Toward Generalized Psychovisual Preprocessing For Video Encoding

Aaron Chadha

Mohammad Ashraful Anam

Matthias Treder

Ilya Fadeev

Yiannis Andreopoulos

iSIZE, London, UK, www.isize.co

Written for presentation at the

SMPTE 2021 Annual Technical Conference & Exhibition

Abstract. *Deep perceptual preprocessing has recently emerged as a new way to enable further bitrate savings across several generations of video encoders without breaking standards or requiring any changes in client devices. In this paper, we lay the foundations toward a generalized psychovisual preprocessing framework for video encoding and describe one of its promising instantiations that is practically deployable for video-on-demand, live, gaming and user-generated content. Results using state-of-the-art AVC, HEVC and VVC encoders show that average bitrate (BD-rate) gains of 11% to 17% are obtained over three state-of-the-art reference-based quality metrics (Netflix VMAF, SSIM and Apple AVQT), as well as the recently-proposed non-reference ITU-T p.1204 metric. The proposed framework on CPU is shown to be twice faster than x264 medium-preset encoding. On GPU hardware, our approach achieves 714fps for 1080p video (below 2ms/frame), thereby enabling its use in very-low latency live video or game streaming applications.*

Keywords. Perceptual optimization, deep neural networks, video delivery.

Introduction

Numerous studies in the last 10 years have shown that signal-to-noise (SNR) ratio is a poor indicator of visual quality in video coding. Instead, perceptual metrics that encapsulate elements of human perception [1]-[3], perceptual modelling of encoding artifacts [4], as well as viewing setup awareness [3][5] have emerged as strong contenders for the best means to objectively characterize visual quality. This has also led to the research community now moving away from SNR-optimization [6] in favour of structural similarity (SSIM [5]), video multimethod assessment fusion (VMAF) and Apple's advanced video quality tool (AVQT) optimization [1][3][7]-[9]. However, all current perceptual optimization approaches have one or more of the following detriments: ■ they require multiple encoding passes or in-loop implementation within a specific encoder; ■ they only optimize for a single metric like VMAF or SSIM and are shown to be detrimental in other quality metrics; ■ they comprise hand-crafted shallow models of low-level human perception and fail to encapsulate several characteristics of human vision in a data-driven and learnable manner; ■ their inference complexity makes them prohibitively costly for real-time application on live content or game streaming systems.

These limitations lead us to postulate that a generalized video preprocessing framework must include the following five principles in order for it to be practically applicable in video coding and streaming systems: *(i) psychovisual tuning* – i.e., inspired by (and encapsulating) known principles of human vision in a data-driven (i.e., learnable) manner; *(ii) multi-metric gains* – it must allow for gains over multiple quality metrics (which should include well-established distortion-oriented metrics like SSIM, as well as perception-oriented metrics like VMAF and AVQT), and/or be shown to lead to perceptual improvement in P.910 (or similar) tests; *(iii) cross-content and cross-codec applicable* – offering compounded gains over other optimization frameworks like content-adaptive or convex-hull encoding; *(iv) low delay* – i.e., it must allow for single-pass processing per encoding resolution/bitrate, or, even better, comprise a single-pass model for multiple resolutions/bitrates; *(v) low complexity* – its inference complexity must be analogous to low-complex encoding, e.g., AVC x264 medium-preset encoding or similar, in order to allow for generalized applicability on a variety of hardware.

To this end, we present an instantiation of a deep neural network based framework that meets the above five principles. Our experiments with four different perceptual quality metrics show that it not only offers consistent bitrate savings over multiple state-of-the-art encoders for various content types, but it also allows for very efficient inference runtime on CPU and GPU hardware, thereby making it applicable for a wide range of deployment scenarios.

Proposed Generalized Psychovisual Preprocessor

We present the training and deployment phases of our proposed psychovisual preprocessing framework in Figure 1. The purpose of the training phase is to optimize a preprocessor on a set of fidelity, perceptual and rate-oriented loss functions, such that the output of the preprocessor is a rate-constrained and perceptually-aligned representation of the input. Then, during the deployment phase, only the trained preprocessor needs to be deployed as a front-end to any standard codec, in order to provide rate savings on top of the encoder.

The training phase is shown in Figure 1 in its open loop configuration. Essentially, there are three main components: the preprocessor that we wish to optimize, a virtual codec and a psychovisual model. The preprocessor is simply a convolutional neural network, that processes

a video sequence $\boldsymbol{V} = \{\boldsymbol{x}_1 \cdots \boldsymbol{x}_t, \boldsymbol{x}_{t+1} \cdots \boldsymbol{x}_N\}$ in chunks of size $m + 1$. By processing a chunk of frames, the preprocessor is able to learn both spatial and temporal dependencies. The temporal information allows for the preprocessor to better distinguish between different types of content, from static and unchanging to fast motion. The preprocessor only operates on the luminance channel of frames, scaled to range [0,1]. During training, n replicas of the preprocessor are used to map n reference frames to preprocessed reference frames $\boldsymbol{R} = \{\boldsymbol{p}_{t-1}, \cdots, \boldsymbol{p}_{t-n}\}$. These reference frames are passed, together with the preprocessed current frame $\boldsymbol{p}_t$, to the next training component; the virtual codec. The purpose of the virtual codec is to provide some level of encoder awareness to the preprocessor, such that rate savings achieved during training translate to deployment (where the virtual codec would be replaced with any standard codec). To this end, we emulate the core components of a standard video coding pipeline: block-based motion compensated prediction and transform and quantization of residual frame information. Block based prediction is performed between the current frame and multiple hypotheses $\boldsymbol{R}$, where essentially the spatial displacement vector is extended temporally over more than one reference frame during motion compensation. The importance of multiple hypotheses has been long established, and generally increases codec efficiency. We further note that setting $n = 1$ defaults to a single reference setting. The residual frame output of multi-hypothesis prediction can thus be transformed, quantized and encoded with a differentiable entropy model that acts as a proxy to context-adaptive entropy encoding for rate computation. We denote the rate loss function L_R and this is optimized on the preprocessor weights. The reconstructed frame $\hat{\boldsymbol{p}}_t$ is generated by summing the predicted and quantized residual frames.

The objective of our preprocessor is to jointly remove information from the content that costs rate, while also maintaining areas of perceptual importance, such as those with high motion or contrast. In this way, we not only save rate, but ensure visual imperceptibility between the preprocessed and reconstructed frame $\hat{\boldsymbol{p}}_t$ and the source $\boldsymbol{x}_t$. To this end, we implement a psychovisual model, which attempts to filter the frames and isolate key areas of perceptual distortion in the reconstructed frame. As shown in the left part of Figure 1, the frames are first mapped to a sparse block-based frequency representation with the discrete wavelet transform (DWT). The frequency subbands are filtered with a contrast sensitivity function (CSF) model, which weights down low-contrast pattern stimuli and higher spatial frequencies that would not be visible to an observer. The subbands are additionally processed with a contrast-masking function that isolates areas of high contrast where codec distortions would be less visible. Finally, we apply a learned transform on top of the filtered source and reconstructed subbands, which maps the subbands to a measure L_P that represents the perceptual distance between the source and our preprocessed (and reconstructed) frames. The preprocessor is optimized on L_P during training together with the rate loss L_R and a fidelity loss L_F, such as mean squared error (MSE), which ensures that $\hat{\boldsymbol{p}}_t$ does not deviate too far from the source $\boldsymbol{x}_t$.

Once the preprocessor has been jointly optimized over the perception-rate-distortion plane, both the virtual codec and psychovisual model can be discarded. Given that the preprocessor has been trained in a generalized manner with the virtual codec, it can now be deployed with any standard codec and settings, thus offering compounded gains over other optimization methods such as content-adaptive or convex-hull based encoding. The multi-frame nature of the input means the preprocessor is more adaptive to different types of content and can operate as a single model with single-pass over all resolutions and bitrates. This ensures that its inference complexity is amortized over multiple encodings of the same content. In this work, all results are obtained with a single model with $m = n = 3$. We are able to lower the complexity of the

preprocessor substantially via quantization-aware training, where both weights and activations are quantized to int8; this also allows for deployment on neural processing units (NPUs).

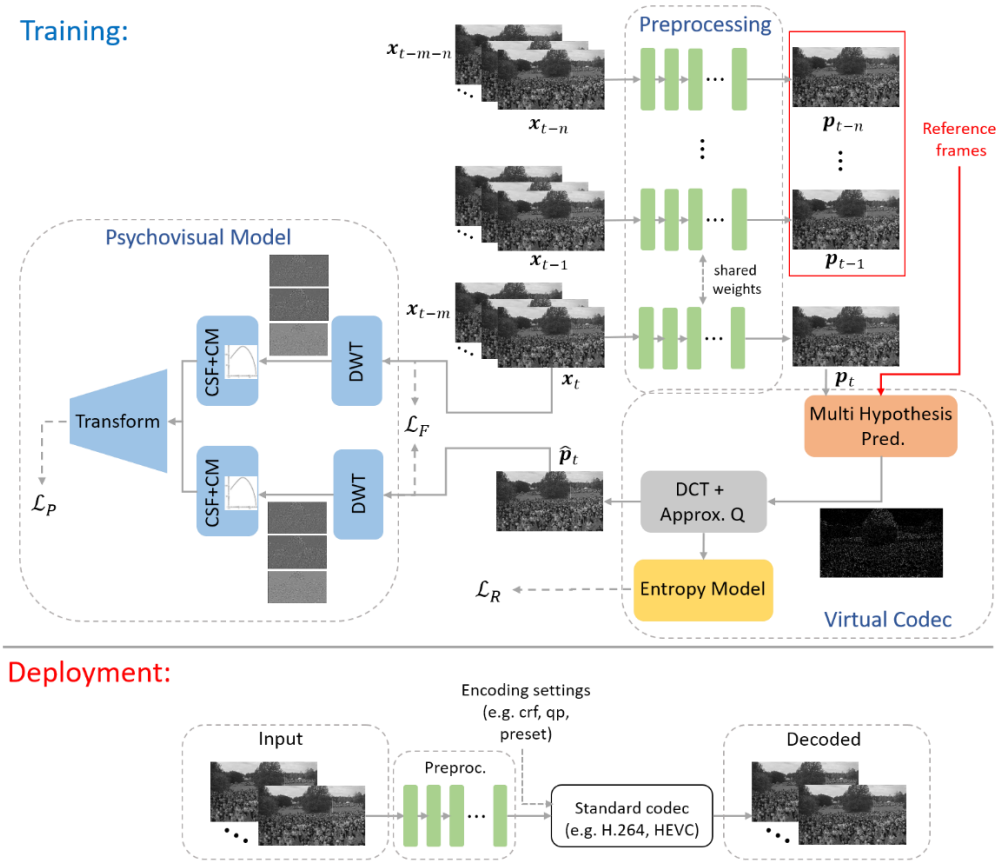

Figure 1. Generalized psychovisual preprocessor framework: training and deployment phases.

Experimental results

We evaluate with state-of-the-art implementations of AVC (x264), HEVC (x265) and VVC (vvenc1.0.0), representing three generations of hybrid encoders. There is no customization required per encoder and similar results have been obtained with AOM encoders but are omitted for brevity of exposition. Since psychovisual preprocessing can be applied to many categories of content and under diverse encoding conditions, we opt to report results with three categories of content: XIPH 1080p video (park_joy, red_kayak, controlled_burn, pedestrian_cross, ducks_take_off, touchdown_pass, rush_hour, crowd_run), YouTube UGC 1080p 'LiveMusic' and 'Sports' [representing user-generated content (UGC), 56 sequences] and UGC 1080p 'Gaming' representing game streaming applications (43 sequences)[1]. The x264 and x265 recipes utilized the veryslow preset, and we used the slow preset for vvenc. For all cases we used multiple CRF values and fixed resolution (1080p). We have verified that similar gains are found when applying in tandem with VoD optimizations like scene-adaptive or convex-hull encoding [10], but opt to report on a simpler test setup that applies to both live and VoD streaming. The use of 1080p

[1] XIPH test material from media.xiph.org/derf/, UGC content from media.withyoutube.com

resolution also represents the typical resolution users tend to receive from premium online services. All Bjontegaard delta-rates (BD-rates) are produced by first measuring with the libvmaf, AVQT and p.1204 tools (as provided by Netflix, Apple and TU Ilmenau) and then using the BD-rate calculator of libvmaf. To avoid skewing the BD-rate averages from the use of very-low or very-high bitrate points (which would not be used in practical streaming), we keep results within the range of 50% to 96% of each metric's range by limiting the BD-rate integration limits. In order to expand the narrow scoring range of SSIM, we rescale it according to the FB-MOS proposal [11], which was shown to appropriately scale SSIM to the entire [0,1] range.

In terms of preprocessor settings, the CSF filtering assumes 1080p resolution viewed at 3H distance, Daubechies-2 filters for the DWT, CNN kernel sizes are 3x3, our network structure is 4 conv layers with relu, and the bitrate range we train for maps roughly to 0.5-10mbps. Finally, the the utilized fidelity loss is a trained regularization between MSE, SSIM, and a DWT-based detail loss measure, while the perceptual loss is a measure over a learned transform on top fo the filtered source and the reconstructed subbands of the perceptual model.

The results for each case of content are presented in Table 1. All experiments show that the proposed psychovisual preprocessing approach leads to BD-rate gains for all metrics, all encoders, and all content types. Importantly, metrics include: the distortion-oriented SSIM and VMAF-NEG scores, perception-oriented metrics (VMAF, AVQT), and no-reference metrics (p.1204). Within each content type, savings are higher for more challenging content (sports, gaming, high-motion music videos), where motion-compensated prediction in encoders struggles to encapsulate the scene dynamics.

BD rates (%)	XIPH VoD/Live Premium Content			UGC Music/Sports	UGC Gaming
Metric/Encoder	AVC	HEVC	VVC	AVC	HEVC
SSIM	-8.1	-8.4	-5.7	-16.6	-16.8
AVQT	-6.2	-7.9	-10.2	-16.5	-17.8
p.1204	-6.1	-8.9	N/A	-0.8	-17.5
VMAF-NEG	-11.4	-10.1	-6.8	-17.4	-1.1
VMAF	-20.2	-21.6	-19.6	-34.0	-26.8
Numerical average:	-10.4	-11.4	-10.6	-17.1	-16.0

Table 1. BD-rate averages in percentages (%) for different types of 1080p content and various encoders. More negative numbers indicate higher average saving versus the corresponding encoder. The ITU-T p.1204 metric does not support VVC bitstreams.

Importantly, these gains are obtained with a single-pass, low-delay, model prior to encoding with all encoders at all encoding CRFs/bitrates. Tests on an Intel Xeon 8275cl 24-core CPU using the OpenVINO inference engine and the int8 quantized model showed that we can process 218 frames per second (fps), while AVC x264 (preset=medium, crf=20) reaches up to 114 fps, both operating with over 92% CPU utilization. Equivalent inference experiments on an NVIDIA T4 GPU reach 714 fps for the proposed preprocessing, which corresponds to less than 2ms inference time per 1080p frame. These properties demonstrate that our approach meets the five principles postulated for a generalized psychovisual preprocessing system applicable to both VoD and live content.

In order to compare our proposal with existing methods from the literature, we compare with a recently-used sharpening recipe [9] that obtains similar BD-rate results to tune-vmaf recipes of encoders like aomenc [7], as well as our previous proposal that uses two models for different regularization coefficients between the loss functions. The results, shown in Table 2, demonstrate that sharpening approaches cannot provide for universal BD-rate gains across different metrics since, unlike the proposed approach, they change the perceptual characteristics of the source. Moreover, while previously-proposed preprocessing [8] offers BD-rate gains for SSIM and VMAF, it is outperformed by the present approach.

BD rates (%)	AVC			HEVC		
Metric	Sharpen [9]	Low-λ [8]	High-λ [8]	Sharpen [9]	Low-λ [8]	High-λ [8]
SSIM	24.8	-4.1	-0.7	34.4	-3.2	-0.4
AVQT	15.4	-3.7	-0.5	35.4	-0.9	-2.1
p.1204	21.1	-5.5	-4.8	23.7	-6.1	-4.3
VMAF-NEG	9.5	-6.1	-7.3	22.9	-4.4	-4.7
VMAF	-12.7	-20.2	-13.6	-11.6	-16.7	-14.4

Table 2. BD-rate averages on the XIPH content for sharpening and two previous deep perceptual preprocessing models [8]. Positive numbers indicate average bitrate increase and more negative numbers indicate more saving versus the corresponding encoder.

Finally, on-going work is showing that the obtained savings are also validated by rigorous controlled tests with human viewers. Specifically, we have recently carried out P.910 absolute category rating (ACR) tests with hidden reference. Our current P.910 experiments used: 5 resolutions (1080p, 720p, 540p, 360p, 216p, multiple CRFs per resolution), two different encoder technologies (AVC and VP9, x264 and libvpx-vp9, both at their slowest preset), and 21 test videos from the AV2 CTC test set (all at 1080p resolution, encapsulating a wide variety of UGC, entertainment, sports and gaming content). In terms of preprocessing+encoder vs. encoder, our current results show that the proposed psychovisual preprocessing framework obtains -11.5% BD-rate over AVC and -20.5% over VP9 on the selected test content and encoders+recipes, with the best agreement to experimental mean opinion scores obtained by the VMAF metric. We plan to elaborate further on these tests in a future publication.

Conclusion

Deep perceptual preprocessing has the potential to bring significant bitrate savings in video delivery without imposing any changes in the encoding, packaging, transport and decoding side. We postulate that generalized forms of such preprocessing should allow for: psychovisual tuning, multi-metric gains across content types and for various coding standards, and low-delay & low-complexity implementation. A promising instantiation of such preprocessing is shown to offer over 10% average bitrate reduction over premium, gaming and UGC video. Unlike sharpening-based approaches, its learnable psychovisual aspects ensure its BD-rate savings are indeed confirmed by four different quality metrics that score for a wide range of perception-distortion characteristics. The low-delay and low-complexity characteristics of our proposal make it attractive for large-scale deployment over CPU, GPU or NPU hardware.

Acknowledgement

The work for the P.910 experiments was supported in part by Innovate UK [SEQUOIA project (grant number 96984)]. Yiannis Andreopoulos is Director at iSIZE and also Professor at University College London (UCL), U.K. All technical proposals of this paper comprise intellectual property of iSIZE that is under patent-protected or patent-pending status.

References

[1] "VMAF, code repository." online at: github.com/Netflix/vmaf.
[2] Z. Li. Video @Scale 2020: VMAF. *At Scale Conference*, 2020.
[3] "Apple AVQT presentation", *Apple Worldwide Developers Conference* (WWDC) 2021.
[4] A. Raake, *et al*. "Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P. 1204," *IEEE Access* 8: 193020-193049 (2020).
[5] A. K. Venkataramanan, *et al*. "A hitchhiker's guide to structural similarity," *IEEE Access* 9 (2021): 28872-28896.
[6] Q. Huynh-Thu, and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters* 44(13), 800–801 (2008).
[7] S. Deng, *et al*. "VMAF based rate-distortion optimization for video coding," *Proc. IEEE 22nd International Workshop on Multimedia Signal Processing* (MMSP). IEEE, 2020.
[8] A. Chadha, and Y. Andreopoulos, "Deep perceptual preprocessing for video coding," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 14852-14861, 2021.
[9] A. Chadha, *et al*., "Escaping the complexity-bitrate-quality barriers of video encoders via deep perceptual optimization," *Applications of Digital Image Processing* XLIII. Vol. 11510.
[10] I. Katsavounidis, and L. Guo, "Video codec comparison using the dynamic optimizer framework," *Applications of Digital Image Processing* XLI , 10752, 107520Q (2018).
[11] S. L. Regunathan, *et al*., "Efficient measurement of quality at scale in Facebook video ecosystem," *Applications of Digital Image Processing* XLIII. Vol. 11510, 2020.
[12] ITU-T P.910, *Subjective video quality assessment methods for multimedia applications*, Recommendation ITU-T: 910.