IMPROVING THE VERIFICATION PROCESS OF VIDEO QUALITY METRICS

Christian Keimel, Tobias Oelbaum and Klaus Diepold

Technische Universität München, Institute for Data Processing, Arcisstr. 21, 80333, Munich, Germany
christian.keimel@tum.de, tobias.oelbaum@mytum.de, kldi@tum.de

ABSTRACT

The most important step in the development process of a video quality metric is its verification with regards to the subjective quality experience. Even though guidelines in the form of standards and recommendations are well known, there are still quite often shortcomings in the verification process of many metrics. In this contribution we revisit these rules, point out important details and review contributions to video quality metrics for typical shortcomings. We will highlight in detail five steps that should be followed in order to improve the overall quality of the verification process of video quality metrics: using a large and diverse data base, planing and conducting subjective tests carefully, using different data for calibration and verification of a metric, avoiding unnecessary data fitting steps, and a clear and meaningful presentation of the results. Also we will provide short examples how an improper verification might affect the results of a video quality metric.

Index Terms— Visual quality, video quality metric, cross verification

1. INTRODUCTION

Video quality metrics (VQM) have been a subject of research for quite some time. A remarkable number of different metrics have been proposed so far that aim at producing results similar to the perception of human observers. Still, the problem of predicting the quality of distorted video well enough has not been completely solved and only few VQMs are used in real-life applications and products.

We can only verify VQMs by using data gained from subjective tests, as a mathematical proof is not possible. How this should be done is well known and has been intensively studied by the ITU and also the Video Quality Experts Group (VQEG). Guidelines are given in the form of standards or recommendations. The most prominent one is ITU-R BT.500 [1], others are [2–5]. Unfortunately, the verification for many VQMs does not comply completely with the recommendations outlined in these documents. The main reason is certainly the high effort that is needed for such a verification, another that people developing VQMs do not necessarily have experience in conducting subjective experiments and thus are unable to produce trustworthy data. If data bases with results from subjective tests were publicly available, this problem could be solved. So far, however, the only data base publicly available is the rather old VQEG Phase I data base [6] for standard definition TV (SDTV). Also no data base for current coding technologies is available up to now. This insufficient verification may very well be one reason for the low acceptance of the VQMs, as it does not allow to determine the real prediction capabilities of a metric. This also makes it nearly impossible to compare the accuracy of different metrics only by analyzing literature.

In this contribution we highlight the most important steps in the verification process in order to gain a solid verification of VQMs. These steps are: using a large and diverse data base, planing and conducting subjective tests carefully, using different data for calibration and verification of a metric, avoiding unnecessary data fitting steps, and finally a clear and meaningful presentation of the results. We will discuss common errors in the verification process, and explain possible pitfalls in each step by providing examples of our own or from other contributions. We will see that the most common problems are insufficient data bases, consisting of only very few videos, subjective tests, not done properly and therefore not delivering valid results, and using the same data for calibration and verification of a metric. Also the presentation of the results may not be clear enough.

The remainder of this contribution is organized as follows: in the sections 2 to 6 we highlight the steps of a proper verification process, but also describe often made mistakes and how they affect the verification before we conclude in section 7.

2. VERIFICATION DATA BASE

The data base used for the verification of a VQM consists of a set of source video sequences that are processed using a set of hypothetical reference circuits (HRC) and should be large enough. The HRCs can be realized using video encoders, (error prone) transmission systems, or even by adding a artificial distortion e.g. noise to the video. Thus we can increase the size of the data base by either increasing the number of source videos or HRCs.

2.1. Verification video sequences and HRCs

During the VQEG Phase I tests, 20 different source video sequences were used [6], for the VQEG Phase II tests, 26 different videos were used [7]. For testing the coding efficiency of a newly developed video codec in a verification test, the Moving Picture Experts Group (MPEG) used 15 different videos for the verification of SVC [8], and 14 different videos for the verification of AVC [9]. Using less than ten different videos for the verification probably results in a verification set that includes only a very limited subset of all possible videos. Thus the generality of the presented results is limited. The used videos should cover a wide range of content. They should have different properties and represent different levels of detail, motion and color.

The HRCs should also represent real world scenarios. To this end, we can use typical settings for video encoders in practical applications e.g. broadcasting or video conferencing. Also one could use a simulated transmission system including video encoders, transmission channels, transcoders etc. Using HRCs which are generated by adding artificial distortions, e.g noise or blur, to the video may be a good approach during the the development of VQMs. But results based on such HRCs are probably of limited use for the evaluation

 Digital Object Identifier: 10.1109/QOMEX.2009.5246966

Table 1: Pearson correlation coefficients for small data bases

	Data points	PSNR	Bit rate
All data	56	0.67	0.64
Single videos	3-7	0.87 - 1.0	0.73 - 0.98
Subset 1[(a)]	24	0.95	0.70
Subset 2[(b)]	19	0.60	0.82

[(a)] City, Crew, Football, Foreman, Head

[(b)] Bus, City, Crew, Harbour, Husky, Ice

of the capabilities of a VQM in a real world scenario. Thus such HRCs should not be used for the final verification of a VQM. The HRCs should also be diverse enough to show that the VQM is independent of a certain technology or codec. Although it may be sensible to build VQMs that are desigend for certain video codecs, they should still be verified using different encoders or at least different encoder settings. This should include different quality levels, prediction structures or rate control strategies. Varying the bit rate or frame rate only is not enough, if only one encoder is used, especially if the VQM uses bit stream parameters to calculate the visual quality of the video. The HRCs should be chosen to deliver a broad range of quality levels, ranging from bad to very good visual quality. Preferably, the resulting data points should be equally distributed across the whole quality range.

2.2. Too small data bases

Often the verification data base is far too small to draw reasonable conclusions. Only one single source video is used in [10]. In [11–13] results are presented for two videos only, three different videos are used in [14] . [15] uses five very similar videos, all showing some outdoor nature scenes. In [16–18] it is proposed to use different models for different content. However, only five different videos are used to verify five different models for five different content classes.

Only very few contributions present results using different encoders or at least different encoder settings apart from varying bit rate. This is especially critical for [16–23], as these contributions use the bit rate or other bit stream parameters without testing different prediction structures or encoders.

We can demonstrate the problem of using only a very small data base by using the subjective data gained in [9, 24]. Combining the CIF datasets from these two tests results in 13 different videos and 56 different data points. The overall Pearson linear correlation between the PSNR and subjective ratings is 0.67. If we select only one video, the prediction accuracy for PSNR varies between 0.87 and 1.0. Even if we select set of five different videos, the prediction accuracy of PSNR can be as high as 0.95 (see Table 1). This shows that the real prediction accuracy can be far below compared to what can be achieved with a very small data base. The same can be seen if we use bit rate as a quality predictor. Here, the overall correlation is 0.64, for single source videos the prediction accuracy varies between 0.73 and 0.98 and if six different videos are selected, a Pearson correlation coefficient of 0.82 can be achieved, using bit rate as the only quality measurement. Of course the selected subsets are hand-picked, but the numbers show that if we do not use a data base that is large enough, this may result in overestimating the prediction accuracy of a VQM.

3. SUBJECTIVE TESTING

A first step to test the design of a VQM could be an objective evaluation that basically checks, if the output of the metric changes when different encoding parameters are used. However, a final verification can only be done using the results from subjective tests. Therefore the subjective tests must be conducted very carefully, and the test documentation should contain all information needed to reproduce the tests. Subjective testing is not an easy process and requires accurate execution of every single step, as otherwise the subjective results can be meaningless.

3.1. Test methodology

Initially one has to decide whether to use a double stimulus (DS) method that uses a explicit reference, or a single stimulus method (SS) that only shows the distorted videos. The decision which test method should be used depends on the application area, the quality level of the videos, and the number of different data points to be evaluated. DS tests are thought to be less sensitive for contextual effects and are preferred, if high quality videos should be evaluated. SS methods are preferred, if no common reference exists, or if videos at a comparably low quality should be evaluated. The use of a SS method in the case of comparably low quality videos is motivated by the idea that showing a high quality reference, all distorted videos will be perceived as equally bad, and no distinction between different levels of low quality will be made by the observers. The used continuous or discrete rating scale should be detailed enough to allow discrimination between small quality differences, and easy enough to be used in a meaningful way. Recent tests within MPEG made extensive use of a discrete scale with eleven grades ranging from 0 to 10, which provided stable results.

Next we have to consider the used equipment and viewing conditions. Calibrated equipment and a well defined testing environment obviously deliver more accurate and reproducible results, than if consumer grade equipment, arbitrary lighting and variable viewing conditions were to be used. ITU-R BT.500 [1] specifies not only the viewing conditions, including lighting conditions, background color, viewing distance, viewing angle and more, but at least for SDTV also the monitors to be used. Clearly, other monitors than professional grade interlaced TV monitors may be used for subjective testing, especially if videos at different resolutions and scanning systems than SDTV are to be evaluated, as suggested in ITU-T P.910 [3]. Still, the selection and calibration of the monitor or projector has to be done with much care. Displays should not be standard consumer products, but high class professional equipment. They should be color calibrated, black level and gamma settings should be correct. In [25] Tourancheau *et al.* showed, that CRT displays are still perceived to deliver a better visual quality for video due to the superior reproduction of motion when compared to LCD monitors for the case of high quality HDTV content. However, relative quality differences were constant.

Lastly, we have to take into account our test subjects. In order to reach statistical significance, at least 15 valid viewers should be available. For competitive tests, it is proposed to have at least 20 test subjects to increase the stability of the results. But if too few test subjects take part, results will possibly depend too much on one single subject. A simple check if the number of participants is sufficient can be done by dividing all participants into two equally large groups: the results for these two independent groups should be very similar. Naïve viewers are preferred compared to experts viewers, as these are usually paid for their participation in the tests, compared to

experts, as those quite often would prefer to do other things. This difference in motivation results in more reliable results if naïve viewers are employed. Due to better eye-sight younger people are preferred. All viewers have to be screened for normal visual acuity and correct color vision using standard test charts e.g. Snellen and Ishihara. In addition, the test subjects should be able to completely understand the task of the test and the comments made by the presenter during the training phase.

3.2. Conducting the tests

Each tests consists of three main phases: training of the test subjects, the actual test and processing of the results. The training phase aims at making the test subjects familiar with the test procedure, the type of video they will see, and the range of quality they can expect in the test. It is probably the most important part of the whole test. The training phase can be further split up into two parts: explanations that are given to the test subjects and a (short) training session that is similar to the actual test. The training session should not only use the same testing procedure, but also the quality of the presented videos and artifacts present in them should be similar to the quality that will appear in the test. In particular, the lowest and the highest visual quality that will occur in the actual test should be included. The test itself should be split into several sessions of not more than 25 minutes each, having equally long breaks between single sessions.

After finishing the tests, the processing begins by removing the outliers and checking if all subjects were able to carry out the test sensibly. A formal assessment of the test subjects is proposed in Annex 2 of ITU-R BT.500. Viewers that produce too many outliers votes should be removed accordingly. Outliers can also be detected by arbitrarily assigning the test subjects to different groups and comparing the results of these groups. To detect if the viewers are able to reproduce their own results, several test cases should appear more than once in one test. Viewers that are not able to assign (roughly) the same vote to the same test case at two different time instances should be removed. Finally, the votes have to be processed by calculating a mean value for each test case and the 95% confidence interval of the votes. The mean value is also known as mean opinion score (MOS) and serves as quality value that is assigned to one video. The size of the 95% confidence interval tells us how similar the votes of different subjects for one video actually were and gives a good indication about the accuracy of the tests. Whereas no general rule can be given, the number of outliers that come from subjects that are considered for the final results should be significantly below 5%. The 95% confidence intervals themselves should not be above 0.1 on a 0 to 1 range for every single test case. Additionally, an analysis of variance (ANOVA) can be done.

3.3. Subjective testing problems

Although the above mentioned steps are well known, many contributions neglect important aspects of subjective testing. Often the number of people that took part in the tests is significantly too low: [14] uses only three people for a test in a very informal environment. Other contributions, where not more than ten people were used are [16–18, 26, 27]. Also the tests are sometimes conducted too inaccurately e.g. in [28] the subjective tests are done without training the subjects and the video is captured from low quality VHS tapes.

The danger of not using standardized methods, or not using appropriate equipment becomes obvious analyzing the data from two published tests. In [29] standard interlaced TV videos are first deinterlaced and then displayed using standard PC equipment. The authors used 24 test subjects, but the reported confidence intervals are higher than what would be expected. [16–18] are all based on the same tests, where 26 test persons were used for the calibration data base, but only 10 people rated the verification videos. Instead of using high quality displays, test were done using a handheld PDA. According to [17], the percentage of removed outlier votes is above 12%. This is a clear indication, that something went wrong either in the design of the test, or in conducting the test itself.

One additional problem, that almost all contributions do have in common, is that very little information about the subjective tests is given. Only a few contributions report the percentage of removed outliers or achieved confidence intervals, and often relevant information about room or equipment is not provided. One example where no information about the related subjective tests is given is [30].

4. CALIBRATION AND VERIFICATION

Nearly all VQMs combine parameters extracted from the videos into one quality value. Careful research suggests to use different data for calibration and verification. We have two options to calibrate or train the metric: the use of different data sets for calibration and verification or the use of the same data set for calibration and verification in conjunction with a cross validation.

4.1. Cross validation

A cross validation for four videos A, B, C and D is done by using the data points from videos A, B and C for calibrating the metric later used to verify the data points from video D and vice versa. The cross verification approach allows using a bigger data base both for calibrating and verifying the metric. But the calibration phase must now be done separately for every video. Such a cross validation step is used very rarely in the field of visual quality metrics. So far, only two contributions apart from our own contribution [31] could be identified: [32, 33].

The importance of keeping the data used for calibrating the metric separate from data used in the verification, is shown using data from [9]. A subset of this data (calibration data) is used to build a simple PSNR based FR VQM that delivers quality ratings in the range of 0 to 1. We calculate a combined PSNR using a weighted sum of the PSNR on all three color channels of the YCbCr color space. This PSNR^3 metric can be calculated according to

$$\text{PSNR}^3 = 0.99 + 0.11 * \text{PSNR}_Y - 0.21 * \text{PSNR}_{Cb} + 0.11 * \text{PSNR}_{Cr}. \quad (1)$$

According to the Pearson correlation coefficients as reported in Table 2, the new PSNR^3 metric performs significantly better than standard luminance PSNR_Y. The same was also done for a second subset of the original data (verification data). Correlation for these unknown videos is only slightly higher than for standard PSNR_Y which shows, that the PSNR^3 metric was tailored to the videos used for the calibration step, but does not provide a real benefit for other videos. Without verifying a new metric on previously unknown data, the danger of having a metric that is fitted to special videos is quite high. Unfortunately, the calibration and verification data is not always separated explicitly [10, 15, 19, 20, 22, 23, 34–42].

Very similar videos are used for [15] questioning the generality of the proposed metric and in [43] probably the same videos are used for calibration and verification, but at least only small parts of the videos are used for the calibration phase. Still, a cross validation approach would better.

Table 2: Pearson correlation coefficients for $PSNR^3$

	Calibration Data	Verification Data
$PSNR_Y$	0.644	0.684
$PSNR^3$	0.953	0.709

Table 3: Calibration and validation sets for $PSNR^3$

Test Set	Slope	Offset	Mean Absolute Error data fitting	Mean Absolute Error no data fitting
Calibration	1.00	0.00	0.05	0.05
Validation	0.70	0.22	0.10	0.13

4.2. Use of unknown videos and HRCs

In [44] Lubin pointed out, that there is a high correlation between subjective quality and objective quality for single source videos, even if only the mean squared error (MSE) is used to rate the visual quality. Due to this linear relationship, it is relatively easy to predict the visual quality for a processed video, if this source video is known. Therefore it is important to use previously unknown source videos for the verification step. The importance of using unknown source videos is also demonstrated in [28]. The authors do use different data for training and verification but compose the calibration data base in two different ways: first in excluding a set of videos and second in excluding a set of HRCs. The remaining part is then used for verification. The prediction results for the case where the source videos were unknown are significantly behind the results for unknown HRCs.

If VQMs use parameters from the bit stream, it is insufficient to only vary the bit rate to generate an unknown HRC. Instead, different encoders, or at least different prediction structures and rate control strategies should be used to generate new HRCs. In most cases different HRCs were generated only by varying the bit rate of a certain encoder using fixed settings [16–23].

5. DATA FITTING

Fitting the output of the metric to the results of the subjective tests is quite common in the field of objective visual quality evaluation, but contrasts with reality where such a fitting step is not possible. Sigmoid (or logistic) fitting was proposed in [6] and [7]. It was reasoned that subjective tests themselves do not produce results that are linear for the whole quality range. In subjective tests typically compression appears at the very ends of the quality range and the extremes of the quality scale are rarely reached. Hence the sigmoid fitting function should be more or less constant for one subjective test and that therefore different metrics should have the same sigmoid fitting function. Also differences between the fitting functions of different subjective tests should be very small. It should have the following characteristics: saturation toward the ends of the quality range with a large middle section that is close to be linear. Examining the fitting functions used for the metrics in [45] shows that only two out of eight functions are similar to the required shape. This fitting step resulted in a significantly reduced outlier ratio for all four metrics [46].

The problem of data fitting after having done the subjective tests is again shown by our $PSNR^3$ metric from section 4. For the calibration data, the first order fitting line should have no offset and a slope of 1.0, as this was the goal of the bilinear regression applied to find the weights for the three PSNR values in (1). This is achieved for the given example. In contrast, the fitting line for the set of validation videos has a offset of about 0.2 and the slope is not as steep as desired (0.7). Thus, the error between the predicted quality and the actual quality increases significantly if no final fitting step is allowed (Table 3). A special case is given in [47]. Here the fitting is done separately for each original video. This obviously increases the correlation values as we demonstrated in [48], where this step lead to a correlation value of 0.99 for PSNR to the results of subjective tests, whereas the real correlation value would be 0.67. If a fitting step is to be included in the VQM, the consequences i.e. optimistic correlation values should be pointed out e.g. [47]. Therefore we propose to not apply any data fitting, as there are no clear advantages for using such a fitting step, but the true prediction accuracy is hidden by the use of a fitting step.

6. PRESENTATION OF THE RESULTS

6.1. Statistical representation

The statistical tool used most often to demonstrate the performance of a visual quality metric is the Pearson correlation. It gives an indication about the prediction accuracy of the metric. A similar task is solved by the Spearman rank order correlation. This rank order correlation gives an indication how much the ranking between the videos under test changes for the metric's values compared to the subjective values (prediction monotonicity). Both statistical metrics should be calculated for the whole verification data set, and not for each video or each HRC separately. The results should be presented in one common plot instead of providing different plots for different source videos or HRCs. As neither these two give an indication about the absolute error between the predicted and the actual values, they are supported by the MSE between the subjective data and the objective values. Also an outlier ratio may give an indication about the accuracy of a metric. Calculation of the outlier ratio may be based on the confidence intervals of the single data points, e.g. consider every data point that does not fall into the 95% confidence interval as an outlier. For simplicity, also a fixed deviation may be allowed.

6.2. Comparison metrics

Each new method should compare itself to the state of the art. Firstly, however, it is hard in the absence of accepted standards to decide what the state of the art in video quality evaluation is. As we have shown so far, the verification for many VQMs is not completely sufficient. This makes the selection of powerful state of the art methods even more difficult. Secondly, there are only very few public available implementations of VQMs. Considering the complexity of implementing such a metric, it is not feasible to implement a metric only to generate comparison data. The metrics with public available implementations are the SSIM [49], the VQM according to [50], and the VQM developed by the NTIA [51].

Comparison from literature is also only possible to a small extend. The only available data base of videos and related subjective ratings is the comparably old VQEG Phase I data base [6]. Apart from the metrics tested there, only very few metrics were verified on the complete data base [23, 41, 52, 53].

Therefore a comparison with PSNR should be sufficient in most cases, even if the limitations of PSNR to serve as a visual quality metric are well known. PSNR still is the metric in this area most often used. Also experts in the field of visual quality metrics can judge the capabilities of a new VQM if useful results are provided for the new VQM and PSNR. PSNR may be a low quality anchor, but at least it is a known anchor. The comparison to PSNR should be supported using the other metrics where implementations are available.

7. CONCLUSION

We have reviewed the basic steps needed to adequately verify video quality metrics and discussed the important details to be considered in every step in order to provide a proper verification. While none of the proposed steps in this contribution is new in itself, we have seen that details are nevertheless quite often neglected in the verification process of many contributions. Regardless if too few video sequences are used, the subjective testing is done poorly, calibration and verification data is mixed, misleading data fitting is applied or no comparison to other metrics is given, the provided examples show the possible consequences of a insufficient verification process on the presented results.

Therefore we propose that the following five steps should be considered in order to improve the overall quality of the verification process in the research on video quality metrics:

1. Use a sufficiently large data base.
2. Use a public data base or perform extensive and good subjective tests.
3. Use different data for calibration and verification.
4. Use no subsequent data fitting.
5. Report useful and comparable results.

We strongly believe that without solid verification of new video quality metrics, only small steps toward efficient objective evaluation of video quality will be made.

8. REFERENCES

[1] *ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for Television Pictures*, ITU-R Std., Rev. 11, Jun. 2002.

[2] *ITU-R BT.1676 Methodological framework for specifying accuracy and cross-calibration of video quality metrics*, ITU-R Std., Rev. 1, Feb. 2004.

[3] *ITU-T P.910 Subjective video quality assessment methods for multimedia applications*, ITU-T Std., Rev. 1, Sep. 1999.

[4] *ITU-R BT.710 Subjective assessment methods for image quality in high-definition television*, ITU-R Std., Rev. 4, Nov. 1998.

[5] *EBU Project Group B/VIM (Video In Multimedia) EBU BPN 056: SAMVIQ Subjective Assessment Methodology for Video Quality*, EBU Std., May 2003.

[6] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," VQEG, Tech. Rep., Mar. 2000.

[7] ——, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase 2," VQEG, Tech. Rep., Aug. 2003.

[8] MPEG Test Subgroup, "SVC verification test report," ISO/IEC JTC1/SC29/WG11, Tech. Rep. N9577, Jan. 2008. [Online]. Available: http://www.chiariglione.org/mpeg/quality_tests.htm

[9] ——, "Report of the formal verification tests on AVC (ISO/IEC 14496-10 ITU-T Rec. H.264)," ISO/IEC JTC1/SC29/WG11, Tech. Rep. N6231, Dec. 2003. [Online]. Available: http://www.chiariglione.org/mpeg/quality_tests.htm

[10] M. Carli, M. C. Farias, E. Drelie Gelasca, R. Tedesco, and A. Neri, "Quality assessment using data hiding on perceptually important areas," in *Proc. IEEE International Conference on Image Processing 2005, (ICIP2005)*, vol. 3, Sep. 2005, pp. 1200–3.

[11] H.R.Wu and K.R.Rao, Eds., *Digital video image quality and perceptual coding*. CRC, 2006, ch. No-Reference Quality Metric for Degraded and Enhanced Video, pp. 305–324.

[12] A. Leontaris, P. Cosman, and A. Reibman, "Quality evaluation of motion-compensated edge artifacts in compressed video," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 943–956, Mar. 2007.

[13] M. Montenovo, A. Perot, M. Carli, P. Cicchetti, and A. Neri, "Objective quality evaluation of video services," in *Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2006.

[14] J. Hu, S. Choudhury, and J. D.Gibson, "$PSNR_{r,f}$: Assessment of delivered AVC/H.264 video quality over 802.11a WLANs with multipath fading," in *MultiComm 2006*, Jun. 2006.

[15] F. Meng, X. Jiang, H. Sun, and S. Yang, "Objective perceptual video quality measurement using a foveation-based reduced reference algorithm," in *IEEE International Conference on Multimedia and Expo*, Jul. 2007, pp. 308–311.

[16] M. Ries, O. Nemethova, and M. Rupp, "Performance evaluation of video quality estimators," in *Proc. European Signal Processing Conference EUSIPCO*, Sep. 2007, pp. 159–163.

[17] ——, "Video quality estimation for mobile H.264/AVC video streaming," *Journal of Communications*, vol. 3, no. 1, pp. 41–50, 2008.

[18] M. Ries, C. Crespi, O. Nemethova, and M. Rupp, "Content based video quality estimation for H.264/AVC video streaming," in *Proceedings of IEEE Wireless and Communications & Networking Conference*, 2007.

[19] S. Péchard, D. Barba, and P. Le Callet, "Video quality model based on a spatio-temporal features extractions for H.264-coded HDTV sequences," in *Proc. Picture Coding Symposium*, Nov. 2007.

[20] M. Ries, O. Nemethova, and M. Rupp, "Reference-free video quality metric for mobile streaming applications," in *Proceedings of the DSPCS 05 & WITSP 05*, Sunshine Coast, Australia, Dec. 2005, pp. 1–5.

[21] P. Gastaldo, G. Parodi, and R. Zunino, "DSP-based neural systems for the perceptual assessment of visual quality," in *Proc. IEEE International Conference on Neural Networks*, vol. 1, Jul. 2005, pp. 663–668.

[22] N. Suresh, O. Yang, and N. Jayant, "AVQ: A zero-reference metric for automatic measurement of the quality of visual communications," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2007.

[23] L. Lu, Z. Wang, and A. C. Bovik, "Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video," in *Proc. IEEE International Conference on Multimedia and Expo*, vol. 1, Aug. 2002, pp. 61–64.

[24] MPEG Test Subgroup, "Subjective test results for the CfP on scalable video coding technology," ISO/IEC JTC1/SC29/WG11, Tech. Rep. N6383, Apr. 2004.

[25] S. Tourancheau, P. Le Callet, K. Brunnström, and D. Barba, "Display awareness in subjective and objective video quality evaluation," in *Proc. European Signal Processing Conference EUSIPCO*, Sep. 2007, pp. 164–168.

[26] N. Jayant and N. Suresh, "Objective measurement of video quality: Prediction of mean time between failures," in *National Association of Broadcasters*, Apr. 2006.

[27] Y. Jia, W. Lin, and A. A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 820–829, Jul. 2006.

[28] T. Morris, K. Angus, R. Butt, A. Chilton, P. Dettman, and S. McCoy, "CQA - subjective video codec quality analyser," in *BMVC99*, 1999.

[29] J. Okamoto, T. Hayashi, A. Takahashi, and T. Kurita, "Proposal for an objective video quality assessment method that takes temporal and spatial information into consideration," *Electronics and Communications in Japan*, vol. 89, no. 12, pp. 97–108, Jun. 2006.

[30] P. Ndjiki-Nya, M. Barrado, and T. Wiegand, "Efficient full-reference assessment of image and video quality," in *Proc. IEEE International Conference on Image Processing*, vol. 2, Sep. 2007, pp. 125–128.

[31] T. Oelbaum and K. Diepold, "A reduced reference video quality metric for AVC/H.264," in *Proc. European Signal Processing Conference EUSIPCO*, Sep. 2007, pp. 1265–1269.

[32] P. Le Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.

[33] P. Le Callet, C. Viard-Gaudin, S. Péchard, and E. Caillault, "No reference and reduced reference video quality metrics for end to end QoS monitoring," *IEICE Communication Transactions*, vol. E89B, no. 2, pp. 289–296, Feb. 2006.

[34] M. Masry and S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Elsevier, Signal Processing, Image Communication*, vol. 19, no. 2, pp. 133–146, Feb. 2004.

[35] E. Ong, W. Lin, Z. Lu, and S. Yao, "Colour perceptual video quality metric," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Sep. 2005, pp. III: 1172–1175.

[36] H.-H. Ho, T. Wolff, M. Salatino, J. M. Foley, S. K. Mitra, T. Yamada, and H. Harasaki, "An investigation on the subjective quality of H.264 compressed/decompressed videos," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2007.

[37] R. R. Pastrana-Vidal and J.-C. Gicquel, "A no-reference video quality metric based on a human assessment model," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2007.

[38] R. Feghali, F. Speranza, D. Wang, and A. Vincent, "Video quality metric for bit rate control via joint adjustment of quantization and frame rate," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 441–446, Mar. 2007.

[39] Q. Li and Z. Wang, "Video quality assessment by incorporating a motion perception model," in *Proc. IEEE International Conference on Image Processing*, Sep. 2007.

[40] M. C. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," in *Proc. IEEE International Conference on Image Processing*, vol. 3, Sep. 2005, pp. 141–144.

[41] I. P. Gunawan and M. Ghanbari, "An efficient reduced-reference video quality metric," in *Proc. Picture Coding Symposium*, Nov. 2007.

[42] E. Ong, W. Lin, Z. Lu, S. Yao, and M. Loke, "Perceptual quality metric for H.264 low bit rate videos," in *Proc. IEEE International Conference on Multimedia and Expo*, May 2006, pp. 677–680.

[43] M. Masry, S. S.Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 260–273, Feb. 2006.

[44] J. Lubin, "A human vision system model for objective picture quality measurements," in *Proc. IBC International Broadcasting Conference*, Sep. 1997, pp. 498–503.

[45] *ITU-T J.144. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*, ITU-T Std., Mar. 2004.

[46] T. Oelbaum, "Design and verification of video quality metrics," Ph.D. dissertation, TU-München, Apr. 2008. [Online]. Available: http://www.ldv.ei.tum.de/forschung/publikationen

[47] K. Seshadrinathan and A. C. Bovik, "An information theoretic video quality metric based on motion models," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2007.

[48] T. Oelbaum, K. Diepold, and W. Zia, "A generic method to increase the prediction accuracy of visual quality metrics," in *Proc. Picture Coding Symposium*, Nov. 2007.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[50] F. Xiao. DCT-based video quality evaluation. [Online]. Available: http://compression.ru/video/quality_measure/vqm.pdf

[51] M. H. Pinson and S. Wolf, "Application of the NTIA general video quality metric (VQM) to HDTV quality monitoring," NTIA, Tech. Rep., 2007.

[52] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[53] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2005.