Understanding Banding - Perceptual Modelling and Machine Learning Approaches for Banding Detection

Hojatollah Yeganeh[1], Kai Zeng[1] and Zhou Wang[1, 2]
[1] SSIMWAVE Inc., and [2] University of Waterloo
Waterloo, Ontario, Canada
hojat.yeganeh@ssimwave.com, kai.zeng@ssimwave.com, zhou.wang@uwaterloo.ca

Abstract – Banding is an annoying visual artifact that frequently appears at various stages along the chain of video acquisition, production, distribution and display. With the thriving popularity of ultra-high definition, high dynamic range, wide color gamut content, and the increasing user expectations that follow, the banding effect has been attracting a growing deal of attention for its strong negative impact on viewer experience in visual content that could otherwise have nearly perfect quality. Here we present two different types of frameworks to detect the banding artifact. The first is knowledge-driven and is built upon computational models that account for the characteristics of the human visual system, the content acquisition, production, distribution and display processes, and the interplay between them. The second is data-driven, and is based on machine learning methods, by training deep neural networks with large-scale datasets.

Keywoards: Banding impairment, contouring impairment, perceived video quality, human visual system

INTRODUCTION

Banding is a visual artifact that appears frequently at many stages in video acquisition, production, distribution and display systems. Banding typically appears as perceived discontinuities or false contours in large and smooth image regions of slow color or intensity gradients. An example is given in Figure 1, where severe banding effect is observed in the sky region. Although heavy video compression is a potential source of banding, banding may also occur in the absence of any lossy compression, and may create annoying visual quality degradations in otherwise pristine quality images or video content.

What often frustrates many industrial practitioners is that simply increasing the bit-depth or bitrate of a video does not necessarily lead to removal or even reduction of banding. Indeed, with the recent accelerated growth of ultra-high definition (UHD), high dynamic range (HDR), wide color gamut (WCG) in content production, distribution services, and consumer display devices, severe banding occurs even more frequently than before and the visual effect is often much stronger. This is because UHD/HDR/WCG content typically covers a wider range of luminance levels and color variations than those of the traditional standard dynamic range (SDR) content. This, together with the limited and varying capabilities of display devices, creates major challenges to maintain smooth visual transitions simultaneously across all luminance levels and color variations. Significant effort has been made over the years on removing or reducing banding effects in video distributions. Depending on where these banding reduction techniques are applied, they may be classified into pre-processing, post-processing and banding-aware encoding methods. However, without having a reliable objective measure to detect banding, improving dithering or de-banding efforts are quite cumbersome and directionless. Therefore, the industry is in an urgent need of innovative approaches that are able to detect, control, and remove/reduce banding in an automated fashion.

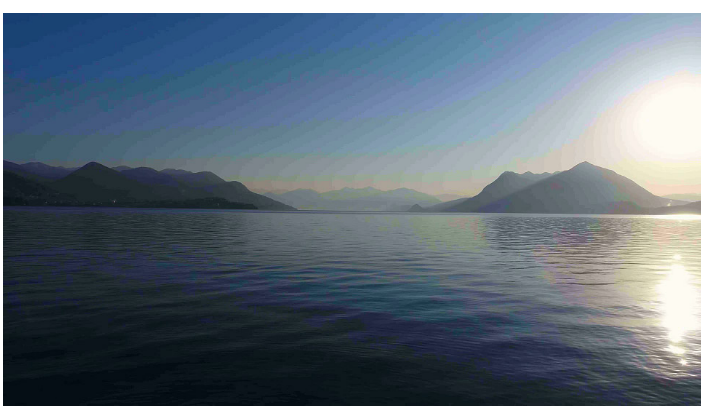


FIGURE 1 – A SAMPLE IMAGE OF VISUAL BANDING EFFECT (IN THE SKY REGION).

Automatic or objective image/ video quality assessment (IQA/VQA) has been a highly active topic in the past two decades. However, popular IQA/VQA methods such as PSNR [1], SSIM [2], MS-SSIM [3], SSIMPLUS [4], and VMAF [5], often require full access to a pristine reference when assessing a test image or video – and pristine references are rarely available in real-world testing environments. Moreover, although the quality maps created by SSIM types of approaches [2][3][4] often successfully capture local banding artifacts, the overall assessment of these approaches mixes all visual distortion types together and there is no simple mechanism to single out the banding effect. Therefore, it is desirable to develop novel IQA/VQA methods dedicated to detecting and assessing banding without access to a pristine-quality original image/video as reference.

Recently, two substantially different types of approaches have shown notable success at banding detection. The first is based on domain knowledge gained through deep and thorough understandings of the human visual system (HVS) and the video acquisition, production, distribution and display processes. Computational models derived from such domain knowledge are then combined to construct an overall banding detection and assessment model [12][13].

In contrast to the first type of domain knowledge-driven methods, the second type of approaches are data-driven, with no or little domain knowledge assumed. Instead, a large number of images/videos and their ground-truth labels (with or without banding) are collected, and machine learning methods are then used to train black-box models such as the deep neural networks (DNNs) using the image/video dataset, so that the learned model may make good banding predictions on unseen image/video content.

KNOWLEDGE-DRIVEN METHOD

There is a rich literature on computational modelling of HVS characteristics and the individual components in the sophisticated video acquisition, production, distribution and display processes. Knowledge-driven banding detection methods select relevant models and combine them in a systematic way, so as to produce a prediction of the perceived banding effect.

A knowledge-driven method [6] is illustrated in Figure 2, where the input is an image or a video frame at the pixel level, and the outputs are a banding score together with a banding map. The banding score denotes the overall level of perceived banding by considering two important factors: banding spread and banding strength. The spread of banding impairment impacts viewing experience but does not solely represent banding annoyance. The contrast sensitivity of the HVS varies based on multiple signal components and viewing conditions, and thus not every banded edge or false contour is perceived equally. The severity of banding is captured by the banding strength component. In other words, the goal here is to detect abrupt local activities in smooth image regions, and then analyze the visibility of such activities from the perspectives of HVS characteristics.

Pixels that correspond to abrupt activities deemed visible as banding are then marked, which collectively constitute a banding map of the image or video frame. The banding map illustrates the presence of banding impairment in an image or a video frame, and does not reflect the banding strength. Examples of such banding maps are shown in Figure 3 (right), where the banding artifacts are highlighted by the white pixels. It appears that this knowledge-driven approach not only detects banding, but also precisely localizes the banding impairment at pixel-precision.

Banding regions may be determined by pixels in the image or video frame that have significant local signal activity while the signal activity in a majority of its surrounding regions is not significant. Therefore, classifying pixels into significant and non-significant categories is the first step in detecting banding impairment. To classify pixels, a significant threshold is determined based on the characteristics of HVS as well as a series of signal and system properties. Figure 4 depicts the workflow of marking pixels based on their signal activity.

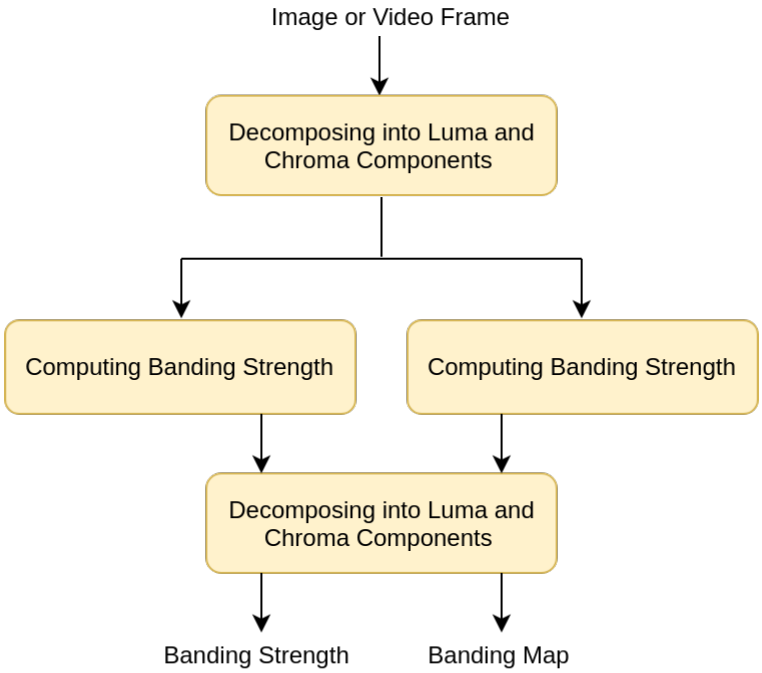


FIGURE 2 – DIAGRAM OF A KNOWLEDGE-DRIVEN METHOD.

Determining the significance threshold in Figure 4 is the key to detecting banding impairment. It requires a deep understanding of the HVS properties such as the contrast sensitivity function (CSF) [7] and various visual masking effects [8].

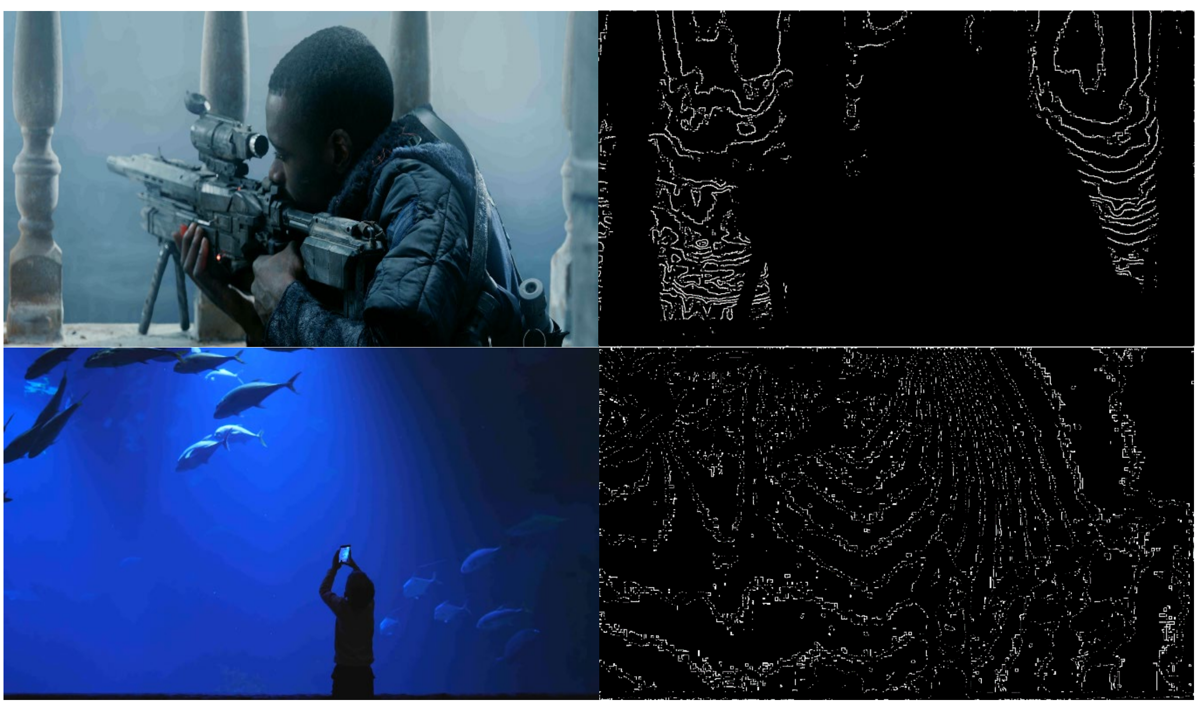


FIGURE 3 – SAMPLE IMAGES AND BANDING MAPS CREATED BY KNOWLEDGE-DRIVEN METHOD.

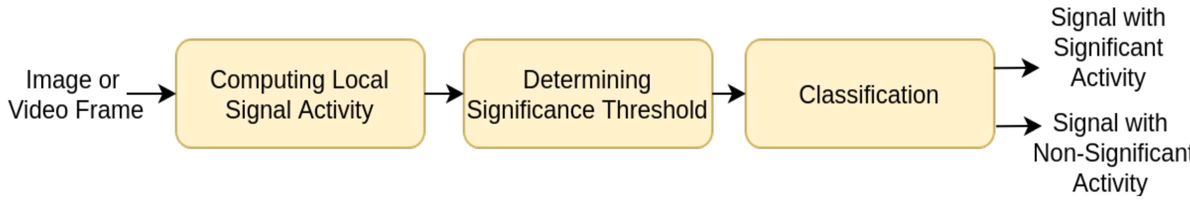


FIGURE 4 – CLASSIFICATION OF LOCAL SIGNALS BASED ON AN ACTIVITY THRESHOLD.

Figure 5 shows how the significance threshold is generated, starting with HVS modelling, and followed by adjustments based on important video workflow and display factors. Banding is a local activity in smooth image regions that is visible under certain conditions and viewing environments. Therefore, not only signal properties, but also display devices and viewing conditions affect perceived banding.

Modelling CSF and visual masking of the HVS provides a starting point in determining the significance threshold, which would need to be tuned further to determine precisely the visibility of banding as shown in Figure 5. The CSFs are typically derived based on psychophysical studies on the visibility of patterns with a varying luminance level and spatial frequency [7]. Chroma component has a significant impact on the visibility of a signal contrast [9], and may be used to adjust the initial significance threshold. Signal dynamic range is often interpreted as the ratio of the brightest to the darkest luminance. Different Opto-Electrical Transfer Functions (OETFs) and Electro-Optical Transfer Functions (EOTFs) are designed to accommodate signals with standard and high dynamic range using certain bit depths i.e. 8 and 10 for standard dynamic range (SDR) videos and 10 and 12 for high dynamic range (HDR) videos. The bit depth of the content determines signal precision and in conjunction with transfer functions impacts on the visibility of banding impairment and suggest adjustments to the significance threshold. Further adjustments to the significance threshold may also be required based on the maximum and minimum values of contents and the capabilities of display devices in producing an adequate range of luminance that is needed to avoid banding.

When all these content, perceptual, chroma, bit-depth, transfer function, and display factors are properly modelled, a precise prediction of visual banding may be achieved. Recently, technologies following this path have emerged and enjoyed growing adoption. The advantages of such knowledge-driven approaches is not only the high pixel-precision accuracy (that allows for the creation of banding maps) and low computational complexity, but also the high explainability – meaning that when banding happens, a deep investigation is plausible to find the cause of banding and then localize the problem to be fixed. The disadvantage of this approach is mainly in the difficulty of the modelling process itself, as precise models of the contributing factors are difficult to develop and parameters of such models are hard to calibrate.

DATA-DRIVEN/MACHINE LEARNING METHOD

Machine learning, and especially deep learning approaches, have attracted a great deal of attention in recent years and have achieved remarkable success in many application areas. These approaches are generally data-driven, with black box models being trained by data samples. When these data samples are sufficiently representative of the real world data distribution, the trained model may be strong enough to make good predictions on novel data samples unseen in the training dataset. The data-driven approach becomes a desirable option in the case of visual banding detection because it avoids the difficulty in developing and calibrating knowledge-driven models.

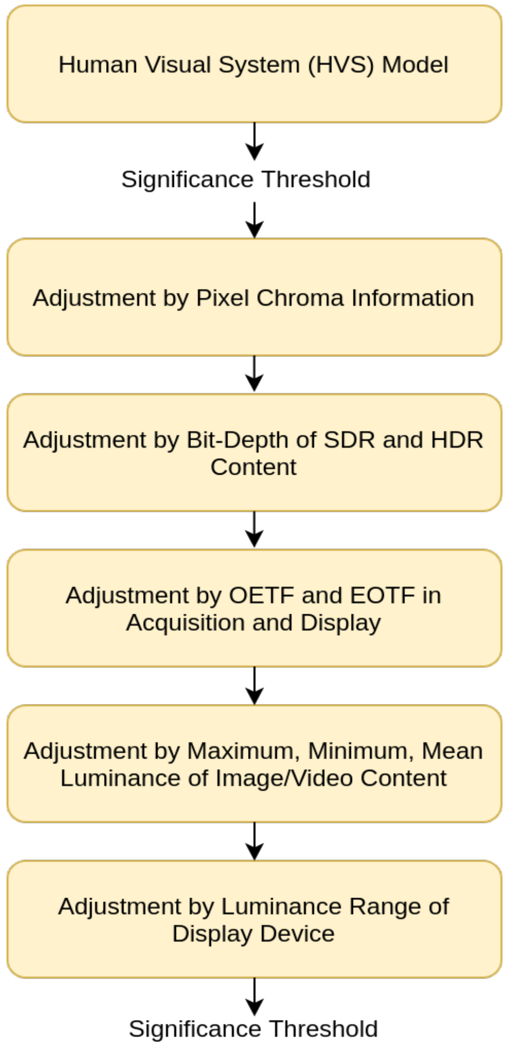


FIGURE 5 – COMPUTATION AND ADJUSTMENT OF SIGNIFICANCE THRESHOLD FACTORS

The first step in building a deep learning based method is to obtain "big data", i.e., to construct a large-scale dataset for training, validation and testing. Fortunately, such datasets have emerged recently. A dataset has been constructed [10], which is composed of nearly 17,000 image patches, together with their ground-truth labels, i.e., each image patch has been labelled to be either containing or not containing banding. This allows us to train a deep neural network (DNN), more precisely a convolutional neural network (CNN), to classify a given image patch as either with or without banding in an end-to-end manner. Such a method is end-to-end because the DNN takes a raw image patch of pixels as input and directly produces a classification result as output. As such, the feature extraction process and the classifier are trained or optimized all together (as opposed to being constructed separately in traditional image classification methods) by learning from data samples. By doing so, a model is constructed with no explicit domain knowledge. In other words, all knowledge is learned from data and stored in the weights of the trained CNN.

Figure 6 shows a diagram of how DNN-based banding patch classifier may be applied to assess visual banding of a given image or video frame [11]. Image patches are first extracted from the test image using a sliding window that moves pixel by pixel (or a larger stripe when computational cost is a concern) across the image space. The extracted patches are then fed into the DNN based banding patch classifier.

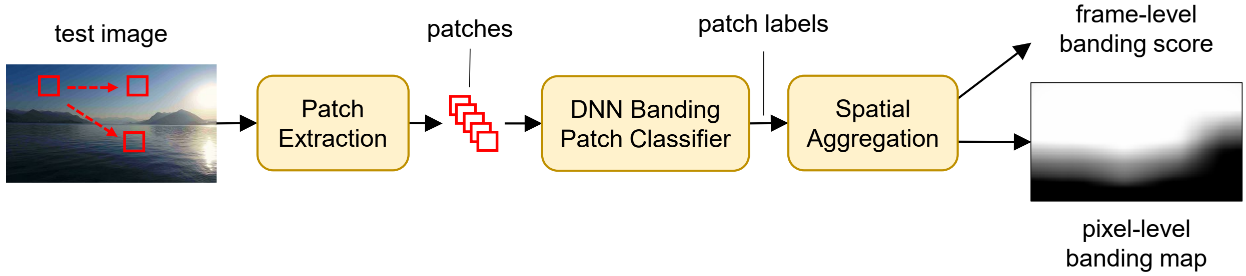


FIGURE 6 – DIAGRAM OF DNN-BASED VISUAL BANDING ASSESSMENT.

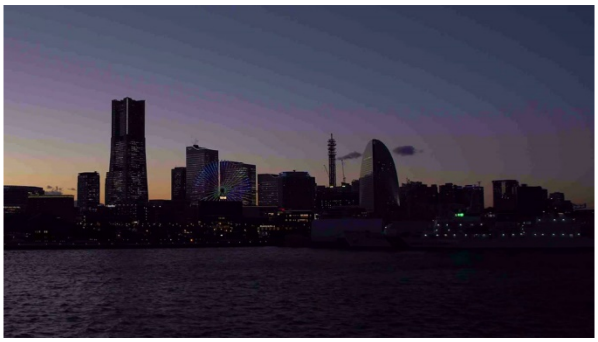
The DNN is typically implemented using a CNN structure, which contains multiple convolutional layers, followed by a fully connected neural network. In each convolutional layer, there are multiple linear kernels that convolve with the input signal, followed by activation and pooling processes. The outcome of the layer is considered intermediate features that are subsequently taken as the input to the next layer. Due to the pooling process, the number of features reduces over the layers. At the starting point of the fully connected layer, the features are aligned into a vector, which is fed into a neural network of multiple layers of node and with full connection between all nodes of adjacent layers. The final output of the fully connected layers produces the classification results. Once the CNN architecture is constructed, all that remains is to determine the parameters (including kernel weights in the convolutional layers and the connection weights in the fully connected layers) at each layer. This is achieved by a training process, in which ground truth outputs are compared with CNN output and the errors are back-propagated and used to adjust the weights in the CNN. When we have sufficient data samples for training, the CNN will converge to a stage that may make accurate classification results. The classification results obtained for individual local image patches through the CNN classifier are laid out spatially. They are then aggregated with pre-defined smoothness constraints into two outcomes, as shown in Figure 6. The first is a scalar frame-level banding score for the whole test image, and the second is a pixel-level banding map.

Sample images and their corresponding DNN-based banding scores and banding maps are shown in Figure 7. Images (a), (c), and (e) have ascending levels of visual banding, which are well reflected by the banding scores. The banding maps created by smoothed local patch level banding assessment are given in (b), (d) and (f). It is interesting to compare the banding maps created by the knowledge-driven approach and the DNN-based data-driven approach, as exemplified by Figure 3 and Figure 7, respectively. Both maps offer good predictions on the existence and the spatial locations of banding. The spatial information is very important, especially for subsequent methods that may be used to fix or reduce the banding problem. Comparatively, the maps created by the knowledge-driven method give much more precise localization of banding at the pixel level.

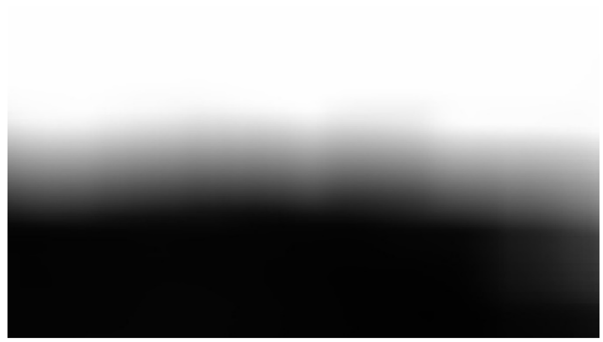
The advantage of the data-driven approach is mainly in the alleviation of the necessity to fully comprehend the domain knowledge, which is sophisticated and evolving over time. On the other hand, its disadvantages are manifold. First, it relies heavily on the quality and quantity of training data. The training sample images need to cover all test cases and the labels assigned to these images need to be very accurate. In general, the behavior of CNN models are difficult to predict on novel data samples, especially when the data samples contain novel content, distortion type or distortion combinations. This becomes a major issue when the setup of the video acquisition, production, distribution and display workflow changes, in which case we often need to collect new data and retrain the DNN model. Second, the quality prediction results may not be as precise as knowledge-driven models, as demonstrated by the comparison of banding maps between Figures 3 and 7. Third, the deployment of DNN models in practical systems may demand high computational and memory resources, especially when the use of the model requires precise localization and diagnosis, such as generating the banding maps shown in Figure 7, in which the DNN patch classifier needs to be applied to all sliding windows. Nevertheless, significant progress has been made in the past decade on accelerating DNN performance through advanced hardware and software design. Although the training process is usually time-consuming, once the model is trained, the application of the

(A) IMAGE WITH BANDING

(B) BANDING SCORE = 20



(C) IMAGE WITH BANDING



(D) BANDING SCORE = 67

(E) IMAGE WITH BANDING

(F) BANDING SCORE = 94

FIGURE 7 – SAMPLE IMAGES AND BANDING MAPS CREATED BY DNN METHOD.

trained model in real-world testing may be made very fast, especially when localized assessment such as production of banding maps is not required.

Challenges and Future Works

Despite the exciting progress made in the past decades on banding detection and reduction, there are still significant gaps in practice. Some of the root causes of these gaps are summarized as follows:

- First, there is no definite way to differentiate the real contours in the video content and the false contours of banding. Contours of various types exhibit in real-world videos: some are from camera acquisition of the visual world, and some are artificial, e.g., in animation and computer screen content. How to reliably and efficiently differentiate them remains a challenging problem.
- Second, while it is proven that dithering does reduce banding, there is a dilemma between banding reduction and preserving fine details in an image, because dithering involves adding noise to an image, and noise reduces the visibility of fine texture details. This dilemma becomes even stronger with UHD/HDR/WCG content, which is supposed to bring in finer details than SDR content of lower resolution and smaller colour gamut, but the details embedded in deeper depth bit-planes are even more sensitive to noise contamination. Furthermore, the noise in the

dithered content makes video encoding more difficult, as the noisy pixels consume a large number of bits to encode, leaving much fewer bits for true fine details of the original content.

- Third, there has been a debate about the objective of video distribution – whether we should aim for the preservation of the creative intent of the content producers or for creating appealing visual results of the end viewers. It is important to be aware that these two goals may not always align. If preserving the creative intent is the goal, then techniques such as dithering or pre-processing based banding reduction are problematic, because they purposely change the content.

All of these are entangled with the never-ending progress of camera and display technologies, and the recent development of scene-adaptive and device-adaptive HDR/WCG processing in the new HDR standards such as HDR10+ and Dolby Vision. Consequently, banding detection and reduction will remain an open problem in the future and will evolve with the technology front of the digital video industry.

CONCLUSION

In this paper, we focus on the banding effect, an annoying visual artifact that appears in all stages of the life cycle of digital videos, and that has been drawing an increasing amount of attention with the recent growing popularity of UHD/HDR/WCG video content. We discuss the technical details of two promising but substantially different types of approaches for banding detection – knowledge-driven approaches that are built upon deep understandings of the HVS and each component in the video acquisition, production, distribution and display workflows, and data-driven approaches that learn to detect banding by training DNN models with big data of labelled image samples. Our experiments and analysis demonstrate promising results and prediction using both of the mentioned frameworks.

REFERENCES

1. Wang, Z. and Bovik, A. C., "Mean squared error: love it or leave it? – a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98-117, Jan. 2009.
2. Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity". *IEEE Transactions on Image Processing*, 13(4), 600612, (2004).
3. Wang, Z., Simoncelli, E. P. and Bovik, A. C., "Multi-scale structural similarity for image quality assessment". *IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2003.
4. Rehman, A., Zeng, K. and Wang, Z., "Display device-adapted video quality-of-experience assessment," *IS&T/SPIE Electronic Imaging: Human Vision & Electronic Imaging*, Feb. 2015.
5. Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A. and Manohara, M., "Toward a practical perceptual video quality metric," *Netflix TechBlog*, Jun 2016.
6. Wang, Z., Yeganeh, H., Badr, A., and Zeng, K., "Image and video banding assessment," *US patent application*, 2020.
7. Barten P. G. J., "Formula for the contrast sensitivity of the human eye" *Proc. SPIE-IS&T* Vol. 5294:231-238, Jan 2004
8. Wang, Z. and Bovik, A. C., Modern image quality assessment, *Morgan & Claypool Publishers*, Mar. 2006.
9. Denes, G., Ash, G., Fang, H., Mantiuk, R., "A visual model for predicting chromatic banding artifacts", *Human Vision and Electronic Imaging (HVEI)*, 2019.
10. Kapoor, A., Sapra, J., and Wang, Z., "HD images dataset with banded and nonbanded region information," *https://zenodo.org/record/4513740*, 2021.
11. Kapoor, A, Sapra, J., and Wang, Z., "Capturing banding in images: database construction and objective assessment", *IEEE International Conference on Acoustic, Speech and Signal Processing*, Jun 2021.
12. Wang, Y., Kum, S., Chen, C. and Kokaram, A. ,"A perceptual visibility metric for banding artifacts." *In IEEE Int. Conf. on Image Processing*, pp. 2067-2071, 2016.
13. Tu, Z., Lin, J., Wang, Y., Adsumilli, B., and Bovik, A. C., "Bband index: A no-reference banding artifact predictor," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., pp. 2712–2716, 2020.