Signal Processing: Image Communication 63 (2018) 125–140

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Visually lossless coding in HEVC: A high bit depth and 4:4:4 capable JND-based perceptual quantisation technique for HEVC

Lee Prangnell

Department of Computer Science, University of Warwick, England, UK

ARTICLE INFO

Keywords:
HEVC
H.265
JND
Visually lossless coding
Perceptual video coding
Perceptual quantisation

ABSTRACT

Due to the increasing prevalence of high bit depth and YCbCr 4:4:4 video data, it is desirable to develop a JND-based visually lossless coding technique which can account for high bit depth 4:4:4 data in addition to standard 8-bit precision chroma subsampled data. In this paper, we propose a Coding Block (CB)-level JND-based luma and chroma perceptual quantisation technique for HEVC named Pixel-PAQ. Pixel-PAQ exploits both luminance masking and chrominance masking to achieve JND-based visually lossless coding; the proposed method is compatible with high bit depth YCbCr 4:4:4 video data of any resolution. When applied to YCbCr 4:4:4 high bit depth video data, Pixel-PAQ can achieve vast bitrate reductions – of up to 75% (68.6% over four QP data points) – compared with a state-of-the-art luma-based JND method for HEVC named IDSQ. Moreover, the participants in the subjective evaluations confirm that visually lossless coding is successfully achieved by Pixel-PAQ (at a PSNR value of 28.04 dB in one test).

1. Introduction

Just Noticeable Distortion (JND)-based visually lossless coding is presently of considerable interest in video coding and image coding research; for example, visually lossless compression is a core consideration in the emerging JPEG-XS still image coding standard. Focusing on video compression in the HEVC standard, JND-based video coding can profoundly reduce the perceptual redundancies that are present in raw YCbCr video data. Therefore, the number of bits required to store each pixel can be considerably reduced without incurring a decrease in the perceptual quality of the reconstructed video data. As such, burdens related to data storage, transmission and bandwidth can be reduced to an extremely high degree. JND is generally defined as the maximum visibility threshold before lossy compression distortions are perceptually discernible to the Human Visual System (HVS) [1,2]; JND has its roots in the Weber–Fechner law [3]. Even without considering JND, it is well known that raw YCbCr video data, for example, contains a high level of perceptually redundant information. To this end, the HEVC standard [4,5] includes a multitude of advanced video coding algorithms to achieve high efficiency spatiotemporal compression of raw video data. In the lossy video coding pipeline, spatial image coding (intra-frame coding) and also Group Of Pictures (GOP)-based spatiotemporal video coding (inter-frame coding) are initially employed to dramatically reduce spatiotemporal redundancies typically inherent in all raw video sequences. Intra prediction errors [6] and inter prediction errors [7] produce luma and chroma residual values [8]. The residual values are subsequently transformed into the frequency domain by integer approximations of the Discrete Cosine Transform (DCT) and the Discrete Sine Transform (DST) [9]. The transformed residual values are then quantised using a combination of Rate–Distortion Optimised Quantisation (RDOQ) and Uniform Reconstruction Quantisation (URQ) [10]. The DC transform coefficient and the low frequency and medium frequency AC transform coefficients contain the energy which is deemed as the most important in terms of reconstruction quality. Therefore, quantisation is designed to discard the least perceptually important AC coefficients (i.e., the high frequency, or low energy, AC coefficients); the degree to which high frequency AC coefficients are zeroed out is contingent upon the Quantisation Step Size (QStep). Lossless entropy coding of the quantised transform coefficients is performed by the Context Adaptive Binary Arithmetic Coding (CABAC) method; this is the stage at which the actual data compression takes place [11]. If high levels of quantisation are applied, this gives rise to a decrease in non-zero quantised coefficients, which means that the CABAC entropy coder can compress the quantised coefficients more efficiently; that is, the compressed bitstream after entropy coding will contain fewer bits.

With a focused concentration on lossy video coding in the JCT-VC HEVC HM reference codec [12], the video coding algorithms in HEVC HM are based primarily on rate–distortion theory. Consequently, visual quality measurements in HEVC lossy video coding applications are founded upon the Mean Squared Error (MSE) [13]; that is, the MSE of the reconstructed pixel data compared with the raw pixel data. It is a

E-mail address: l.j.prangnell@warwick.ac.uk.

https://doi.org/10.1016/j.image.2018.02.007
Received 18 August 2017; Received in revised form 12 February 2018; Accepted 12 February 2018
Available online 16 February 2018
0923-5965/

well established fact that the Peak Signal-to-Noise Ratio (PSNR) – which is a logarithmic visual quality metric based on MSE – has a very poor correlation with human visual perception. This is primarily due to the fact that MSE is categorised as a simple statistical risk function; it is often employed in the field of statistics for calculating the average of the squares of the deviations [13]. Therefore, it is considered to be an overly simplistic measuring tool for computing the perceptual quality of compressed video data.

In addition to the primary objective of improving coding efficiency, most lossy video coding algorithms employed in HEVC HM are designed with an emphasis on increasing the PSNR values in the compressed video data. These algorithms include Rate–Distortion Optimisation (RDO) [14], RDOQ [15], Deblocking Filter (DF) [16] and Sample Adaptive Offset (SAO) [17]. Note that RDO, RDOQ, DF and SAO are effective methods in terms of increasing PSNR values for the reconstructed video; however, the PSNR-based mathematical reconstruction quality improvement attained by these techniques is perceptually negligible in terms of how the human observer interprets the perceived quality of the compressed video data. For instance, several studies have shown that a compressed video with a PSNR measure of 40 Decibels (dB), or above, typically constitutes visually lossless coding. That is, a coded video with a PSNR $\geq$ 40 dB is perceptually indistinguishable from the raw video data. Furthermore, using the example of PSNR $\geq$ 40 dB for visually lossless coding, this also implies that targeting a reconstruction quality of PSNR $\geq$ 40 dB (e.g., PSNR = 50 dB) is superfluous; i.e., unnecessary bits would be wasted by achieving the superior mathematical reconstruction quality required for the PSNR = 50 dB measurement.

The key difference between JND-based video coding and video coding based on rate–distortion theory is as follows: JND techniques prioritise, above all else, the human observer with respect to assessing the reconstruction quality of a coded video. That is, instead of focusing purely on mathematically-orientated visual quality metrics including PSNR. This is because, in the end, the human observer is the ultimate judge of the visual quality of a compressed video sequence. As such, human subjective quality evaluations are critically important in terms of assessing the reconstruction quality of video sequences coded by JND-based methods. JND techniques are primarily concerned with the following core objective: To reduce bitrates, as much as possible (i.e., reduce the number of bits required to store each pixel), without incurring a perceptually discernible decrease in visual quality in the compressed video data. Note that with JND and visually lossless coding, PSNR measurements are not considered to be important in terms of quantifying the perceptual quality of a reconstructed sequence. In such cases, the PSNR metric is utilised for quantifying the degree to which PSNR values can be decreased before the associated compression-induced distortions in the coded video are perceptually discernible.

The vast majority of JND techniques in video compression applications target the spatiotemporal domain, the frequency domain or a combination of the two. Mannos' and Sakrison's pioneering work in [18] formed a useful foundation for all frequency domain luminance Contrast Sensitivity Function (CSF)-based JND techniques which target HVS-based redundancies in luminance image data. Chou's and Chen's pioneering pixel-wise JND method in [19,20] formed the basis for several spatiotemporal domain JND contributions. The primary means by which Chou and Chen achieved pixel-wise JND are luminance-based spatial masking, contrast masking and temporal masking.

1.1. Overview of related work

In [21], Ahumada and Peterson devise the first DCT-based JND technique, in which a luminance spatial CSF is incorporated. In [22], Watson expands on Ahumada's and Peterson's work by incorporating luminance masking and contrast masking into the spatial CSF (in the frequency domain); note that power functions corresponding to Weber's law are utilised in this method. Chou and Chen develop a pioneering pixel-wise JND profile in [19], in which luminance masking and contrast masking functions are proposed for utilisation in the spatial domain (8-bit precision luma component); this method is based on average background luminance and also luminance adaptation. The authors further expand on this method in [20] by adding a temporal masking component, in which inter-frame luminance is exploited. Yang et al. in [23] propose a pixel-wise JND contribution to eradicate the overlapping effect between luminance masking and contrast masking effects. This technique also includes a filter for motion-compensated residuals, in which they employ a modified version of Chou's and Chen's spatiotemporal domain JND methods. In [24], Jia et al. present a DCT-based JND technique founded upon on a CSF-related temporal masking effect. Wei and Ngan in [25] introduce a novel DCT-based JND method for video coding, in which the authors incorporate luminance masking, contrast masking and temporal masking effects into the technique. The luminance masking component is modelled as a piecewise linear function. The contrast masking aspect is contextualised as edge and texture masking; the temporal masking component quantifies temporal frequency by taking into account motion direction. Chen and Guillemot in [26] propose a spatial domain foveated masking JND technique, which is the first time that image fixation points are taken into account in JND modelling. Moreover, this method also incorporates the luminance masking, contrast masking and temporal masking functions from Chou's and Chen's methods in [19,20]. In [27], Naccari and Mrak propose a JND-based perceptual quantisation method (named IDSQ) which exploits luminance CSF-related spatial masking. IDSQ exploits the decreased perceptual sensitivity of the HVS to quantisation-induced compression artefacts in areas within YCbCr video data that contain high and low luma sample intensities. Y. Zhang et al. in [28] expand on Naccari's and Mrak's IDSQ technique by applying it to High Dynamic Range (HDR)-related tone-mapping applications.

As is evident in the overwhelming vast majority of JND contributions that have been previously proposed, the JND of chrominance data is typically neglected. Several JND methods reviewed in the previous paragraph share one or more of the same features including luminance masking, luminance-based contrast masking, luminance-based temporal masking and luminance-based spatial CSF. As such, if the corresponding JND techniques were to be applied to contemporary video coding applications, the JND threshold for chrominance data would be treated as identical to the JND threshold for luminance data. This is a major drawback because chrominance data is considerably different from luminance data; therefore, this leaves room for improvement. It is important and desirable to develop a comprehensive JND method that accounts for both luminance and chrominance data.

In addition to the absence of accounting for chrominance JND, other issues exist that are not considered in contemporary JND techniques. For example, the method proposed by Yang et al. and also the technique proposed by Chen and Guillemot both employ the luminance masking and contrast masking functions derived by Chou's and Chen's techniques in [19,20]. The issue here is as follows: the psychophysical experiments undertaken by Chou and Chen were conducted in 1995–96 on obsolete visual display technologies (i.e., an SD and low resolution 19 inch CRT monitor). Therefore, Chou's and Chen's corresponding luminance masking and contrast masking functions may require revisions. This is because the derived JND visibility thresholds may prove to be significantly different if the corresponding subjective evaluations were to be performed on contemporary visual display technologies (e.g., a state-of-the-art TV or monitor which supports HD, Ultra HD, HDR, WCG and 4:4:4 video data).

Another potential issue with previously proposed JND methods – with the exception of Y. Zhang's HDR-related tone-mapping extension [28] of Naccari's and Mrak's JND-based IDSQ technique – is the fact that they are designed for raw 24-bit YCbCr data (i.e., 8-bits per channel data). This equates to the fact that most of the aforementioned empirical parameters in the luminance masking, contrast masking and temporal masking functions are designed to work with 8-bit precision data only. This may prove to be a significant issue because high bit depth data