Perceptual Video Quality Assessment for Adaptive Streaming Encoding

Estêvão C. Monteiro, Ricardo E. P. Scholz, Carlos A. G. Ferraz, Tsang I. Ren, Roberto S. M. Barros

Centro de Informática, Universidade Federal de Pernambuco

Cidade Universitária, Recife, 50740-560, Brazil

{ecm3, reps, cagf, tir, roberto}@cin.ufpe.br

Abstract— **Adaptive video streaming has become prominent due to the rising diversity of Web-enabled personal devices. Common limitations in bandwidth and decoding power challenge the efficiency of content encoders to preserve visual quality at reduced data rates over a wide range of display resolutions. Objective assessment of perceptual video quality has greatly improved in the past decade but remains an open problem. Among the most relevant metrics are the many variations of the Structural Similarity (SSIM) index. In this work, several SSIM-based metrics are compared, optimized and improved towards better correlation with human perception by testing the HD content of the LIVE Mobile Video Quality Database. A shifted gradient is proposed to preserve more image feature information for similarity comparison, thus increasing accuracy, along with a down-sampling box pooling filter that coherently emulates Gaussian pooling while reducing computation complexity by a factor of four and providing broader scalability.**

Index Terms— **image quality, rate-distortion optimization, video perceptual fidelity, adaptive video streaming, structural similarity**

I. Introduction

The increasing diversity of personal devices with varying bandwidth and video decoding capabilities, along with the billionaire entertainment industry and the recent cloud services paradigm, have led to a prevalence of adaptive video streaming services in the Internet. In such systems, several low-data-rate versions of each video content are produced according to targeted devices and bandwidth, while attempting to maintain a threshold of visual quality against human perception [1]. However, accurately modelling the human perception of image quality remains an open research problem [2].

The Structural Similarity (SSIM) index for image quality assessment (IQA) has gained wide adoption and has been improved by various techniques [2]-[12]. However, such metrics are often developed for general-purpose IQA, whereas video versioning for adaptive streaming poses a specific, reduced set of problems for which the effectiveness of general-purpose metrics may substantially differ. Typical problems are loss of visual information by blurring, and addition of artifacts, such as banding, blocking and ringing. An effective encoding implementation minimizes blur and artifacts by means of rate-distortion optimization (RDO) and adaptive quantization (AQ) methods [13], [14]. Also, encoding and streaming require low latency, so restricting computational complexity for higher encoding performance is a primary concern as well.

Particularly effective implementations of RDO and AQ are found in the X264 encoder for H.264 video, one of the most efficient publicly available video encoders [15]-[17]. Dubbed psychovisual RDO or Psy-RDO, they have been calibrated to subjective quality by prioritizing visual detail retention over traditional metrics that favor blurring such as mean squared error (MSE) and even the original SSIM. These features, coupled with the fact that H.264 is the most ubiquitous video format in connected devices, make X264 encoding a relevant case study.

Most modern video encoders employ basic forms of the original SSIM for RDO, but Psy-RDO has been shown to give better subjective quality [18]. By studying optimization decisions in X264 which improve upon simple SSIM-based decisions, we propose a new SSIM-based index that better correlates with subjective video quality assessment and that is more computationally efficient.

The rest of the work is organized as follows. Section II discusses the psychovisual RDO that motivates the Shifted Gradient Similarity (SG-Sim) proposal. Section III reviews the SSIM metric and many of its derivations. Section IV explains and defines SG-Sim. Section V presents pooling filter optimizations. Section VI provides the experimental results. Finally, Section VII gives conclusions and proposes further investigation.

II. Psychovisual Rate-Distortion Optimization

Every lossy video encoder faces a critical decision: which information to discard with the least impact on visual quality. Mean squared error (MSE) is the most widely and longest adopted metric for quantization decisions during encoding, as well as for video quality assessment (VQA), such as in the form of peak signal-to-noise ratio (PSNR). However, it is long proven a poor perceptual metric, favoring blur over fine detail and completely disregarding image structure and spatial coherence [19], a deficiency upon which SSIM-based metrics improve. In general, poor encoder implementations maximize PSNR whereas good implementations maximize SSIM.

Psy-RDO is an encoding mode of X264 which employs fine-tuned techniques such as adaptive quantization, visual energy retention, Trellis quantization and in-loop de-blocking filter [18] to achieve better subjective image quality than RDO by SSIM. Particularly, Psy-RDO produces less blur than SSIM, which the developers argue improves subjective assessment due to preservation of complexity. Further, AQ defines regions of homogeneous complexities within the spatial transform blocks

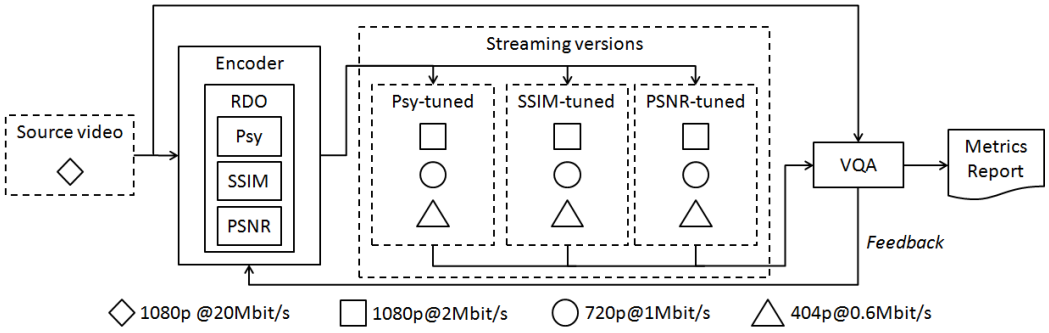

Fig. 1. The low-data-rate streaming versions of different RDO modes may be compared to the source by VQA metrics in order to identify correlations between RDO modes and VQA metrics. Results allow improvements to both RDO implementations and VQA metrics.

and pays particular attention to smooth gradients, such as the sky, which may easily suffer blocking and degrade subjective quality disproportionally to actual information loss.

Psy-RDO is a coding implementation, not actually a VQA metric such as SSIM. A metric that is analogous with the contributions of Psy-RDO would be useful to improve RDO implementations as well as VQA in general. SSIM has been modified towards a wide range of applications, and can also be modified to this task. Such application is represented in Fig. 1.

III. SSIM and Related Metrics

Wang and Bovik's original universal image quality index [20] analyses the luminance channels of two images x and y in terms of contrast (variance of intensities, actually the standard deviation σ_x), structure (correlation of variances, σ_{xy}) and luminance (mean intensities, μ_x). Pixel results are spatially pooled by an 8x8 box filter, which promotes index coherence and stabilization, consistently with the human visual system (HVS). Later improvement of this metric by Wang et al. [3] resulted in the SSIM index, defined in (1), with division stabilization through addition of constants proportional to the image's dynamic range, and the contrast and structure terms combined and simplified into a single structure term that defines the index. Further, the box filter is replaced by an 11x11 Gaussian filter with $\sigma = 1.5$. Thus, SSIM may be understood as a combination of a structure statistic, a division stabilization strategy and a pooling filter. In video, the overall SSIM result is usually the simple mean of the indexes for all individual frames in the stream.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

Rouse and Hemami [4] argue, however, that the luminance term offers insignificant contribution; this is corroborated by the Multi-scaled SSIM (MS-SSIM) [5] and Gradient Magnitude Similarity Deviation (GMSD) [2] indexes.

MS-SSIM achieves better subjective correlation because HVS itself is multi-scalar. This index pools the responses of five scales of the image structure, through dyadic down-sampling, and computes luminance only for the smallest scale.

Chen, Yang and Xie [6] applied SSIM to the gradients of the evaluated images, and this was found more effective. Inversely, 3-SSIM [7] and 4-SSIM [8] propose to segment and weight SSIM by the gradients. Chen and Bovik [9], however, discard variance altogether in their optimized Fast SSIM, comparing only the gradients, also optimizing computation speed by an integer approximation of the quadratic magnitude. Gradients are particularly effective in encoding VQA because they are sensitive to both blurring and compression artifacts. Fast SSIM also proposes to perform multi-scaling with only the four reduced scales, ignoring the original, which contributes the least to the final index.

GMSD [2] is another metric that substitutes variance by gradient, and also discards the luminance and structure terms, comparing only the simple contrasts (gradients). Spatial pooling in GMSD is global instead of local, and by standard deviation instead of mean. It is one of the fastest SSIM-based metrics and achieves generally higher correlation with subjective scores for visual quality than MS-SSIM.

Temporal-based SSIM variations such as Motion-based Video Integrity Evaluation (MOVIE) [10] and Spatio-Temporal Video SSIM [11] were not tested in this work due to their higher complexity which is detrimental to performance in both RDO and live streaming.

IV. The Shifted Gradient Similarity

Preliminary experiments showed that comparisons between only the gradients produce more zero dividends in the SSIM equation than variance does, which results in loss of useful information and an overall decrease in the index and its effectiveness. We propose to mitigate this pitfall by shifting the gradient's magnitudes by +1. We find this improves the index's responses compared to stabilizing with a distorting constant, because numeric proportions between the compared pixel intensities are preserved. Also, amplitudes between similar indexes are generally widened, facilitating comparison. This corroborates with criticism by Rouse and Hemami [12] to the distortions arising from stabilization by constants and also improves adherence to Weber's law of light adaptation [3]. Indeed, Wang et al [3] admit the constant stabilizer's value to be "somewhat arbitrary".

To illustrate SSIM behavior, suppose an original gradient of 1 that is distorted to 2: without stabilization, (1) produces an index of 0.8; whereas the usual stabilization constant of 58.5 produces an index of 0.9843. When 1 is distorted to 8, the index is 0.6032 with constants and 0.2462 without. At higher intensities, any distortion affects less the index: 200 distorted

to 230 produces 0.9903, both with and without a stabilizer constant. As the gradient may be considered analogous to contrast for MPEG-based VQA purposes, increasing the response to differences in lower magnitudes better correlates to light adaptation and may improve the index's effectiveness.

The approximation of the shifted gradient of the source video, ∇S, is defined below in (2), where ∇x and ∇y are the vertical and horizontal responses to the Roberts cross operators [9]. ∇V is its counterpart for the low-data-rate version. Spatial pooling, represented by μ for N pixels, is defined below in (3) and (4), and may represent the simple mean of a Box filter, the normal distribution of a Gaussian filter, or the entire frame, unfiltered. The Shifted Gradient Similarity index (SG-Sim) is defined below in (5).

$$\nabla S_i = \max(|\nabla x_i|, |\nabla y_i|) + {}^1\!/_4 \min(|\nabla x_i|, |\nabla y_i|) + 1 \quad (2)$$

$$\mu_{\nabla S} = \frac{1}{N}\sum_{i=1}^{N} \nabla S_i \quad (3)$$

$$\mu_{\nabla SV} = \frac{1}{N}\sum_{i=1}^{N} (\nabla S_i \, \nabla V_i) \quad (4)$$

$$SG\text{-}Sim(S,V) = \frac{2\mu_{\nabla SV}}{\mu_{\nabla S}^2 + \mu_{\nabla V}^2} \quad (5)$$

V. Optimizing the Pooling Filter

Fast SSIM identifies the greatest computational cost in SSIM-based indexes as the spatial convolutions used for pooling local intensities [9]. Digital video typically streams at least 24 frames per second, so efficiency in computations is a primary concern. To minimize the impact of convolutions, that metric avoids all floating point operations, typical of image filtering, by employing a rough integer approximation of the Gaussian filter and normalizing the response by dividing by the sum of the coefficients for weights. Further, the window size is reduced from 11x11 to 8x8, for 48% less operations.

The coefficients chosen for Fast SSIM, however, are significantly imprecise and 37.5% null. We propose, instead, to trim the 11x11 Gaussian filter to its core 7x7 coefficients, normalized so that the first equals 1. With σ =1.5, an 11x11 filter includes 3.3 standard deviations of the normal distribution at the horizontal and vertical rows, whereas a 7x7 filter covers exactly 2 deviations, which give 96.6% of the weights of the former using merely 40.5% of the coefficients. Further, a 5x5 filter would cover 1.3 deviations and 86% of the weights using merely 20.7% of the coefficients. Our experiments show that, due to spatial coherency of natural images, the index responses to those approximations are over 99% similar to the response from the full 11x11 filter, while computing respectively 23% and 30% faster in an well-known optimized implementation by two perpendicular separated 1-dimensional filters instead of the single 2-dimensional filter.

Spatial coherency also allows replacing the 5x5 Gaussian filter by a 5x5 box filter, then down-sampling by such filter instead of sliding the window, while retaining 98% similarity and computing 250% faster than the 11x11 Gaussian filter. The down-sampling box window strategy also enables scalability: for an arbitrarily larger video frame, a proportionally larger window down-samples to the same size for assessment so that the only increases in computation cost are during the gradient filtering and box down-sampling. This strategy may be particularly effective for Full HD and Ultra HD content.

VI. Perceptual Correlation and Complexity Tests

The video samples of compression distortion in the LIVE Mobile Video Quality Database [21], [22] were used for evaluating the perceptual performances of the reference and proposed metrics. These consist of four versions of increasing distortion for each of ten 15-second HD video sequences. Degradation mean opinion scores (DMOS) are given for each of the forty distorted versions in order to test metric correlation. The methodology of the ITU-T Video Quality Experts Group [22] was applied, which consists of evaluations of accuracy, monotonicity and consistency of the metrics predictions of DMOS by respectively the Pearson linear correlation coefficient (LCC), the Spearman rank order correlation coefficient (ROCC) and the root mean squared error (RMSE), all over the non-linear regression of the data by a logistic fitting function. Finally, the metrics' computation time in seconds was also measured for relative performance comparison.

The experiments were implemented as an extensible framework of numerous SSIM-based tools designed for VQA experimentation with 1080p support called Video Quality Assessment in Java – JVQA, available online at http://sourceforge.net/p/jvqa. All tests were conducted in a single processing thread on a 64-bit Windows 7 system on an Intel Core i5-4690 CPU.

Table 1 presents the most representative results. The shifted gradient metrics are always without stabilizing constants, whereas all others always use the constants, except where conditional is specified. The best three results for each column are shown in bold; our propositions are in italic. Fig. 2 gives the scatter plots for the logistic-fitted regression data for SG-Sim.

Table 1. Performance comparison of VQA metrics over the LIVE database.

Metric	LCC	ROCC	RMSE	Time
Four-scaled SG-Sim, Gauss	**0.907**	**0.917**	**0.48**	0.37
Five-scaled SG-Sim, Gauss	**0.904**	**0.913**	**0.49**	0.99
Four-scaled SG-Sim, Down-sampling Box	**0.901**	**0.908**	**0.49**	**0.19**
Five-scaled SSIM, Gauss [5]	0.839	0.840	0.62	1.40
SG-Sim, Gauss	0.812	0.823	0.67	0.69
Fast SSIM, Gauss [9]	0.803	0.807	0.68	0.68
GMSD [2]	0.804	0.782	0.68	0.37
SG-Sim, Box	0.797	0.781	0.69	0.43
SG-Sim, Down-sampling Box	0.797	0.781	0.69	**0.24**
Fast SSIM, conditional [9], [12]	0.773	0.747	0.72	0.68
3-SSIM, Gauss [7]	0.761	0.731	0.74	1.48
SSIM, Gauss [3]	0.743	0.708	0.76	1.00
SG-Sim, unfiltered	0.690	0.675	0.83	**0.14**

Our proposed shifted gradient image enhancement statistic stands out as a perceptual improvement over the SSIM index as well as the recent GMSD, while 45% faster than the first and 86% slower than the latter. Although GMSD retains a

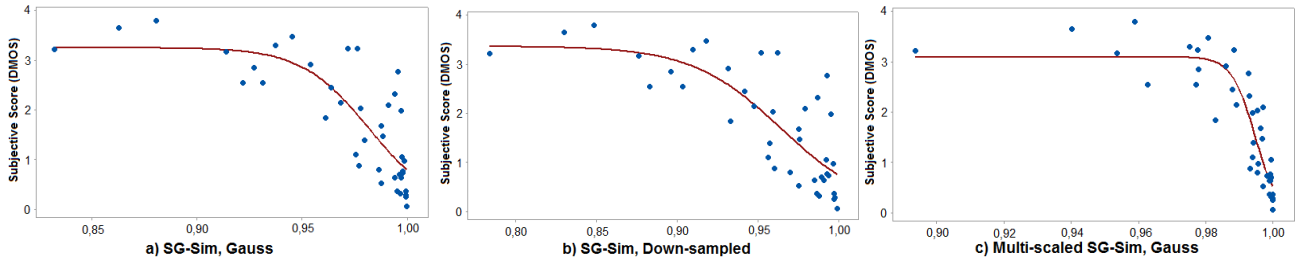

Fig. 2. Scatter plots of quality scores predicted by three SG-Sim variations versus the actual subjective scores (DMOS). (a) SG-Sim with 7x7 Gaussian pooling. (b) SG-Sim with 5x5 Down-sampled Box pooling. (c) Five-scaled SG-Sim with 7x7 Gaussian pooling.

significant balance of perceptual correlation and complexity, the shifted gradient pooled by our second proposal, the down-sampling box filter, is 54% faster and statistically equivalent in perceptual correlation to GMSD, while also four times faster than variance with Gaussian.

Multi-scaling was found to improve the shifted gradient statistic as well as its established improvement to variance, and four scales achieves the best quality and some of the best computing times. In fact, four-scaled SG-Sim pooled by down-sampling is only matched in speed by the unfiltered metric, which has poor quality, and produces significantly higher quality than MS-SSIM.

Substitution of constants in division stabilization by logical treatment, however, was found to improve neither computational complexity nor perceptual correlation. Also, the poor results for unfiltered shifted gradient in all criteria highlight the importance of spatial pooling in the index.

VII. Conclusions

This paper proposes the Shifted Gradient Similarity VQA metric, along with the optimized down-sampling box pooling filter, as a significant improvement over established SSIM variations in correlation to subjective opinion scores as well as in computational complexity, both of which are critical concerns for quality of service in Web streaming. We demonstrate that studying the behavior of the responses of VQA metrics yields mathematical improvements which, however simple in nature, significantly improve performance both in quality assessment and in computational complexity. This work has also produced a flexible, light-weight and open-sourced VQA framework to facilitate experimentation.

The down-sampling box pooling filter is expected to allow scalability over a broad range of resolutions and data rates, though this remains to be verified. Future work may also investigate how SG-SIM compares to different, also effective metrics such as VQM [23] and FSIM [24].

References

[1] M. Levkov, *Video encoding and transcoding recommendations for HTTP Dynamic Streaming on the Flash Platform*: preliminary recommendations for video on demand, Adobe Systems, Oct. 2010.

[2] W. Xue, L. Zhang, X. Mou, A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index", *IEEE Trans. Image Processing*, vol. 23, pp. 684-695, Feb. 2014.

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

[4] D. Rouse and S. Hemami, "Understanding and simplifying the structural similarity metric," in *IEEE International Conference in Image Processing*, pp. 1188-1191, Oct. 2008.

[5] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment", in IEEE Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1398-1402, Nov. 2003.

[6] G. Chen, C. Yang and S. Xie, "Gradient-based structural similarity for image quality assessment," in *IEEE International Conference in Image Processing*, pp. 2929-2932, Oct. 2006.

[7] C. Li and A. C. Bovik, "Content-weighted video quality assessment using a three-component image model", *Journal of Electronic Imaging*, vol. 19, 011003, Jan.–Mar. 2010.

[8] C. Li and A. C. Bovik, "Content-partitioned structural similarity index for image quality assessment", *Signal Processing: Image Communication 25*, pp. 517-526, 2010.

[9] M. Chen and A. C. Bovik, "Fast structural similarity index algorithm," *Journal of Real-Time Image Processing*, vol. 6, pp. 281-287, Dec. 2011.

[10] K. Seshadrinathan and A. C. Bovik, "Motion-based perceptual quality assessment of video", in *IS&T/SPIE Electronic Imaging 2009*, Feb. 2009.

[11] A. K. Moorthy and A. C. Bovik, "Efficient Motion Weighted Spatio-Temporal Video SSIM Index", in *SPIE 7527*, Human Vision and Electronic Imaging XV, 75271I, Feb. 2010.

[12] D. Rouse and S. Hemami, "Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM," in *SPIE Human Vision and Electronic Imaging XIII*, Feb. 2008.

[13] Ben Waggoner, *Compression for Great Video and Audio: Master Tips and Common Sense*, Focal Press, Nov. 2009.

[14] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-Motivated Rate-Distortion Optimization for Video Coding", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, Apr. 2012.

[15] L. Merrit, *X264: a high performance H.264/AVC encoder*, Washington University, 2006.

[16] J. R. C. Patterson. (2012, April). *Video encoding settings for H.264 excellence* [Online]. Available: http://www.lighterra.com/papers/videoencodingh264

[17] D. Vatolin, D. Kulikov, and M. Arsaev, *MPEG-4 AVC/H.264 Video Codecs Comparison*, Graphics & Media Lab Video Group, Moscow State University, May 2012.

[18] J. Garret-Glaser. (2008, May 31). *Psy RDO: Official testing thread* [Online]. Available: http://forum.doom9.org/showthread.php?t=138293

[19] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, pp. 98-117, Jan. 2009.

[20] Z. Wang, and A. C. Bovik, "A universal image quality index", *IEEE Signal Processing Letters*, vol. 9, pp. 81-84, March 2002.

[21] A. K. Moorthy, L. K. Choi, A. C. Bovik and G. deVeciana, "Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies", *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652-671, Oct. 2012.

[22] Video Quality Experts Group, "Draft final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II", Mar. 2003.

[23] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312-313, Sep. 2004.

[24] L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment", *IEEE Transactions on Image Processing*, vol. 20, pp. 2378-2386, Aug. 2011.