reasonable reduction of the input parameters we achieved more generic one.

$$\begin{aligned} D_p = &-5.10 * WorstSq(1s) - 0.077 * NoLoss \\ &+0.0031 * SA(1s) + 4.65 \end{aligned} \tag{3}$$

Table 2. Parameters of the models

Parameter	**Factor**	**p-value**
Model 1		
Intercept	-31.1491	0.000000
Avg(SSIM)	35.7490	0.000000
Model 2		
Intercept	38.7208	0.001798
Avg(SSIM)	-46.0408	0.000699
Worst(2s)	12.2914	0.000000
NoLoss	-0.2026	0.000000
NoF(0.9)	-8.1995	0.018188
Model 3		
Intercept	4.64973	0.000000
WorstSq(1s)	-5.09941	0.000000
NoLoss	-0.07747	0.000028
SA(1s)	0.0030831	0.018266



(a) "Model 1" (b) "Model 2"

(c) "Model 3"

Fig. 2. Correlation of the models with DMOS

6 Conclusions

In the paper we describe in detail how to build a model for packet loss effect on Full HD video content. Important aspects of video pool selection and the subjective experiment design were discussed. We pointed out some problems related to the streaming HD videos using non professional setup. We also explained why the average SSIM calculated over all video frames is not the best quality estimator. We present step-by-step model evolution from the simples one to the final one. Our final model is generic and shows high correlation with the subjective results across diverse content characteristics and network loss patterns. It was achieved by applying proper temporal pooling strategy and considering content characteristics. In the future work the proposed final model will be verified upon another video test set affected with packet loss artifacts.

Acknowledgment

The work presented in this paper was supported by the European Commission under the Grant No. FP7-218086 and also by the Polish State Ministry of Science and Higher Education under Grant No. N N517 4388 33.

References

1. Greengrass, J., Evans, J., Begen, A.C.: Not all packets are equal, part 2: The impact of network packet loss on video quality. IEEE Internet Computing **13**(2) (March 2009) 74–82
2. Verscheure, O., Frossard, P., Hamdi, M.: User-oriented QoS Analysis in MPEG-2 Delivery. Journal of Real-Time Imaging (special issue on Real-Time Digital Video over Multimedia Networks) **5**(5) (October 1999) 305314
3. Shengke, Q., Huaxia, R., Le, Z.: No-reference Perceptual Quality Assessment for Streaming Video Based on Simple End-to-end Network Measures. International conference on Networking and Services, ICNS '06 (2006) 53–53
4. Lopez, D., Gonzalez, F., Bellido, L., Alonso, A.: Adaptive Multimedia Streaming over IP Based on Customer-Oriented Metrics. ISCN06 Bogazici University, Bebek Campus, Istanbul (June 16 2006)
5. Liang, Y., Apostolopoulos, J., Girod, B.: Analysis of packet loss for compressed video: Effect of burst losses and correlation between error frames. IEEE Transactions on Circuits and Systems for Video Technology **18**(7) (July 2008) 861 – 874
6. Dosselmann, R., Yang, X.D.: A Prototype No-Reference Video Quality System. Fourth Canadian Conference on Computer and Robot Vision, CRV '07 **2007** (May) 411–417
7. Pinson, M., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Trans. on Broadcasting **50**(3) (Sept. 2004) 312–322
8. Wolf, S., Pinson, M.H.: Application of the ntia general video quality metric (vqm) to hdtv quality monitoring. In: Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-07), Scottsdale, Arizona (January 25-26 2007)

9. Issa, O., Li, W., Liu, H., Speranza, F., Renaud, R.: Quality assessment of high definition tv distribution over ip networks. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (13-15 May 2009) 1–6
10. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing **13**(4) (April 2004) 600–612
11. Garcia, M., Raake, A., List, P.: Towards content-related features for parametric video quality prediction of iptv services. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008 (April 2008) 757 – 760
12. Wang, Z., Li, Q.: Video quality assessment using a statistical model of human visual speed perception. Journal of the Optical Society of America A **24**(12) (December 2007) B61–B69
13. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? - a new look at signal fidelity measures. IEEE Signal Processing Magazine **26**(1) (January 2009) 98–117
14. VQEG: VQEG HDTV TIA Source Test Sequences. `ftp://vqeg.its.bldrdoc.gov/HDTV/NTIA_source/`.
15. VQEG: The Video Quality Experts Group. `http://www.vqeg.org/`.
16. Webster, A.A., Jones, C.T., Pinson, M.H., Voran, S.D., Wolf, S.: An objective video quality assessment system based on human perception. In: SPIE Human Vision, Visual Processing, and Digital Display IV. (1993) 15–26
17. Fenimore, C., Libert, J., Wolf, S.: Perceptual effects of noise in digital video compression. In: 14th SMPTE Technical Conference, Pasadena, CA (October 1998) 28–31
18. VQEG: Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment. (March 2000) `http://www.vqeg.org/`.
19. Wang, Z., Lu, L., Bovik, A.C.: Video Quality Assessment Based on Structural Distortion Measurement. Signal Processing: Image Communication **19**(2) (2004) 121–13
20. Wang, Z.: Rate Scalable Foveated Image and Video Communications. PhD thesis, Dept. Elect. Comput. Eng. Univ. Texas at Austin, Austin, TX (December 2001)
21. Wang, Z., Bovik, A.C., Lu, L.: Why is Image Quality Assessment so Difficult. in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing **4** (May 2002) 3313–3316
22. VQEG: Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content. (2009)
23. ITU-T: Subjective Video Quality Assessment Methods for Multimedia Applications. ITU-T. (1999)
24. ITU-T: Methods for subjective determination of transmission quality. ITU-T, Geneva, Switzerland. (1996)
25. : Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500 (2000)
26. et al., Z.W.: The SSIM Index for Image Quality Assessment. (2003) `http://www.cns.nyu.edu/~zwang/`.
27. Janowski, L., Papir, Z.: Modeling subjective tests of quality of experience with a generalized linear model. In: First International Workshop on Quality of Multimedia Experience, California, San Diego (July 2009)
28. : NIST/SEMATECH e-Handbook of Statistical Methods. (2002) `http://www.itl.nist.gov/div898/handbook`.