ON ACCURACY OF OBJECTIVE METRICS FOR ASSESSMENT OF PERCEPTUAL PRE-PROCESSING FOR VIDEO CODING

*† *Madhukar Bhat,* †*Jean-Marc Thiesse,* **Patrick Le Callet*

*LUNAM University, University of Nantes, Group IPI Lab LS2N, Rue Christian Pauc , Nantes 44306
†VITEC, 99 rue Pierre Semard, 92320 CHATILLON, France.

ABSTRACT

The objective assessment of encoding performance is a key aspect of video delivery optimization. Objective metrics typically do not address fully the different viewing distances and behavior of compression artifacts being subjected to perceptual changes in video. This poses a daunting task of optimizing compression for video delivery systems for specific viewing conditions and perceptual optimization. This paper introduces a perceptual pre-processing and then discusses accuracy of typically used objective metrics for judging performance of pre-processing for observers at different viewing distances. To this end, this paper reports an in-depth analysis to check if objective metrics can successfully match subjective results at critical pairs i.e. pre-processed and original video at same QP.

Index Terms— Perceptual pre-processing, Objective quality metrics, Paired comparison, critical pairs, perceptual performance

1. INTRODUCTION

Video coding in recent years have invested a considerable research in perceptual optimization of encoded video. Human perception of video consumptions are considered in video coding [1]. Application of human perception in video compression can be approached both within and outside encoding loop. In particular, video pre-processing is a conventional way to apply perceptual consideration outside video encoding loop.

Pre-processing models intends to remove as much information as possible within noticeable threshold from source video before encoding. Some of the earliest applications of pre-filtering were denoising filter [2] [3] to remove noise with edge preservation, which could reduce encoding performance. In [4] advanced analysis lead to application of perceptual anisotropic filters based on contrast sensitivity map to consider Human Visual System (HVS) while removing information from video. Another perceptual modeling parameter considered was Just Noticeable Difference (JND) to control pre-filter proposed in [5].

When designing such perceptual pre-processing, having a reliable measure of performance of the algorithm is critical for calculating bit-rate savings at similar quality. Usually subjective tests are more suitable for quantifying improvements in perceptual optimization [6]. However, they are time consuming and costly. Objective quality metrics are therefore a relevant substitute for subjective test to reduce time, cost and impracticality of test design. But, quality metrics intended for performance evaluation of perceptual improvement should be verified by subjective data to reliably depend on them.

Evaluation of objective metrics for standard video compression quality measurement has been carried out in several studies [7] [8]. In these methods, classical performance evaluation frameworks like Pearson Linear Correlation Coefficient (PLCC) and Root-Mean-Square Error (RMSE) are used. In these frameworks, metrics are fitted by regression to map them to Mean Opinion Scores (MOS) which closely matches ITU-T recommendations [9] [10]. These methods use direct scaling methods for subjective test. But in case of pre-processing it is very important to judge stimuli whose visual difference are small. Indirect scaling methods like paired comparison (PC) which has higher discriminatory power is more suitable [6].

Reibman et al. [11] proposed a method to jointly test image quality estimators using paired comparison and many metrics. Hanhart et al. [6] proposed Receiver Operating Characteristic (ROC) based analysis to use indirect scaling data and Krasula et al. [12] used direct scaling data for objective quality benchmarking. These frameworks test for finding accuracy of quality metrics in detecting significant difference between pairs of stimulus. They also allow finding accuracy for detection of better or worse stimulus in a specific pair.

This paper details the accuracy of commonly used objective metrics in assessment of perceptual pre-processing for video compression. The considered metrics include *PSNR*, *MS-SSIM*, *VIF*, *VMAF* [13], *PQR* [14] and *HDR-VDP* [15]. First, a perceptual pre-processing optimized for multiple viewing distances below JND threshold is used for treatment of video before encoding. Then both original and pre-processed videos are encoded using a professional HEVC encoder and are subjected to paired comparison test using square design [16]. Finally accuracy of objective metrics for

judging critical pairs of pre-processed and original videos encoded at same QP are tested for two viewing distances using state of the art tools introduced in [6] and [12]. This analysis provides details regarding accuracy of metrics for assessing RD-performance of pre-processing optimized for multiple distances. In addition an analysis is conducted for all possible pairs tested in subjective test to assess performance of metrics to evaluate RD-performance of general video compression.

This paper is organized as follows. Section 2 details the perceptual pre-processing algorithm used for the study. Section 3 explains the tools used for measuring accuracy of objective metrics. Experimental setup for paired comparison subjective test is given in section 4. Performance of objective metrics for pre-processing is discussed in section 5. Conclusions are drawn in section 6.

2. PERCEPTUAL PRE-PROCESSING ALGORITHM

Perceptual pre-processing used in this paper is built on accurate modeling of HVS for filtering within JND threshold. Relevant visual tools of HVS are involved, such as optical transfer function which is based on [17] and Contrast Sensitivity Function (CSF) which is based on [18]. JND threshold is considered in applying masking effect on source video by controlling their shape. Both signal dependent (N_{CSF}) [19] and independent noises from CSF (N_{Nmask}) are removed from video. Thanks to HVS base, pre-processing is naturally optimized for multiple viewing distances from the intended screening. Optimization for multiple viewing distance are conducted to study weather lack of modeling of distances in objective metrics affect their judgment at different intended viewing distances.

Two distances are used for analysis in this paper at 3H (height of the display) which is recommended minimum distance of the observer from full HD display and 4.5H which is the reference viewing distance for consumer. The optimization of masking effect for two distance changes because visual resolution of human eye can catch less errors from video moving away from the screen. Normalized CSF is applied to source video in multi-scale complex steerable pyramid [20] domain.

Pre-processed band $B_{pre}(f,o)$ in f^{th} scale and o^{th} orientation band of steerable decomposition is obtained by applying N_{Nmask} and N_{CSF} with gain controlling factor p on source band $B_{src}(f,o)$. It is expressed as,

$$B_{pre}(f,o) = B_{src}(f,o)\left(1 - \frac{1}{\sqrt{N_{CSF}^{2p} + N_{Nmask}^{2}}}\right) \quad (1)$$

Visual display considered for optimizing pre-processing is standard LCD monitor at maximum brightness of 250cd/m2.

3. ANALYSIS OF ACCURACY OF OBJECTIVE QUALITY METRICS

Raw preference data from paired comparison test has to be processed as a first step before any analysis can be conducted. Paired comparison test used in this paper uses square design for choosing pairs within one video sequence. First raw subjective test data of each video sequence is processed to get relative score for each stimuli using Bradley-Terry method. This is followed by analyzing if the two pairs are significantly different or similar subjectively. If they are significantly different then from relative subjective score, better stimuli from the pair needs to be selected.

Then, respective objective quality metrics data are gone through ROC analysis and classification errors. Performance difference between quality metrics can be found using accuracy measure. Analysis performed here are inspired by [6] and [12].

3.1. Producing significantly different subjective pairs

First step in producing significantly different stimulus pairs in a video sequence is to process raw comparison data from paired comparison test. Raw preference scores are processed using Bradley-Terry model to obtain relative subjective scores (RS) of each stimuli with confidence index (CI) for a particular video sequence. It should be noted that one needs to be careful using these scores as relative scores between two different video sequences are meaning less. However, this remains relevant for this study since Bradley-Terry scores are only used to tell if the pairs are subjectively significantly different. From relative scores, two pairs S1 and S2 are significantly different if,

$$||RS(S1) - RS(S2)|| > CI \quad (2)$$

If they are significantly different then the better stimuli is noted by comparing their scores.

3.2. Quality metric data processing

Objective quality for stimuli used for paired comparison can be obtained. Then for pairs used in subjective test difference in objective quality metric is taken as prediction of that metric for the pair. For each pair S1 and S2,

$$\Delta_{OM}(S1,S2) = score_{OM}(S1) - score_{OM}(S2) \quad (3)$$

where, $score_{OM}(Si)$ are the quality prediction of a specific metric for stimuli pair. Difference of metrics obtained for each pair is compared with data from section 3.1 and then analyzed by methods described in the following sections.

3.3. Different v/s similar performance

In this analysis how well objective quality metrics detect significantly different and similar pairs compared to subjective

model from section 3.1 is assessed. Usually ROC analysis is carried out to determine abilities of binary classifier [16]. Performance indicator in ROC curve can be obtained by Area Under Curve (AUC) [6]. Comparison between metrics can be obtained using their AUC values.

3.4. Better v/s worse performance

Another analysis useful in measuring accuracy of metrics is when pairs are different, whether a metric can also detect better/worse compared to subjective model from section 3.1. For this, AUC from ROC is calculated for better/worse analysis.

3.5. Classification Errors

Classification error for a stimulus pair(S1,S2) happens when a particular objective quality metric leads to different conclusion compared to subjective evaluation. There can be three types of errors. First one is *False Tie* error which occurs when objective metrics says S1 and S2 are identical but subjective models say that they are different. *False Differentiation* occurs when objective metrics classifies S1 and S2 as different when subjectively they are the same. *False ranking* which is most offensive error, which happens when objective metrics says S1(S2) is better than S2(S1) when it is the opposite in subjective test.

As explained ITU-T recommendation ITU-T J.149 [9], the percentage of *Correct Decision (CD), False Tie (FT), False Differentiation (FD) and False Ranking (FR)* are recorded from all possible distinct pairs as a function of metric value difference. In this paper classification errors are used to determine these four values for mean objective difference (ΔOM_{mean}) for critical pairs.

4. EXPERIMENTAL SETUP

To measure the accuracy of objective quality metrics five 1080p video sequences are selected. Extensive paired comparison subjective tests and objective tests have been conducted to collect valid data for analysis of tools discussed in the last section. Performances of objective metrics is then calculated in the context of perceptual pre-processing for video compression using professional HEVC encoder. Subjective tests incorporated to evaluate the proposed framework follows ITU-T P.910 [10]. The subjective test protocols are explained in following paragraphs.

4.1. Selection of content and Stimuli

Video sequences used for calculating accuracy of perceptual pre-processing algorithm are five 1080p sequences at 50fps. Two videos, Basketballdrive and BQTerrace are selected from JCT-VC common test conditions [21]. Remaining three videos Crowdrun, Duckstakeoff and Oldtowncross are selected from SVT-HD test set [22] for diversity in spatial and temporal complexity and their wide range of use in video codec performance analysis.

Selection of stimuli for analysis is based on pre-processed and original video encoded at 4 different QP's. Similar quality level for comparison is selected based on expectation that at same QP both pre-processed and original videos behave in a similar way subjectively. It is verified through pre-analysis of All-Intra frames. Eventually, 4 QPs ranging from 20-42 for each sequence have been selected based on their RD-Curve.

4.2. Test environment

The selection of testing environment here is based on Recommendation ITU-T P.910 [10]. Two viewing distance corresponding to height of the display (H) at 3H and 4.5H are chosen. The luminance of the screen used for the test is 250cd/m2.

4.3. Paired comparison design

Paired comparison test conducted here uses squared design [23] with 9 Hypothetical Reference Circuits (HRC's) and 18 pairs to compare for each sequence. These come from one source and eight processed video encoded at different QP's from pre-processed and original videos. 30 naive observers participated in the test for each viewing distance.

5. RESULTS AND DISCUSSION

In this section, the application of analysis tools on perceptual pre-processing introduced in section 3 is reported for experimental setup explained in section 4. Subjective data from section 3.1 is compared with objective metrics data from section 3.2. First analysis of metrics for all possible pairs in video sequences to measure their accuracy is conducted. Then analysis of critical pairs for measuring accuracy of objective metrics for benchmarking pre-processing is carried out. Two viewing distances are separately analyzed to see if lack of modeling distances in objective metrics affect their judgment.

5.1. Overall pair Analysis

Table 1 depicts different/similar AUC and better/worse AUC for all the metrics used for analysis at both distances. Here performance of metrics are expectedly higher as most of the test pairs are with stimuli at different QP. Objective metrics often succeed to rank stimuli with large difference in subjective quality and algorithms without lot of perceptual considerations. Hence, they are popular in measuring performance of general video compression. It is also worth noting that at 4.5H AUC for different/similar analysis is a bit lower than that of 3H. It is because as one move away from screen it is difficult to differentiate pairs which are closer in quality. On the other hand with increasing distance (at 4.5H) if pairs are significantly different one can select better or worse more than

at 3H. Since objective metrics do not model viewing distance, their performance decrease for different/similar analysis for other than optimized viewing distance.

Metrics	Different/Similar		Better/Worse	
	3H	4.5H	3H	4.5H
PSNR	0.9333	0.8808	0.9709	0.9905
MS-SSIM	0.9491	0.8900	**0.9772**	0.991
VIF	0.9539	0.8947	0.9722	**0.995**
VMAF	0.9418	0.8958	0.9752	**0.995**
PQR	0.9515	**0.9074**	0.9659	0.991
HDR-VDP	**0.9673**	0.8938	0.9749	0.9904

Table 1. Different/Similar and Better/Worse AUC for metrics at both distances

5.2. Critical pair analysis

Critical pairs are the pairs of pre-processed and original video encoded at same QP. Correct decision for these pairs from objective metrics is critical in assessment of pre-processing. Table 2 shows different/similar AUC and better/worse AUC for each metrics for both distances. Overall, the metrics have quite small AUC, except HDR-VDP which has substantially better scores compared to other metrics. This means that objective metrics are inefficient to judge quality of perceptual pre-processing. Since subjectively critical pairs are closer, objetive metrics often fail in judging them correctly. On the other hand HDR-VDP performs better than any metric at both distance with respectable AUC for both different/similar and better/worse analysis.

It can be seen from Table 2 that for different/similar judgment of stimuli viewing distance 4.5H is better than 3H for all metrics, which is reverse of what is found for overall pair analysis. This is due to the fact that pre-processing is optimized for two distances and observers found almost same number of pairs to be subjectively different/similar for both distance. But at 4.5H in better/worse analysis except for VMAF and VIF, other metrics perform better than at 3H. This is similar to the overall pair analysis.

Metrics	Different/Similar		Better/Worse	
	3H	4.5H	3H	4.5H
PSNR	0.4633	0.5156	0.4171	0.5
MS-SSIM	0.44	0.475	0.5257	0.625
VIF	0.2667	0.2813	0.3829	0.375
VMAF	0.4783	0.6031	0.4743	0.4681
PQR	0.5133	0.6406	0.4629	0.5
HDR-VDP	**0.5967**	**0.7375**	**0.7489**	**0.8375**

Table 2. Different/Similar and Better/Worse AUC for critical pairs

As explained in section 3.5 classification errors for the ΔOM_{mean} of each metrics at critical pairs for both distance is given in table 3. HDR-VDP is significantly better in giving *Correct Decision (CD)* than other metrics at both distances. This is due to ΔOM_{mean} of HDR-VDP, which is closest to the ΔOM value that maximizes correct decision compared to other metrics. At 3H except PSNR, all metrics can achieve very low *False Ranking (FR)*. At 4.5H VMAF and PQR are prone to FR error. For VMAF it might be due to optimization done for decisions at 3H [13]. Overall, *False Differentiation (FD)* score increases with distance for all metrics except for HDR-VDP. This is because HDR-VDP considers viewing distance in its calculation.

Lack of modeling of perceptual factors and viewing distance is apparent in most metrics except HDR-VDP as they perform differently with viewing distance and judge poorly the perceptual pre-processing for video coding. This poses a problem to reliably depend on objective metrics for judging RD-performance of perceptual pre-processing and its optimization for multiple viewing distances. Overall HDR-VDP is more accurate in judging performance of perceptual pre-processing.

Metrics	CD (%)		FT (%)		FD (%)		FR (%)	
	3H	4.5H	3H	4.5H	3H	4.5H	3H	4.5H
PSNR	55	60	15	**10**	25	30	5	0
MS-SSIM	65	60	20	15	15	25	0	0
VIF	57.5	60	20	15	22.5	25	0	0
VMAF	62.5	65	20	**10**	17.5	25	0	5
PQR	55	65	35	**10**	**10**	20	0	5
HDR-VDP	**75**	**75**	**10**	15	15	**10**	0	0

Table 3. Classification errors for metrics

6. CONCLUSION

In this paper accuracy of objective quality metrics for measuring performance of perceptual pre-processing for video compression is tested. Even though objective metrics are fairly accurate for evaluating video compression algorithms in general, they perform very poorly for critical pair judgments. This study tends to suggest that objective metrics can be improved for accurately measuring RD-performance of perceptual optimizations such as pre-processing which involves tiny subjective differences. Further considerations in viewing distance and perceptual aspects are necessary for their reliability in judging perceptual algorithms. Even though computationally slow, HDR-VDP which takes into account viewing distance and perceptual modeling in judging two stimuli is the most suitable metric for assessing performance of perceptual pre-processing for video compression.

References

[1] Z. Chen, W. Lin, and K. N. Ngan, "Perceptual video coding: Challenges and approaches," in *Multimedia and expo (ICME), 2010 IEEE international conference on*, IEEE, 2010, pp. 784–789.

[2] H.-Y. Cheong, A. M. Tourapis, J. Llach, and J. Boyce, "Adaptive spatio-temporal filtering for video denoising," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, IEEE, vol. 2, 2004, pp. 965–968.

[3] C. Q. Zhan and L. J. Karam, "Wavelet-based adaptive image denoising with edge preservation," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, IEEE, vol. 1, 2003, pp. I–97.

[4] R. Vanam and Y. A. Reznik, "Perceptual pre-processing filter for user-adaptive coding and delivery of visual information," in *Picture Coding Symposium (PCS), 2013*, IEEE, 2013, pp. 426–429.

[5] L. Ding, G. Li, R. Wang, and W. Wang, "Video preprocessing with jnd-based gaussian filtering of superpixels," in *Visual Information Processing and Communication VI*, International Society for Optics and Photonics, vol. 9410, 2015, p. 941 004.

[6] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data?" In *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, IEEE, 2016, pp. 1–6.

[7] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE transactions on broadcasting*, vol. 57, no. 2, p. 165, 2011.

[8] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[9] ITU-T-J.149, "Method for specifying accuracy and cross-calibration of video quality metrics (vqm)," 2004.

[10] ITU-T-P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[11] A. R. Reibman, K. Shirley, and C. Tian, "A probabilistic pairwise-preference predictor for image quality," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, IEEE, 2013, pp. 413–417.

[12] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, IEEE, 2016, pp. 1–6.

[13] Netflix, *Toward a practical perceptual video quality metric*, 2016. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652 [Accessed: 5-Feb.- 2018].

[14] T. AN, *Understanding PQR, DMOS and PSNR measurements*, 2014. [Online]. Available: https://www.tek.com/document/fact-sheet/understanding-pqr-dmos-and-psnr-measurements [Accessed: 5- Feb.- 2018].

[15] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," vol. 30, no. 4, p. 40, 2011.

[16] J. A. Swets, *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press, 2014.

[17] R. J. Deeley, N. Drasdo, and W. N. Charman, "A simple parametric model of the human ocular modulation transfer function," *Ophthalmic and Physiological Optics*, vol. 11, no. 1, pp. 91–93, 1991.

[18] P. G. Barten, *Contrast sensitivity of the human eye and its effects on image quality*. Spie optical engineering press Bellingham, WA, 1999, vol. 19.

[19] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *JOSA A*, vol. 14, no. 9, pp. 2379–2391, 1997.

[20] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International journal of computer vision*, vol. 40, no. 1, pp. 49–70, 2000.

[21] JCTVC-L1100, "Common test conditions and software reference configurations," 2013.

[22] L. Haglund, "The SVT high definition multi format test set," 2006.

[23] O. Dykstra, "Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs," *Biometrics*, vol. 16, no. 2, pp. 176–188, 1960.