A SUBJECTIVE STUDY OF FILM GRAIN SYNTHESIS FOR THE PRESERVATION OF CREATIVE INTENT

Jatin Sapra, Kai Zeng, Hojatollah Yeganeh

(jatin.sapra@ssimwave.com, kai.zeng@ssimwave.com, hojat.yeganeh@ssimwave.com)

SSIMWAVE Inc., Canada

ABSTRACT

Despite the advancements in digital cinematography, numerous artists and filmmakers still adore the look and feel of the content that is shot on film rolls. Specifically, they believe in true film grain as a signature of motion pictures and thus they treat grain as a key part of their artistic intent. The natural randomness of the true film grain comes from the crystallisation of silver halide when exposed to the light, and this natural randomness of true film grain is what fascinates content creators. However, content distributors like OTT providers and streamers always have trouble with such a high entropy signal since randomness possesses challenges to compression. Content distributors have limited bandwidth; they always try to squeeze videos into the pipes as much as possible. A clever and well thought approach to cope with grainy content is to remove the grain at the source side and then synthesise the grain after decoding the compressed videos. Recently developed codecs such as AV1 and VVC provide end-to-end solutions to achieve this goal; however, the faithfulness of the grain with respect to the creative intent is subject to thorough validation and deep investigation.

We believe that while the proposed framework is technically sound, without looking at the problem from a perceptual video quality point of view, the synthesised film grain will likely not satisfy film makers and content creators' pursuit for the look and feel they intend to convey. To support this hypothesis, we have conducted a subjective study using content with film grain. In order to create different Hypothetical Reference Circuits (HRCs), standard film grain synthesis techniques like auto-regression models were used to produce different levels of grain with AV1 codec. The subjective data proves that there is still a big gap in the proposed models by the available codec standards.

INTRODUCTION

Film grain is a characteristic texture that appears in traditional film photography, caused by the random distribution of silver halide crystals in the emulsion layer of the film. It is a result of the light-sensitive chemicals on the film reacting to the light that passes through the camera lens during filming (1). This texture can vary in size and intensity, depending on the type of film used, the lighting conditions during filming, and the camera settings. Film grain can be an important visual element for Hollywood directors as it can contribute to the overall look and feel of a film. It can add a sense of texture, depth, and authenticity to the image,

and can also help to create a certain mood or atmosphere. In addition to its aesthetic qualities, film grain can also be used strategically by directors to control the brightness and contrast of an image. By adjusting the amount of grain present in the image, a director can subtly alter the look and feel of the scene, enhancing certain elements or drawing attention to specific areas of the frame.

A large number of film grain synthesis algorithms have been proposed in the past decades, but little work has been done to quantify the perceptual quality of the synthesised film grain. In practice, researchers often use either subjective evaluations, where a group of viewers are recruited to rate the quality of the grain, or common video quality assessment (VQA) objective metrics, such as the peak signal-to-noise-ratio (PSNR) and the structural similarity index (SSIM), but proper validations of these measures are missing.

Both subjective and objective VQA methods can be employed to assess the quality of the synthesised film grain. In a subjective experiment, multiple human subjects are asked to rate or rank the quality of the synthesised grain for mean opinion score (MOS) collection. Subjective methods are highly valuable in comparing grain synthesis algorithms and in validating objective VQA methods, but they are often very time consuming and costly. Depending on the accessibility to the original source that is assumed to have perfect quality, objective VQA measures can be classified into full-reference (FR), reduced reference (RR) and no-reference (NR) methods. Objective models can be employed to evaluate the grain quality automatically, and can also be embedded into the design and optimization of various grain processing algorithms and systems. Notable success has been achieved in all three categories, especially in the FR case, where a number of state-of-the-art algorithms have been shown to have good correlations with subjective quality ratings. The FR quality metric is the primary study topic of this paper, because it is consistent with the philosophy of preserving creative intent.

In this work, we focus on the perceptual quality assessment of synthesised film grain in terms of its fidelity to the grain in source. We first create a database that contains different levels of synthesised grain, together with multiple compression levels, and carry out a subjective study using the database. Comprehensive subjective score analysis is conducted to comparatively study the behaviour of different grain synthesis methods. We find that state-of-the-art VQA models only moderately correlate with subjective opinions. Closer examinations reveal that popular deterministic VQA approaches such as VMAF and AVQT lack appropriate considerations on the statistical naturalness of the grain. This provides potential guidelines for designing a more effective objective grain fidelity metric in the future.

BACKGROUND

Film grain removal and synthesis has been proved to be an effective tool in the video encoding system to save large bandwidth while maintaining the perceptual quality of the encoded video. Figure 1 shows a general framework of the grain-aware video encoding. The film grain is estimated and removed from the source video before encoding and added back to the encoded video after decoding based on the estimated grain statistics.

In the video encoding industry, the primary film grain synthesis algorithms can be categorised into frequency filtering (FF) and auto-regression (AR) approaches. The grain statistics, that needs to be embedded into the supplemental enhancement information (SEI) message of the encoded bitstream, has been standardised in the ITU-T H.274 (2), so that the synthesis can be done after decoding

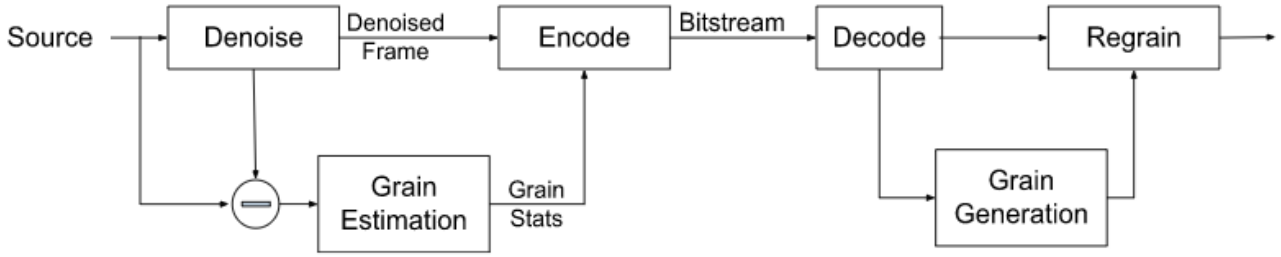

Figure 1: Framework for grain-aware video encoding

on the user device. In AV1, a special AR method (3)(4) was adopted to synthesise the film grain after decoding, given the grain statistics estimated using the Wiener filter on the encoder side. It has to be noted that those methods were adopted in the standards mainly due to their implementation efficiency, where the perceptual quality of the generated grain needs to be examined. In the academic publications, lots of work has been done to generate the film grain with its appearance to be as close to the true film grain as possible. Oh et al. (5) described an advanced method for film grain extraction and synthesis for high-definition video coding. Dai et al. (6) proposed to remove film grain using temporal filtering before encoding and add back to the decoded video after being synthesised using an autoregressive model. Hwang et al. (7) proposed to add inter-color dependency for the grain removal before encoding. Newson et al. (8) presented a film grain rendering algorithm using a Monte Carlo simulation method to generate physically realistic film grain. Ameur et al. (9) used a deep convolutional neural network for synthesising realistic film grain. From the quality evaluation of film grain perspective, most of the film grain synthesis methods have their own implicit metric to quantify the performance of the proposed algorithms. Kim et al. (10) used a SSIM-kind distortion metric to quantify the film grain noise in HEVC coding applications. Visual example inspection had been performed in most of the film grain removal and synthesis publications to justify the effectiveness of the proposed methods (6)(7)(9). Therefore, an objective perceptual quality metric for film grain is urgently needed to quantify the performance of the existing and future grain processing methods.

SUBJECTIVE EXPERIMENT DESIGN

To the best of our knowledge, there is no publicly available dataset that is dedicated to study the preservation of film grain. The purpose of this study was twofold: firstly, to accumulate a library of compressed videos with various levels of artificially generated film grain that could be used as a foundation for developing a texture similarity measure, and secondly, to comprehensively investigate the components of the AV1 film grain synthesis pipeline and their impact on the final decoded frames with the inclusion of film grain. This research was crucial because AV1 was the only available pipeline designed to retain film grain from the original content, and its film grain synthesis feature was met with minimal satisfaction from filmmakers and the video industry, who doubted its capability to conserve creative intent.

Contents

To assess subjective video quality, ten 5-second-long video sequences were selected for the study. The sequences were chosen to have a variety of film grain looks, five from Netflix Open Content (11) and five proprietary content. All but one of the source sequences include true or practically synthesised film grain in the content. The remaining sequence had no film

grain and was included to explore and compare how humans perceive similarity when dealing with content with film grain versus those without it.

The videos were all in UHD (3840x2160) resolution and were in HDR format. The videos had a frame rate of 24 frames per second. The use of UHD and HDR allowed for the inclusion of high-quality content that would be representative of current media consumption trends. The selection of video sequences with a range of film grain levels provides a diverse set of stimuli for participants to rate and allows for a more comprehensive understanding of the perception of film grain similarity. Overall, the selection of video sequences with different film grain shape, size and intensity, along with the use of UHD and HDR, provides a strong foundation for investigating subjective video quality assessment in the current media landscape.

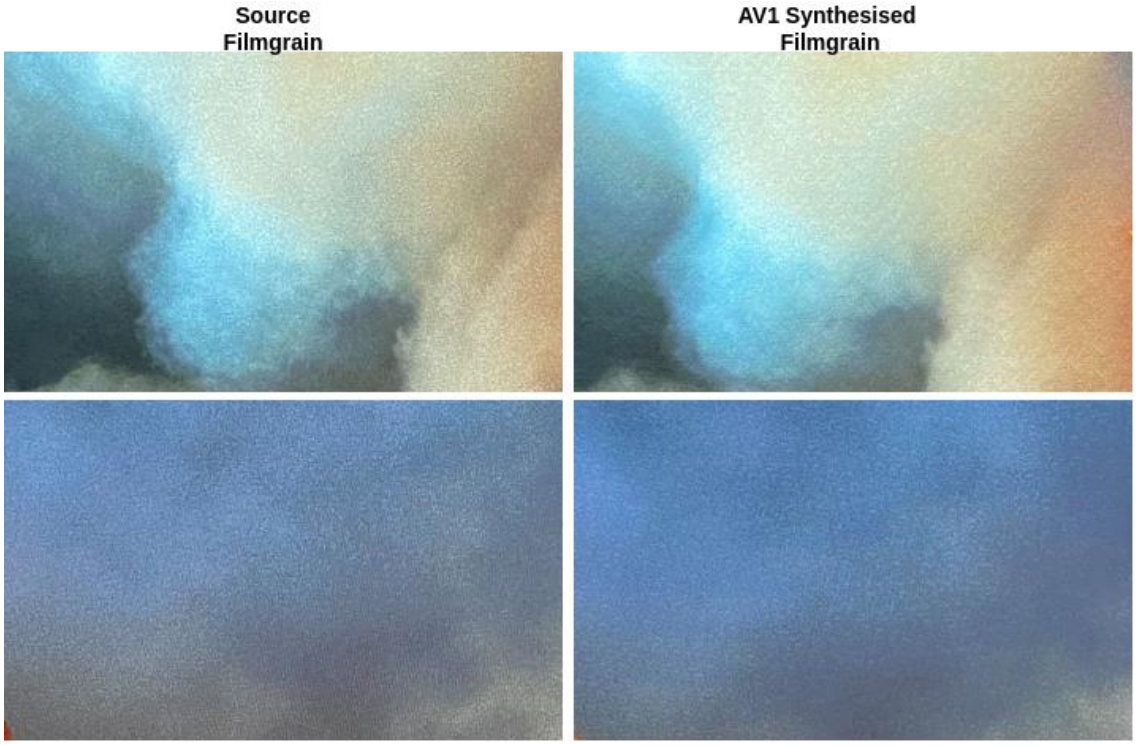

Figure 2: Examples of periodic pattern in AV1 synthesised film grain

Test Sequences

We prepared test videos or Processed Video Sequences (PVSs) while keeping in mind two main objectives of the subjective study. To reduce the number of variables, we limit the codec to be AV1. We generate test videos using the AV1 film grain synthesis framework. The test sequences are generated by using the industry standard AV1 as well as tweaking and modifying the components of the AV1 pipeline manually. While the idea of AV1 film grain synthesis is very elegant, the main assumption here is to not treat AV1 as the finalised and optimum solution. This hypothesis has been validated after our initial observations about the fully automated AV1 film grain synthesis framework, and encouraged us to generate alternative video streams by modifying the two main components in the film grain synthesis pipeline: the denoising/de-graining module and selecting the auto-regression parameters that generate the synthesised film grain on rendition.

Therefore, the test sequences are generated using two main pipelines: industry AV1 standard, and the modified framework where the goal is to replace and modify the components of AV1 film grain synthesis solution and thus generate a variety of perceived film grain.

In the modified framework, the main replacements and changes have been performed on the initial stage that is the noise estimation module and the post decoding stage where Auto-Regression (AR) algorithm is used to synthesise the film grain. More specifically, to try different denoising/de-graining component, we use IMAX's proprietary product known as Digital Media Remastering (DMR) and the Block-Matching and 3D filtering denoising algorithm (BM3D) to denoise frames and then analyse the residual between the original and the denoised frames to understand the statistics of content film grain. We speculate that the AV1 built-in denoising algorithm is not special in this case either. The AV1 codec is used to generate three different compression levels by exploiting the Constant Rate Factor (CRF) rate control module. One of the compression levels is equivalent to lossless. The other two CRF values are chosen so that the perceptual quality of the encodes are in the mid and low range. To restore grain to decoded content, we apply the AV1 codec's auto-regression (AR) algorithm and utilize an open source tool offered by InterDigital (13) to generate three levels of film grain: Less (subtle), Similar (close to source), and More (intense). Obviously, these levels are selected subjectively by the conductors of the study and they do not necessarily reflect the absolute similarity of the grain with respect to the source.

Score	Description for scores
10	Undistinguishable/Same
9	Very Similar
8	
7	Similar
6	
5	Different
4	
3	Very Different
2	
1	Totally Different

Table 1: Impairment scale for grain similarity

In summary, there are three Hypothetical Reference Circuits (HRCs) categories. The first category includes one source with no film grain and includes three compression levels. The second category uses the AV1 automated pipeline, with three compression levels. The third category adopts three denoising algorithms (BM3D, DMR configuration 1 and DMR configuration 2), three compression levels, and three re-graining levels. Therefore, the total number of processed video sequences (PVSs) or equivalently test sequences is 273.

Subjective Study Method

The study was conducted at SSIMWAVE's lab at Waterloo. An LG C2 65" 4K OLED evo with ThinQ AI TV (12) was used to conduct the study. The TV was calibrated for HDR while disabling all advanced processing options.

The double-stimulus impairment scale (DSIS) method was used in this study. The double-stimulus method is cyclic in which subjects would watch the pristine source video first then the impaired version of the video following which they were asked to provide Film Grain

Similarity (FGS) score. The impairment scale here follows Absolute Category Rating (ACR) convention that uses 10-level scores from 1 to 10 corresponding to the intervals ”very different”, ”different”,”similar”, ”very similar”, and ”undistinguishable”.

The experiment setup was in a dark room that corresponded to a lab-study environment, following the recommendation in BT. 2022 (15) for critical viewing of HDR content and the recommendation in BT. 500 (14) for general viewing conditions for a subjective study in a laboratory environment. The viewing distance was about 1.5 x H, where H was the height of the TV. The whole study was divided into 4 sessions which lasted upto 30 minutes and after each session there was a 10 minutes break to avoid eye fatigue.

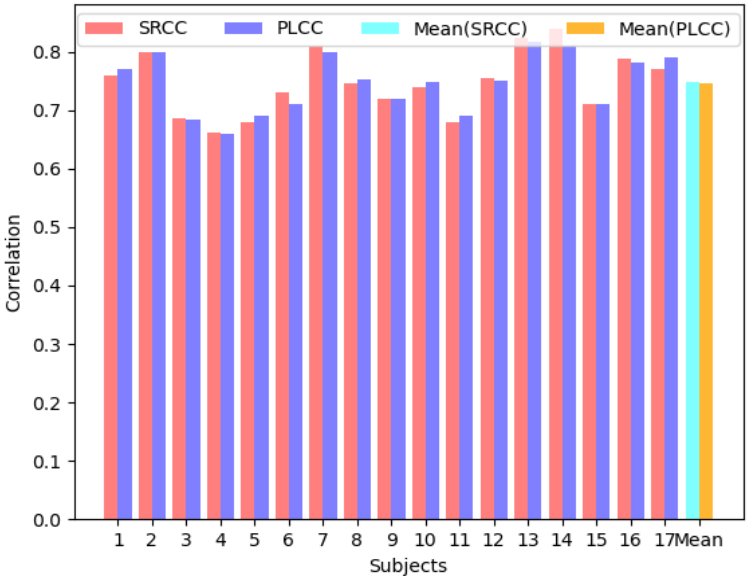

Figure 3: SRCC and PLCC b/w Subject’s scores and MOS

A total of 20 human subjects who are considered as experts in video quality assessment were recruited from SSIMWAVE, IMAX and the University of Waterloo’s IVC lab. A training session is performed before the formal experiment, in which 2 videos different from those in the formal experiment are shown. The same methods are used to generate the videos used in the training and testing sessions. Therefore, before the testing session, subjects knew what distortion types would be expected. Subjects were instructed with sample videos to judge the film grain similarity based on distortion level.

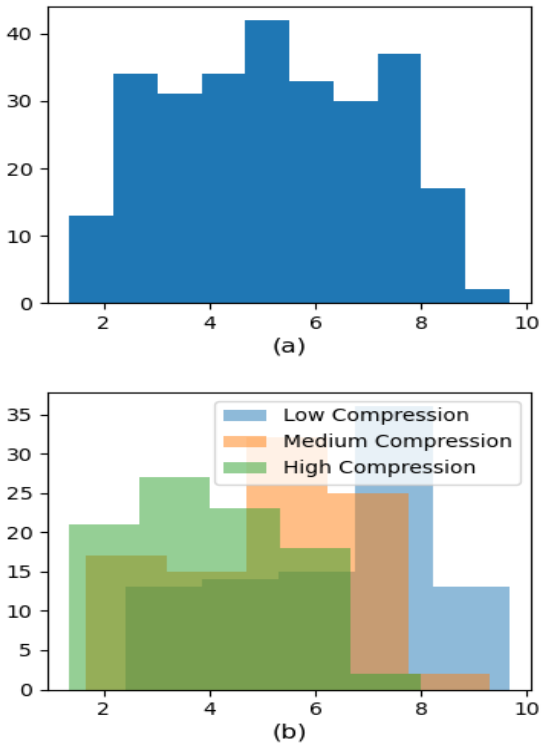

Figure 4: MOS histogram for (a) Full dataset (b) Various compressions

ANALYSIS OF SUBJECTIVE RESULTS

Subjective Score Processing

The present study aims to investigate the perceived film grain similarity of the test video sequences with respect to the sources. To this end, the subjective scores obtained from a group of participants were first converted into Z-scores per subject to normalise any potential variations in the use of the quality scale among the participants. An outlier removal process was then employed as suggested in ITU Rec BT. 500 (14). The resulting Z-scores were linearly re-scaled to fall within the range of 1 to

10. The mean opinion score (MOS) for each individual video was calculated as the average of the re-scaled Z-scores, which were obtained from all valid participants. In addition, we identified three outliers by calculating the Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) between the scores given by each participant and the MOS. These outliers were removed from the analysis since their scores were significantly low compared to the others. Following the removal of outliers, the average PLCC and SRCC across all participants were found to be 0.74 and 0.75, respectively, with standard deviations of 0.05 and 0.04, respectively. Figure 3 demonstrates the agreement among the participants regarding the perceived film grain similarity of the test video sequences.

Observations and findings

The distribution of the MOS is illustrated in Figure 4(a). The plot shows that the test sequences provide a wide range of film grain similarity from low scores that indicates a low grain similarity between the test sequences and the corresponding sources to high scores which means viewers see that the creative intent with a focus on film grain is fairly preserved.

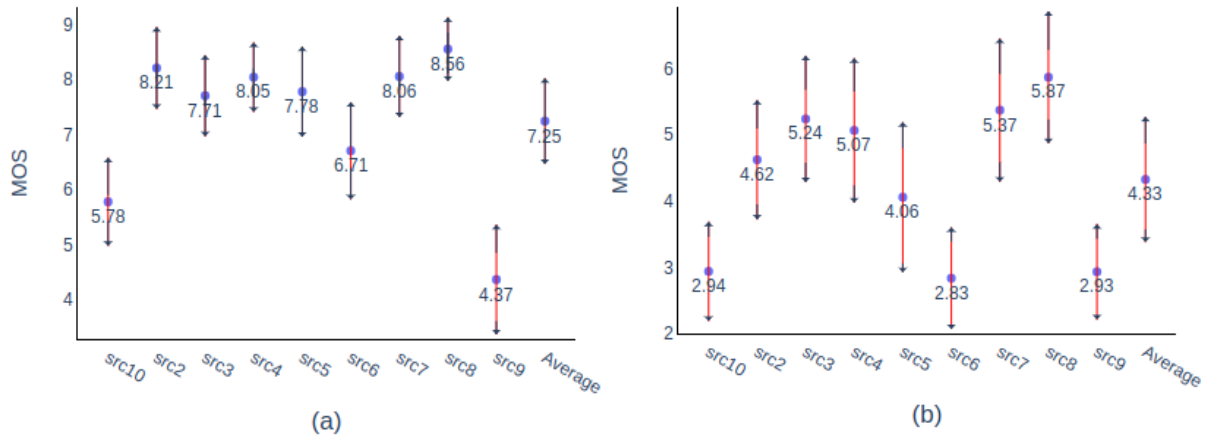

Figure 5: (a) MOS distribution for low compression (b) MOS distribution for high compression

Figure 5(a) and 5(b) illustrate the distribution of the scores given to the same synthesised film grain level along with confidence intervals for two very low and very high CRF values. The plots suggest that even though the subjects have seen a high similarity between the source and the test film grain textures, as the compression level increases, the subjects become more uncertain about the similarity of the film grain. Moreover, in general they rate the similarity of the grain significantly lower when the compression impairments are high. This implies that while the idea of film grain synthesis by adding metadata to the video bitstream is smart, the added value of performing such synthesis becomes questionable when the compression level is relatively high. Basically, the structural distortion will dominate the subjective experience and so the preservation of the film grain will not be of a big concern.

Figure 4(b) also shows that overall the subjects tend to give low scores to film grain similarity when the compression level is high which may indicate that performing a great job in estimating the statistical characteristics of the film grain prior to encoding and carry the metadata to synthesise the film grain in post decoding may not necessarily result in a pleasing experience. In other words, when the compression is high, preserving the look and feel of the intended film grain is not of a concern for end viewers.

Is AV1's FGS suboptimal?

Analysis of the subjective data provides insights about the idea of the film grain synthesis and the AV1 framework as one of the early adopters. The subjective data shows that the AV1 end-to-end solution does not always preserve the intended film grain, and so overall, expert viewers rate the similarity of the synthesised film grain resulting from the modified framework higher. Figure 6 illustrates the MOS for the videos with low compression level. As discussed in the last section, subjects are more confident in rating the similarity of the content with less compression impairment. To evaluate the grain synthesis frameworks, we

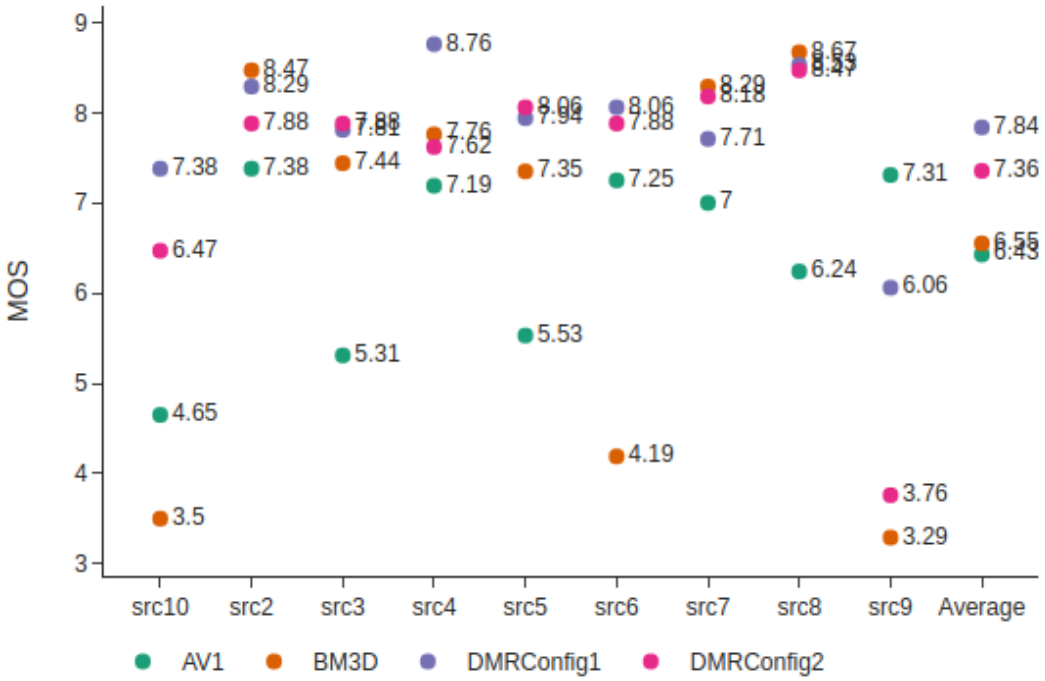

Figure 6: MOS distribution for different pipelines

investigate the MOS given to the processed test videos generated from 4 different pipelines. It is evident that none of the pipelines is superior and so they are all suboptimal. However, the MOS for different contents show that fully automated AV1 framework is overly ranked last compared to the other three.

By interviewing the subjects, we found out that there are two main reasons for giving lower film grain similarity scores to the AV1 framework. First is because of seeing repeated grain patterns. Figure 2 provides a few patches taken from AV1 test videos along with patches extracted from corresponding locations in the source sequences. As is shown in the figure, the extracted patches from AV1 videos exhibit periodic patterns which do not look quite natural. Secondly, subjects unanimously mentioned that the videos resulting from the AV1 framework look more blurry compared to the other test sequences and some details are

missing by comparing the AV1 sequences to the source. This may be because of the upfront denoising/de-graining operation that AV1 uses to estimate the statistical characteristics of the film grain and so strong denoising may compromise some of the fine details or structures as well. Figure 7 demonstrates two extracted patches from the sources that contain details and grain as well as the corresponding patches taken from the AV1 test videos and the modified framework that uses DMR as the denoiser.

Does Denoising technique make any difference?

As discussed in the previous section, the denoiser in the film grain synthesis has an important role and so choosing the right denoiser has a big impact on the final result. Further analysis on Figure 6 reveals that the lowest MOS values are given to the test videos generated by either AV1 framework or the modified framework that uses BM3D algorithm as the upfront denoiser. Please note that Figure 6 is derived from the subjective scores given to the test videos with almost no compression artifact and with the same level of synthesised film grain. We believe that this significant difference is because of the blurriness introduced by the BM3D algorithm and the AV1 denoiser which badly impacts perception of added film grain.

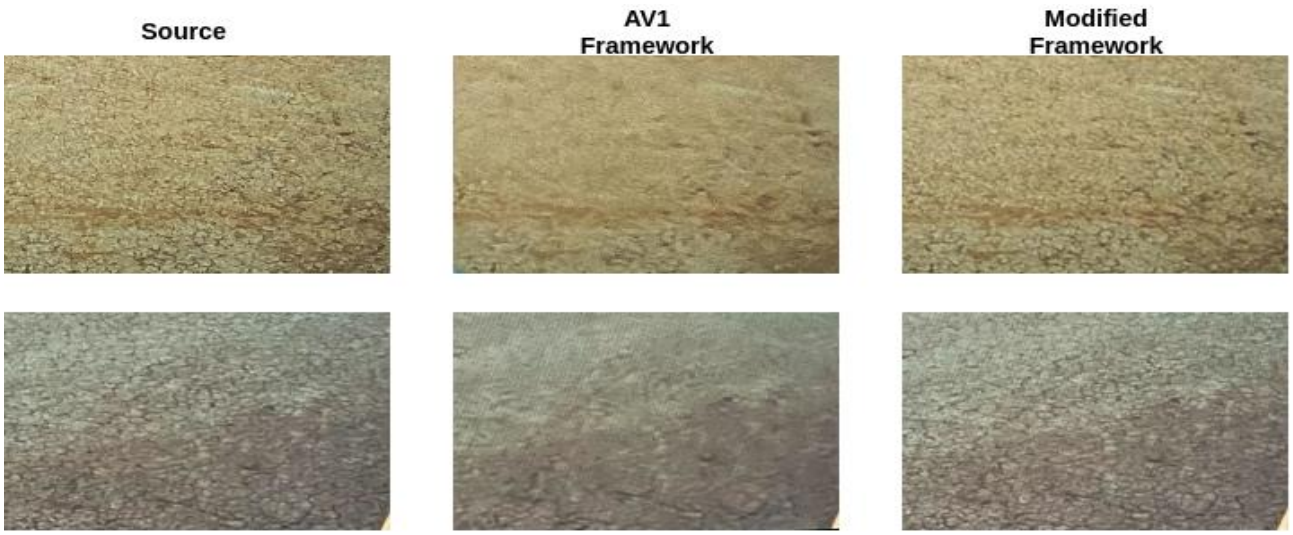

Figure 7: Patches demonstrating detail loss using AV1 Framework

Despite observing deviations in the subjective opinion for the different film grain synthesis pipelines, there are three contents, i.e. src 6, src 9 and src 10, that have received significantly lower scores. By re-watching these sequences, it was noticed that all three of them include heavier grain in terms of the density, shape and the size and apparently the AR algorithm is not doing a decent job at synthesising the original grain look.

EVALUATION OF OBJECTIVE QUALITY METRIC

Classical Full reference metrics are designed based on measuring structural or signal deviations between the source and the test images/videos. They typically operate on pixels or a group of pixels and penalise the calculated distance from the source. Despite the advancements in designing objective quality metrics for images and videos in recent decades, the applicability of such metrics in the context of film grain synthesis is very limited. This is due to the fact that the human visual system may distinguish between the structural signals and the statistical patterns and perceive it separately. More specifically, the human visual system (HVS) is very sensitive when it comes to preserving the structural details in

an image or a video signal, and detects and penalises any structural impairments. However, it tends to comprehend the overall statistical characteristics of texture-like signals such as film grain. To put it differently, the HVS (Human Visual System) penalizes the overall difference between the statistical properties of film grain and does not consider the precise location where the grain appears. Figure 8 illustrates the performance of a few well-known objective metrics that are designed for image and videos on the film grain subjective dataset. Apparently, such metrics are not designed to measure the similarity between two signals with different statistical characteristics and so all of them fail in predicting the MOS.

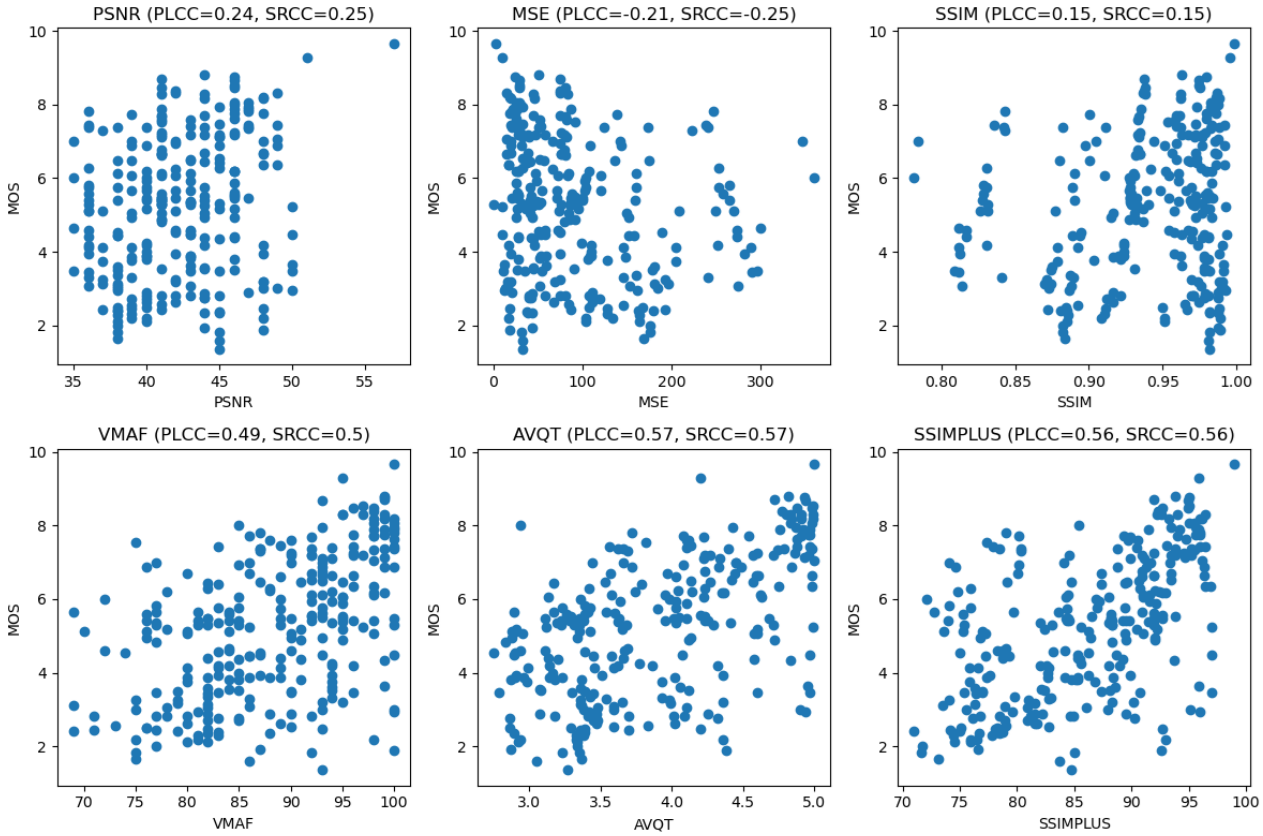

Figure 8: Performance of objective quality metrics

CONCLUSION

A subjective user study was conducted to investigate the effectiveness of the synthesised film grain at different levels of compression and grain characteristics in preserving the creative intent. In particular, we make one of the first attempts that focuses on understanding the interactions between the denoiser module, compression and the synthesised film grain levels. The analysis of the subjective data shows that while the idea of film grain synthesis in post decoder is very smart and has a great potential, the existing approaches are suboptimal. In particular, we believe that the most important missing piece in the film grain synthesis framework is a reliable objective metric that can measure the similarity between two film grain patterns. Without having such objective metric all the efforts in deriving and tuning the parameters of the existing grain synthesis pipelines are suboptimal and ad-hoc.

REFERENCES

1. JC. Kit Yan, and D. Hatzinakos, Signal dependent film grain noise removal and generation based on higher-order statistics. Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics, Banff, AB, Canada, 1997, pp. 77-81.

2. M. Radosavljevic, E. François, E. Reinhard, W. Hamidouche, and T. Amestoy, Implementation of film-grain technology within VVC. Applications of Digital Image Processing XLIV, 2021, vol. 11842.

3. A. Norkin, and N. Birkbeck, Film grain synthesis for AV1 video codec. 2018 Data Compression Conference, Snowbird, UT, USA, 2018, pp. 3-12.

4. A. Norkin, and N. Birkbeck, Technical report on AOMedia film grain synthesis technology. https://aomedia.org/docs/CWG-C051o_TR_AOMedia_film_grain_synthesis_technology_v2.pdf, July 8, 2022.

5. B. T. Oh, S. Lei, and C. C. J. Kuo, Advanced film grain noise extraction and synthesis for high-definition video coding. in IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 12, pp. 1717-1729, Dec. 2009.

6. J. Dai, O. C. Au, C. Pang, W. Yang, and F. Zou, Film grain noise removal and synthesis in video coding. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA.

7. I. Hwang, J. Jeong, S. Kim, J. Choi, and Y. Choe, Enhanced film grain noise removal and synthesis for high fidelity video coding. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2013.

8. A. Newson, N. Faraj, B. Galerne, and J. Delon, Realistic Film Grain Rendering. Image Processing On Line, 2017, pp. 165-183.

9. Z. Ameur, W. Hamidouche, E. Francois, M. Radosavljevic, D. Menard, and C. H. Demarty, Deep-based film grain removal and synthesis. Image and Video Processing eess.IV, 2022.

10. S. Kim, D. Pak, and S. Lee, SSIM-based distortion metric for film grain noise in HEVC, Signal Image and Video Processing 12(2), 2018.

11. Netlix Open Content, https://opencontent.netflix.com/

12. LG C2 65" 4K OLED evo with ThinQ AI, https://www.lg.com/ca_en/tvs/lg-oled65c2pua

13. Versatile film grain synthesis, https://github.com/InterDigitalInc/VersatileFilmGrain

14. ITU, BT.500 : Methodologies for the subjective assessment of the quality of television images, Tech. Rep., Intl. Telecomm. Union, 2019.

15. ITU, BT.2022 : General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays, Tech. Rep., Intl. Telecomm. Union, 2017.