Evaluation of Full-Reference Objective Video Quality Metrics on High Efficiency Video Coding

Glenn Van Wallendael, Sebastiaan Van Leuven,
Jan De Cock, Peter Lambert, Rik Van de Walle
Ghent University - iMinds, ELIS Department - Multimedia Lab,
Ghent, Belgium
Email: glenn.vanwallendael@ugent.be

Nicolas Staelens, Piet Demeester
Ghent University - iMinds, INTEC Department - IBCN,
Ghent, Belgium

***Abstract*—For the purpose of automatic video quality evaluation, Peak Signal-to-Noise Ratio (PSNR) has been the most well-known full reference quality metric since a long time. Improving on PSNR, several other pixel-based quality metrics have been developed, namely Structural SIMilarity (SSIM), Multi Scale-SSIM (MS-SSIM), and Video Quality Metric (VQM). The goal of these objective video quality metrics is to replace time-consuming and expensive subjective quality assessment experiments. These alternative objective metrics have already been evaluated on several video compression schemes, such as MPEG-2 and H.264/AVC, transported over different kinds of network protocols and under a large variation of network characteristics. In January 2013, the successor of the Advanced Video Coding (H.264/AVC) standard, named High Efficiency Video Coding (HEVC), has been finalized. Although HEVC is still a block based hybrid video compression standard, some radical changes are made to subjectively improve the compression efficiency compared to H.264/AVC. Until now, the alternative quality metrics have never been evaluated on this new compression scheme. Therefore, in this paper, we analyze the difference in performance of these full reference metrics. Based on subjective evaluations, a performance analysis is presented which shows the validity of these models when applied to HEVC compressed video content.**

I. Introduction

When offering a digital video service on the market, end users expect a certain quality [1], [2]. For IPTV, this is not straightforward because the IP network, over which the video is transported, is a packet-based best-effort network. Not only IPTV has to cope with these problems, any type of video distribution has to consider the loss and delay characteristics of its network. A small delay or drop of a certain packet of the compressed video stream can have a significant visual impact on the video depending on the type of packet that gets lost or delayed [3], [4]. Additional to the network characteristics, the type of video compression and the properties of the video stream highly influence the visual effect of network errors [5], [6]. In general, consumers expect less than one visual artifact per hour [7]. Therefore, in order to guarantee customer satisfaction, the video quality and the related Quality of Experience (QoE) should be monitored all the time [8], [9]. The best way to monitor the QoE is by human inspection, but evidently this is time consuming, costly, and most of the time impractical. Because of all these restrictions of subjective evaluation, quality metrics are being developed. The quality metrics considered in this paper are Peak Signal to Noise Ratio (PSNR), Structural SIMilarity (SSIM), Multi Scale-SSIM (MS-SSIM), and Video Quality Metric (VQM) as described in Section III. These quality metrics have been evaluated on a large number of compression standards and impairment scenarios, but never on the recently standardized High Efficiency Video Coding (HEVC) standard. With a new compression scheme, different artifacts are introduced when packet loss is introduced. Therefore, in this paper, the behavior of these quality metrics is evaluated on a packet loss environment transporting HEVC instead of H.264/AVC.

HEVC and its main differences with H.264/AVC are described in section II, followed by a description of the different quality metrics. Then, the methodology, as followed in this paper, is described in Section IV. The methodology section starts with a description of the source sequences on which the metrics are evaluated in Section IV-A. Afterwards, Section IV-B describes which compression parameters and loss parameters are applied on these sequences. This is also called the Hypothetical Reference Circuit (HRC). Now that quality degraded video streams are obtained, the procedure followed for subjective quality evaluation is described in Section IV-C. Next, a post processing step on the obtained subjective scores must be performed together with the calculation of the correlation between the quality metrics and the subjective scores as described in Section IV-D. Finally, the results and a conclusion are given in Section V and Section VI respectively.

II. High Efficiency Video Coding

HEVC [10] is a video compression standard realized by a joint collaboration between ISO/IEC and ITU-T and is supposed to be the successor of their earlier defined H.264/AVC [11] standard. With HEVC, only half the bitrate is needed compared to H.264/AVC in order to offer similar subjective quality [12].

Both these standards are block based video compression techniques implying that the pictures are divided in blocks which are used for intra or inter prediction. With intra prediction samples surrounding the block are used to predict the block. Consequently, when an error is introduced in samples surrounding the current block, then the error will drift further in this block. With inter prediction, blocks predict their sample values from previously decoded pictures. Similarly, when an error was introduced in the referenced picture, the error will drift further if the current block predicts from it.

The main difference between H.264/AVC and HEVC is the possibility for increased block sizes. In H.264/AVC, only sizes

978-3-901882-50-0

4x4 up to 16x16 could be used. With HEVC, the maximum block size got increased up to 64x64 for inter prediction. Consequently, when an error occurs in the video stream, this error can get propagated with blocks of up to 64x64 in size resulting in different artifacts compared to H.264/AVC.

III. Considered quality metrics

First of all, PSNR is considered because it is the most applied quality metric in the field of video quality evaluation. The formula for PSNR is as follows:

$$\text{PSNR} = 10 \log_{10} \frac{\text{MAX}^2}{\text{MSE}} \quad (1)$$

with MAX being equal to the maximum sample value (255 for 8 bit component bit depth) and MSE being the mean square error between the sample values of the compared video frames.

With this detailed explanation, we would like to show that PSNR is a frame based method, i.e. there is no notion of block partitioning for any calculation in the metric. The entire frame is considered in its entirety. Similarly, SSIM [13] is also calculated on the entire frame.

For MS-SSIM [14], the SSIM is calculated on the original and four resized versions. These resized versions half the resolution with every iteration. Then, depending on the resolution, a weighted sum is taken from these SSIMs resulting in the MS-SSIM. Because measurements are made on different sizes of the frame, the different block structure of HEVC could have an impact on MS-SSIM measurement behavior.

For VQM, the general model as described in [15] is applied. With the general model, different quality indicators are combined in a weighted way in order to come up with a single number for the entire video sequence. VQM is the most advanced of the different models because it can compensate for temporal shifts, spatial shifts, contrast changes, and brightness changes. Additionally, it can take into account interlacing differences and certain regions to evaluate. Because the other models are not functioning under these effects, they are not being applied to the video sequences. As a quality evaluation technique, the VQM general model is configured to calculate its quality indicators on block sizes of 4x4 and 8x8 samples. With a change to larger block sizes in HEVC, this could impact the performance of VQM.

IV. Methodology

A. Sequences

When evaluating quality metrics through subjective evaluation, it is important to have enough variety in video content because the type of video influences visual effects of the introduced degradations [16]. More specifically, it is the amount of motion and spatial details that affect the visibility of these degradations [17]. In ITU Recommendation P.910 [18], two measures are offered to describe these properties, namely spatial perceptual information measurement (SI) and temporal perceptual information (TI).

SI can be calculated as the maximum value of the standard deviation of the sobel filtered frame at time n [18]:

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\}. \quad (2)$$

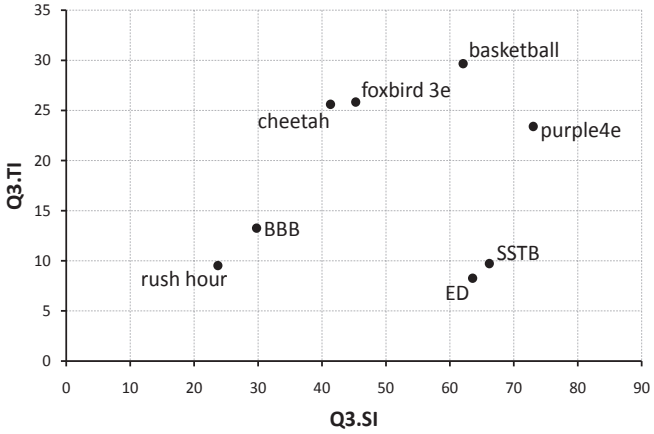

Fig. 1. Calculated Q3.SI and Q3.TI values for our eight selected source sequences.

TABLE I. Short description of test sequences

Sequence	Description
basketball	Camera pans and zooms to follow the action.
BBB	Big Buck Bunny. Computer-Generated Imagery.
cheetah	Camera pans to follow the cheetah.
ED	Elephants Dream. Computer-Generated Imagery.
foxbird3e	Cartoon. Fast camera pan with zoom.
purple4e	Many small objects moving in a circular pattern.
rush hour	High depth of focus. Fixed camera.
SSTB	Sita Sings the Blues. Slight camera zoom in.

For the TI, the maximum value of the standard deviation of the difference of a certain frame with its previous frame is measured [18].

$$TI = max_{time}\{std_{space}[F_n(i,j) - F_{n-1}(i,j)]\}, \quad (3)$$

In [19], it is noted that taking the maximum value in these formulae can result in misleading numbers when local peaks occur. As an alternative, the third quartile value (Q3) is proposed instead of the maximum value resulting in the following measures: $Q3.SI$ and $Q3.TI$.

To have a representative subjective test, it is important that there is enough variation in $Q3.SI$ and $Q3.TI$ values. Therefore, these measures are calculated on test sequences from the Consumer Digital Video Library (CDVL) [20], the Technical University of Munich (TUM), and open source movies. From these databases, eight sequences are selected with $Q3.SI$ and $Q3.TI$ values as indicated in Figure 1. Additionally, a short description of the sequences is given Table I. All the selected sequences are 1920x1080 pixels in resolution and were shot at 25 frames per second. The original sequences were trimmed to 10 seconds to have consistent evaluation time.

B. Hypothetical reference circuit

In this paper the behavior of quality metrics on H.264/AVC and HEVC video content is evaluated. Consequently we selected encoder and decoder software for each of these compression standards. For H.264/AVC, the Joint reference Model JM 16.1 [21] was selected and modified such that any combination of picture losses could be concealed. As an HEVC codec, the HEVC reference Model HM v4.0 [22] was selected. The error

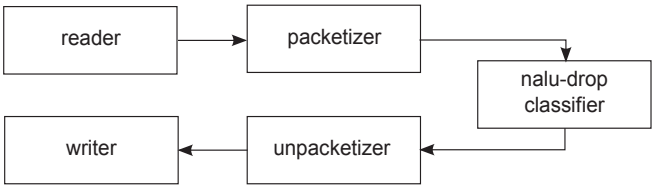

Fig. 2. Sirannon configuration. The raw video stream gets packed in RTP-packets after which losses can be applied by the *nalu-drop classifier* component. Afterwards, the impaired sequence is extracted from the RTP stream and saved to a new file.

concealment method applied in both decoders is the frame copy algorithm.

Next, the configuration of the video streams will be described. To describe which coding structure is used, we define a Group Of Pictures (GOP) as all the pictures between I-frames. Research indicates that the GOP size is typically between 12 and 15 frames in an IPTV environment [23]. Therefore, in our test, intra pictures are inserted every 15^{th} or 16^{th} picture depending on the B-picture configuration.

Three different intervals of B-pictures are selected in this test. In the first configuration, no additional B-pictures are inserted in the video stream. Therefore, the resulting GOP structure would look like this: IPPPPPPPPPPPPP. The second configuration is characterized by a single B-picture between every P-picture (IPBPBPBPBPBPBPB). Finally, as a third variation, between every P-picture, three B-pictures got inserted resulting in the following GOP structure: IPBBBPBBBPBBBPBBB. For all the configurations, only a single slice is encoded for every picture in the video stream. The primary purpose of this test is to evaluate artifacts resulting from network impairments. Therefore, encoding bitrates of the video stream are set high enough such that no compression artifacts could be observed.

After creation of video streams, network impairments should be introduced. This is done by means of the open-source media streamer called Sirannon [24]. The chain with which Sirannon got configured is depicted in Figure 2. With this configuration, RTP packetization is applied on the raw H.264/AVC Annex B bitstream according to RFC3984 [25]. For HEVC, the raw HEVC Annex B bitstream got equivalently packetized, but since an official standard for RTP packetization was not yet finalized, the procedure for H.264/AVC packetization got mimicked. During the RTP packetization, no aggregation is applied. Consequently, an RTP packet does not contain data from more than one encoded picture.

In the *nalu-drop classifier*, RTP packets are discarded from the RTP stream. When a packet gets discarded, all RTP packets containing information from the same picture get discarded as well. Additionally, a restriction is applied to not discard packets from the first and last two seconds of the video stream. This to avoid confusion between the transition to the start of the video or the end and packet loss artifacts.

The impaired video stream then gets umpacketized resulting in an impaired H.264/AVC Annex B compliant bitstream or an HEVC Annex B bitstream. These bitstreams are saved to a file, decoded, and played back for subjective evaluation.

With the packet loss chain described, the only parameters remaining to be clarified are the loss scenarios. Because every loss scenario applied on every sequence should be subjected to viewing, a limited set of loss scenarios is created. In total, 13 loss scenarios are created resulting in a total of 104 impaired sequences to be subjectively evaluated.

TABLE II. LOSS SCENARIOS WITH FIXED CONFIGURATION.

GOP structure	lost frame	position in GOP
IPPPP	I	BEGIN
IPPPP	2P	BEGIN
IPPPP	2P	MIDDLE
IPPPP	P	END
IPPPP	2P	END
IPBPB	I	BEGIN
IPBPB	P	END
IPBPB	B	END
IPBBB	I	BEGIN
IPBBB	2B	END

TABLE III. THREE FLEXIBLE LOSS SCENARIOS.

GOP structure	lost frame	position in GOP
IPBBB	P,B	BEGIN
IPBBB	2P,2B	MIDDLE
IPBBB	2P,2B	END

From these 13 loss scenarios, 10 are created with a fixed configuration (see Table II) and three scenarios are left more flexible (see Table III).

The following parameters are used for the different loss scenarios and explicitly specified in these tables:

- Number of B-pictures between two reference pictures (0, 1 or 3)
- Type of picture in which the first loss is inserted (I, P or B)
- Number of entire picture drops (1 or 2)
- Location within the GOP where the loss is inserted (begin, middle or end)

In the configurations of Table III, randomly one or two P- or B-pictures are removed. This to increase the variety in loss configurations.

C. Subjective evaluation

After selecting sequences, applying compression, and introducing network losses, the impaired sequences are evaluated subjectively on their quality.

In our test, subjective evaluation follows a Single Stimulus (SS) Absolute Category Rating approach with Hidden Reference (ACR-HR) [18].

An Ishihara test and a Snellen chart are performed in order to verify the participants color vision and visual acuity respectively. Then, the participant is made aware of possible artifacts that can be noticed by presenting three sequences from the full set, namely an original one, a very distorted one, and an average distorted one. The subjective evaluation is following the single stimulus approach because from then on no comparison with the original sequence can be made anymore. For every dataset exactly 24 valid participants evaluated the quality

by indicating the quality on an absolute category scale from one to five with the following naming:

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

With the Hidden Reference approach, the original sequence is included in the test set of sequences such that they get rated by the participants without knowing this is the original. Otherwise, when some participants would rate a certain original sequence as *4 Good* instead of *5 Excellent*, then this could not be reproduced by the objective model. A model is certainly not able to reproduce this result when it is a full reference model because such models would get the original as ground truth. Lower ratings of the original sequence can occur when the quality of the original is not good enough or when the content is not appealing for the participant. By rating the original too, the difference with the original can be taken as the effective quality score of a certain participant.

To keep the participants attention, the time for subjective evaluation is restricted to 20 minutes. To accomplish this, the test set of impaired sequences is split into four datasets. The subjective evaluation took place by a single participant sitting in front of a 40 inch full HD LCD screen at four times the display height.

Validity of the subjective scores is checked by a post-experiment screening as described in annex V of the Video Quality Experts Group (VQEG) HDTV report.

D. Evaluation procedure

As described in Section IV-C as the Hidden Reference approach, first the difference of the quality score with the score of the original is taken. This difference is called the Differential Mean Opinion Score (DMOS) from here on and is calculated as follows:

$$DMOS = MOS(\text{degraded seq.}) - MOS(\text{original}) + 5 \quad (4)$$

In this equation, MOS(x) represents the quality score that is given to a certain sequence x.

The next variation where is compensated for is the fact that all participants have a different internal weight for rating certain artifacts. Some optimistic participants would only use the upper range of the scale, while the more pessimistic viewers would rate more to the lower end of the scale. To compensate for this phenomenon, a z-score is calculated from every given rating. The z-score is a normalization of every participants score.

A further mapping that is made is between the metrics output and the DMOS scores. It cannot be assumed that an objective model outputs DMOS scores without further modification needed. Because every model outputs data over different scales with a different distribution, a non-linear mapping is applied. Empirically, it is observed that a cubic polynomial mapping as described in equation 5 performs well [26].

$$DMOS_p = ax^3 + bx^2 + cx + d \quad (5)$$

In this equation, $DMOS_p$ is the predicted DMOS after cubic polynomial mapping of the quality metrics output x. The weighting factors a, b and c and the constant d result from a fitting operation over the DMOS and model output data.

From the MOS of the participants, a DMOS score is generated after the HR compensation and z-score calculation. The output of the objective models is mapped to this data such that the numbers can be compared on the same scale. These numbers are identified as $DMOS_p$. Now that both numbers are on the same scale, a pearson correlation and Root Mean Square Error (RMSE) comparison can be made in order to measure the linear relationship and the accuracy of the model respectively. Along with those measurements, the 95% confidence intervals are calculated to check for significant difference.

TABLE IV. RMSE AND 95% CONFIDENCE INTERVAL OF DIFFERENT OBJECTIVE MODELS FOR THE BOTH THE H.264/AVC AND THE HEVC TESTSET.

		PSNR	SSIM	MS-SSIM	VQM
H.264/AVC	upper	0.88	1.05	1.20	0.74
		0.77	0.92	1.05	0.65
	lower	0.69	0.82	0.93	0.58
HEVC	upper	0.71	0.88	1.06	0.70
		0.62	0.77	0.93	0.62
	lower	0.55	0.69	0.82	0.55

V. RESULTS

RMSE and 95% confidence intervals of the different metrics on the H.264/AVC and HEVC content are enumerated in Table IV and visualized in Figures 3 and 4. The RMSE results represent the accuracy of the different metrics. For H.264/AVC, it can be observed that VQM is significantly more accurate than SSIM and MS-SSIM. This because their 95% confidence interval does not overlap. For this test set, it can also be observed that PSNR is significantly more accurate than the MS-SSIM metric. Now, the question we asked ourselves is whether with HEVC a different trend could be observed. In both the table and the figure with HEVC results, we observe that VQM is again significantly more accurate than MS-SSIM, but there is a small overlap with SSIM. For PSNR, the same observation as with H.264/AVC can be made.

Purely looking at the accuracy change between metrics when changing from H.264/AVC to HEVC, no definite changes can be noticed.

In order to investigate the linear correlation of the MOS scores with the metric results, pearson correlation results are given in Table V and Figures 5 and 6. Although a similar trend can be seen as in the RMSE results, there could not be found a significant difference in linear correlation between the different metrics.

Finally, graphs comparing H.264/AVC with HEVC are visualized in Figure 7. A trend that can be observed is that the RMSE value of the metrics consistently decreases when changing from H.264/AVC to HEVC. There is always

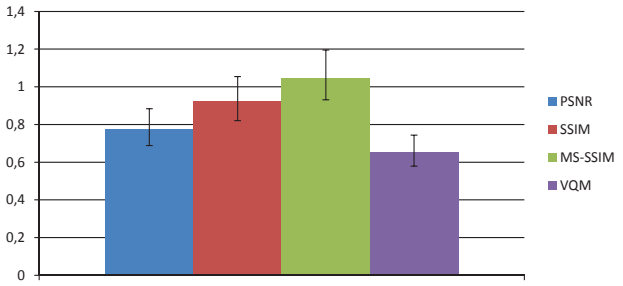

Fig. 3. RMSE of different objective models for the H.264/AVC testset.

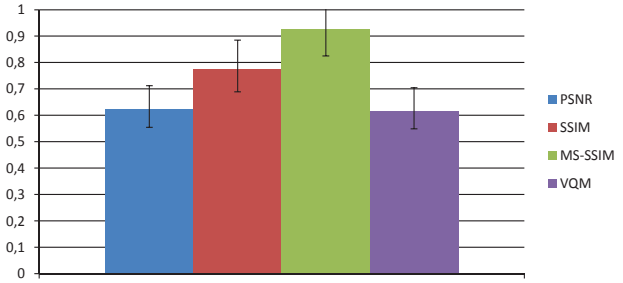

Fig. 4. RMSE of different objective models for the HEVC testset.

TABLE V. PEARSON CORRELATION AND 95% CONFIDENCE INTERVAL OF DIFFERENT OBJECTIVE MODELS FOR THE BOTH THE H.264/AVC AND THE HEVC TESTSET.

		PSNR	SSIM	MS-SSIM	VQM
H.264/AVC	upper	0.87	0.85	0.84	0.90
		0.82	0.80	0.78	0.86
	lower	0.75	0.72	0.70	0.81
HEVC	upper	0.91	0.89	0.86	0.91
		0.88	0.85	0.81	0.88
	lower	0.83	0.79	0.74	0.83

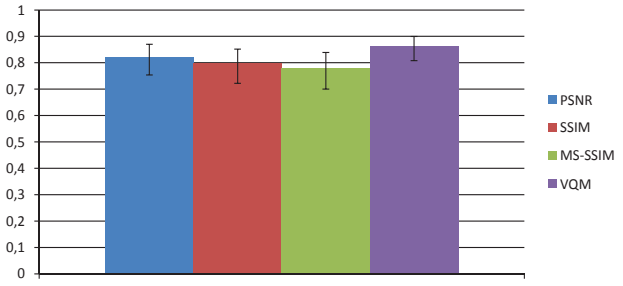

Fig. 5. Pearson correlation of different objective models for the H.264/AVC testset.

an overlap of the HEVC 95% interval with the H.264/AVC accuracy interval, so it can never be regarded as a significant improvement. The fact that VQM improves less in accuracy compared with the other metrics could indicate the block restrictions of this metric. From these results it seems plausible that further improvements to the VQM quality metric to increase performance on HEVC could be made.

From all the results, it can be observed that PSNR as a metric performs particularly well compared to the other

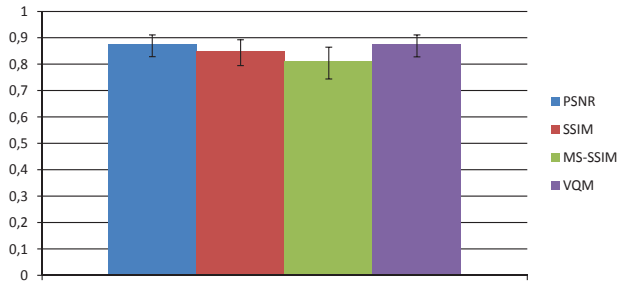

Fig. 6. Pearson correlation of different objective models for the HEVC testset.

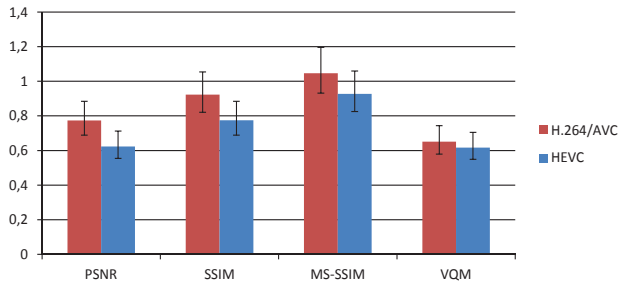

Fig. 7. RMSE of different objective models for the both the H.264/AVC and the HEVC testset.

techniques. Though, it must be noted that special precautions are made such that valid results would come out of the PSNR metric. To do this, pictures were aligned properly after introducing degradations, only progressive content was used, and no special degradations were introduced in the video stream. VQM, on the other hand, has more possibilities to cope with these irregularities because of its build-in functions handling interlace, frame temporal shift, spatial shift, valid region estimation or cropping, and contrast and brightness corrections. Therefore, the test can be considered as being set up more in favor of PSNR and less challenging for VQM.

Additionally, we would like to analyze sequence dependence of the different metrics. It could be observed that the sequence of Elephants Dream showed some strange MOS behavior. In particular, none of the models could predict these subjective scores very well. This may be caused by the combination of random movement and science fiction animation resulting in difficult to observe distortions. Looking closer into this subject, for HEVC the highest correlation between DMOS and metric result could be obtained with the cheetah, purple4e, foxbird3e, and basketball sequences. From SI and TI measurements (see Fig. 1), it appears that higher motion videos are easier to predict with a model.

VI. CONCLUSION

By evolving to a new video compression standard, different effects can be observed when degradations on the network take place. Then, the question arises if existing quality metrics still perform equivalently compared to the systems they were tested and designed on. In this paper, a direct comparison of different commonly used quality metrics (PSNR, SSIM, MS-SSIM, and VQM) is made when these are applied on

an error prone network environment carrying a new video compression scheme like HEVC. As a reference to compare with, packet loss effects on the well investigated H.264/AVC video compression standard is used. From the RMSE and pearson correlation measurements of the metrics with the MOS, it could be observed that the performance difference between metrics does not change significantly. On average, it is observed that all metrics except VQM perform slightly better with a change to HEVC, which could be caused by the fact that VQM is calculated on a 4x4 and 8x8 block basis, which does not scale with the increased block size of HEVC. In general, in both the H.264/AVC and the HEVC test, VQM was able to outperform or be slightly better than SSIM or MS-SSIM. Results from the PSNR metric are notably well compared to other experiments, but it should be noted that experiment conditions were set up more favorable for PSNR evaluation, which could explain this behavior.

ACKNOWLEDGMENT

The research activities described in this paper were funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research–Flanders (FWO–Flanders), and the European Union. Furthermore, this work was carried out using the STEVIN Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government - department EWI.

REFERENCES

[1] K. Yamori and Y. Tanaka, "Relation between willingness to pay and guaranteed minimum bandwidth in multiple-priority service," in *The 2004 Joint Conference of the 10th Asia-Pacific Conference on Communications and the 5th International Symposium on Multi-Dimensional Mobile Communications.*, August 2004, vol. 1, pp. 113 – 117.

[2] M. Ries, O. Nemethova, and M. Rupp, "On the willingness to pay in relation to delivered quality of mobile video streaming," in *International Conference on Consumer Electronics (ICCE 2008)*, January 2008, pp. 1–2.

[3] J. Asghar, F. Le Faucheur, and I. Hood, "Preserving video quality in IPTV networks," *IEEE Transactions on Broadcasting*, vol. 55, no. 2, pp. 386 –395, june 2009.

[4] S.R. Gulliver and G. Ghinea, "The perceptual and attentive impact of delay and jitter in multimedia delivery," *IEEE Transactions on Broadcasting*, vol. 53, no. 2, pp. 449–458, June 2007.

[5] F. Boulos, B. Parrein, P. Le Callet, and D. Hands, "Perceptual effects of packet loss on H.264/AVC encoded videos," *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-09)*, January 2009.

[6] M.H. Pinson, S. Wolf, and G. Cermak, "HDTV subjective quality of H.264 vs. MPEG-2, with and without packet loss," *IEEE Transactions on Broadcasting*, vol. 56, no. 1, pp. 86–91, March 2010.

[7] G.W. Cermak, "Consumer Opinions About Frequency of Artifacts in Digital Video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 336 –343, April 2009.

[8] A. Perkis, "Does quality impact the business model? case: Digital cinema," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2009, pp. 151–156.

[9] R. Serral-Gracià, E. Cerqueira, M. Curado, M. Yannuzzi, E. Monteiro, and X. Masip-Bruin, "An overview of quality of experience measurement challenges for video applications in IP networks," in *Wired/Wireless Internet Communications*, vol. 6074 of *Lecture Notes in Computer Science*, pp. 252–263. Springer Berlin / Heidelberg, 2010.

[10] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649 –1668, dec. 2012.

[11] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," vol. 13, no. 7, pp. 560–576, Jul. 2003.

[12] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards including high efficiency video coding (hevc)," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1669 – 1684, dec. 2012.

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

[14] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," *in IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 13981402, Nov. 2003.

[15] "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.," *ITU-T Rec. J.144*, Mar. 2003.

[16] P. Corriveau, *Video Quality Testing*, chapter 4, pp. 125–153, Digital Video Image Quality and Perceptual Coding. CRC Press, 2006.

[17] Gordon E. Legge and John M. Foley, "Contrast masking in human vision," *Journal of the Optical Society of America*, vol. 70, no. 12, pp. 1458–1471, Dec 1980.

[18] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union (ITU), 1999.

[19] A. Ostaszewska and R. Kloda, "Quantifying the amount of spatial and temporal information in video test sequences," in *Recent Advances in Mechatronics*, pp. 11–15. Springer Berlin Heidelberg, 2007.

[20] M.H. Pinson, S. Wolf, N. Tripathi, and C. Koh, "The consumer digital video library," *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-10)*, January 2010.

[21] VQEG Tools and Subjective Labs Setup, "Modified JM H.264/AVC Codec," Tech. Rep., http://vqegstl.ugent.be/.

[22] K. McCann, B. Bross, S.-i. Sekiguchi, W.-J. Han, "HM4: High Efficiency Video Coding (HEVC) Test Model 4 Encoder Description," *JCTVC-F802*, Torino, Italy, Jul. 2011.

[23] Gerard O'Driscoll, *Next Generation IPTV Services and Technologies*, Wiley-Interscience, New York, NY, USA, 2008.

[24] A. Rombaut, N. Staelens, N. Vercammen, B. Vermeulen, and P. Demeester, "xStreamer: Modular Multimedia Streaming," in *Proceedings of the seventeenth ACM international conference on Multimedia*, 2009, pp. 929–930.

[25] S. Wenger, M.M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer, "RTP Payload Format for H.264 Video," February 2005.

[26] J. Korhonen, N. Burini, Junyong You, and E. Nadernejad, "How to evaluate objective video quality metrics reliably," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2012, pp. 57–62.