Metadata Reduction for Soft Video Delivery

Ticao Zhang and Shiwen Mao, *Fellow, IEEE*

***Abstract*—Soft video delivery allows a graceful degradation of video quality matching to user's channel quality. However, existing schemes require a considerable amount of metadata to be transmitted. In this paper, we propose a blind data detection method that recovers video signals from the squared amplitude of received signals, which is almost metadata free. Simulation results show that our proposed method can significantly reduce the metadata overhead, and outperforms the existing soft video delivery scheme by nearly 2dB. It does not require a time-consuming process to fit signal energy distributions (SED) and video quality can improve gracefully.**

***Index Terms*—Soft video delivery; Softcast; Metadata reduction; Wireless video transmission.**

I. Introduction

Wireless video deliveries nowadays are mostly built on *Shannon's theorem* that separates channel coding and source coding to ensure a reliable transmission. In conventional digital video transmissions (e.g., H.264 [1]), a video is first encoded into bit streams. Then channel coding is performed to reliably transmit the bit streams. However, this framework has several drawbacks. First, the quantization process involved is a lossy process which cannot be recovered at the receiver. Thus, even when the channel condition gets better, the video quality may not be able to be improved. This scheme also suffers the *cliff-effect* [2], which refers to the phenomenon that the video quality drops dramatically due to bit errors incurred in bad channel conditions. Most wireless systems adopt forward error correction (FEC) coding to correct a certain number of bit errors. However, when the number of errors exceeds the correction capability, FEC will fail to correct the errors. The errors will even spread to the correctly demodulated codes due to the nature of the convolutional code.

Joint source and channel coding has the potential to tackle this problem, as demonstrated by a break-through system named SoftCast [3], [4]. In this soft video delivery framework, the process of video compression, data protection, and transmission are integrated. The luminance value of the video content is first processed by a 3-dimensional discrete cosine transform (3D-DCT) to de-correlate the signal. The power allocation process then reduces the end-to-end distortion by optimally scaling the DCT coefficients with a *power scaling factor*, which is encoded as metadata. To combat packet losses, a whitening process is applied by multiplying a Hadamard matrix to ensure that each packet is of equal importance. Finally, The numerical, scaled values are mapped to constellation points of modulation and transmitted. At the decoder, a linear least Square Error (LLSE) estimator is used to decode the signal. Since all the operations involved are linear, this scheme has the potential to overcome the cliff effect in conventional digital solutions. Users can hereafter gracefully improve their video quality matching to their wireless channel conditions.

In conventional soft video delivery (SoftCast), DCT coefficients are divided into chunks and each chunk uses the same scaling factor according to the mean power of the chunk. In [5], it is shown that the end-to-end distortion strongly depends on the chunk size. If the chunk size is large, the distortion would also be large. If the chunk size is small, the metadata overhead would be large. According to the experiments in [3], if each picture is divided into $8 \times 8 = 64$ chunks, the metadata overhead is 0.005bits/pixel, which is acceptable. However, the PSNR gap from the optimality is still quite large under this chunk size. *How to reduce the resource occupied by transmitting the metadata, while also achieving a satisfactory received video quality, is a big challenge.*

To combat the metadata overhead, the authors in [5], [6] adopt an adaptive, L-shaped chunk division method to more accurately model the signal energy distribution (SED) via a simple piece-wise linear function. Simulations show that the proposed scheme can achieve 95% transform gain while reducing the overhead of metadata from hundreds or thousands to only a few parameters. In [7], the authors apply a Gaussian Markov Random Field (GMRF) model with very few parameters to approximate the SED. This way, the received video quality can be improved by 1.2dB with 99.7% reduction in metadata overhead. To our best knowledge, most existing works share a common idea that the transmitter allocates power according to the fitted SED and the receiver reconstructs the SED with fitting parameters. In practice, the fitting process is time-consuming, especially for real time hi-definition (HD) videos. Further, the existing approach suffers from the distortion caused by the mismatch between the empirical SED plane and the fitted SED plane. Once fitted, the distortion cannot be eliminated at the receiver.

In this paper, we proposed a blind data detection algorithm. Different from the existing works, in our design, the transmitter optimally scales each DCT coefficient, so that the distortion caused by improper power scaling can be eliminated. At the receiver side, the video signals are recovered using the squared amplitude of the received signals. Since we no longer need to transmit any information about the power scaling factors, this scheme is simple and almost metadata-free. This paper is organized as follows. In Section II, we introduce the basic framework of soft video delivery with some fundamental analysis. The proposed metadata overhead reduction scheme is

Manuscript received Mar. 17, 2019; accepted Apr. 19, 2019. This work is supported in part by the NSF under Grant CNS-1702957, and the Wireless Engineering Research and Education Center (WEREC) at Auburn University. The associate editor coordinating the review of this paper and approving it for publication was K. Almeroth. (Corresponding author: Shiwen Mao.)

T. Zhang and S. Mao are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA. Email: tzz0031@tigermail.auburn.edu, smao@ieee.org.

Digital Object Identifier 10.1109/LNET.2019.2912831.

presented in Section III and validated in Section IV. Section V reviews related work and Section VI concludes this paper.

II. System Model

A. Encoder and Decoder

The monochrome version of a video sequence is first partitioned into groups of pictures (GOPs), each consisting of F pictures. Let $\mathbf{X} \in \mathbb{R}^{H\times W\times F}$ denote one GOP, where H and W represents the resolution in height and width of each picture, respectively. We then perform a 3D-DCT on $\mathbf{X}$ and obtain the corresponding DCT coefficients matrix $\mathbf{S}$ for each picture of the same size. Projecting the video sequences to an orthogonal basis, we de-correlate the video signals and compact energy in the low frequency components. Matrix $\mathbf{S}$ is then divided into chunks. This is different from the MPEG compression where a picture is first divided into blocks, and then a 2D-DCT is performed on each block.

Suppose each picture is divided into m^2 chunks and there will be $N = Fm^2$ chunks in one GoP. Moreover, each chunk is a rectangle with height $h = H/m$ and width $w = W/m$. Denote the j-th DCT coefficient in the i-th chunk as $s_i[j]$, $i = 1, 2, ..., N$, we scale $s_i[j]$ with a factor g_i for noise reduction. The value of g_i is encoded as metadata and transmitted to the receiver digitally via entropy coding.

Assume an additive white Gaussian noise (AWGN) channel. After demodulation, the receiver receives $y_i[j] = g_i s_i[j] + n_i[j]$, where $n_i[j]$ is AWGN with variance σ_n^2. This process can be written in a matrix form as $\mathbf{Y} = \mathbf{GS} + \mathbf{N}$, where $\mathbf{Y}$ is the matrix of received signal, $\mathbf{G}$ is a diagonal matrix of power scaling factors, $\mathbf{S}$ is the stacked DCT coefficients matrix, and $\mathbf{N}$ is the AWGN noise matrix. The receiver aims to decode the signal and reconstruct the DCT coefficients via an optimal *linear least square estimator* (LLSE) [8], as

$$\hat{\mathbf{S}} = \mathbf{\Lambda}_s \mathbf{G}^T (\mathbf{G}\mathbf{\Lambda_s}\mathbf{G}^T + \mathbf{\Sigma})^{-1}\mathbf{Y}, \tag{1}$$

where $\mathbf{\Lambda}_s$ is a diagonal matrix whose element λ_i is the variance of the i-th chunk, and $\mathbf{\Sigma}$ is a diagonal matrix whose diagonal elements are the variance of AWGN entries.

We can also rewrite (1) in a scalar form as

$$\hat{s}_i[j] = \frac{g_i\lambda_i}{g_i^2\lambda_i + \sigma_n^2} \cdot (g_i s_i[j] + n_i[j]). \tag{2}$$

When the average power of the chunk is much greater than that of the noise, i.e., $p_i = g_i^2\lambda_i \gg \sigma_n^2$, the LLSE estimator reduces to a *zero-forcing* (ZF) estimator, as

$$\begin{aligned}\hat{s}_i[j] &\approx y_i[j]/g_i \\ &= s_i[j] + n_i[j]/g_i.\end{aligned} \tag{3}$$

B. Distortion Analysis

The expected *mean square error* (MSE) is

$$\begin{aligned}\mathrm{MSE}_s &= \mathbb{E}\left[(s_i[j] - \hat{s}_i[j])^2\right] \\ &= \sum_{i=1}^{N} \frac{\sigma_n^2\lambda_i}{g_i^2\lambda_i + \sigma_n^2} \\ &\approx \frac{\sigma_n^2}{N}\sum_{i=1}^{N}\frac{1}{g_i^2}.\end{aligned} \tag{4}$$

When the power scaling factor g_i is large enough, the distortion will be minimized. However, in practice, we cannot set g_i too large since the transmitted power is usually constrained by the total power budget P for one GOP.

The optimal g_i in matrix $\mathbf{G}$ and the power of the scaled coefficients can be derived as [3]

$$g_i = c\lambda_i^{-\frac{1}{4}}, \tag{5}$$

where c is a constant so that the total power budget constraint is satisfied. Since $hw\sum_{i=1}^{N} p_i = P$, we have

$$c = \sqrt{\frac{P}{hw}\cdot\frac{1}{\sum_{i=1}^{N}\lambda_i^{\frac{1}{2}}}}, \tag{6}$$

From (5), we can see that the resulting power of the scaled coefficient is proportional to the squared root of the chunk variance. Substituting (5) into (4), the distortion becomes

$$\mathrm{MSE}_s \approx \frac{\sigma_n^2 hw}{NP}\left(\sum_{i=1}^{N}\lambda_i^{\frac{1}{2}}\right)^2. \tag{7}$$

For one GOP, the total variance of noise is $Nhw\sigma_n^2$. Then the *signal-to-noise ratio* (SNR) can be defined as $\rho = \frac{P}{Nhw\sigma_n^2}$. We can rewrite the distortion of DCT coefficients as

$$\begin{aligned}\mathrm{MSE}_s &\approx \frac{1}{\rho}\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^{\frac{1}{2}}\right)^2 \\ &= \frac{1}{\rho}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[s_i[j]^2]^{\frac{1}{2}}\right)^2 \\ &\geq \frac{1}{\rho}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[\left|s_i[j]\right|\middle| j\in\mathcal{C}_i\right]\right)^2 \\ &= \frac{1}{\rho}\left(\mathrm{E}\left[\left|s_i\right|\right]\right)^2.\end{aligned} \tag{8}$$

The inequality in (8) is because the quadratic mean is no smaller than the arithmetic mean. The equality holds when the absolute value of the coefficients in one chunk are all identical. This does not hold for a natural image/video unless we force each chunk to contain only one coefficient. In this case, the chunk size is one and we have the minimum distortion. However, the number of g_i, which is equal to the number of chunks, would also be large, introducing a considerable transmission overhead.

III. Overhead Reduction

To carry out the MMSE filtering in (2), the sender needs to correctly notify the receiver the value of λ_i of all coefficients, i.e., the metadata. For example, to transmit a video sequence with a resolution of 352×288, there will be $352 \times 288 = 811,008$ DCT coefficients. In this case, the amount of metadata will be approximately 5.8 bits/pixel. This overhead will cause performance degradation, since it consumes extra transmit power during the transmission of analog-modulated symbols. This is why the conventional soft video delivery scheme [3] partitions DCT coefficients into chunks and carries out scaling and MMSE filtering for each chunk.

However, the overhead is still high in general and chunk partition is not optimized, which would incur a performance degradation. Although we can lower the chunk number to reduce the overhead of metadata, the video quality would degrade significantly (as will be shown later). Moreover, the accuracy of metadata plays a vital role in successful decoding of DCT coefficients. To ensure a reliable transmission, more error correction codes have to be added and this even further increases the transmission overhead.

The existing SED modeling based signal detection algorithm leverages the information of an approximated g_i. In contrast, our proposed algorithm decodes the signal without prior knowledge of g_i. Thus we call it *blind data detection*. Note that the SED modeling based method scales each DCT coefficients via a learned and also mismatched modeling function, which may also introduce distortion. In our proposed scheme, we scale each coefficient with an *optimal scaling factor* at the encoder first. This way, each chunk only contains one coefficient. Letting $c = \sqrt{P/\sum_i^N |s_i|}$, we have $\lambda_i = \mathrm{E}[s_i[j]^2] = s_i^2$ and (5) can be reduced to

$$g_i = c\lambda_i^{-\frac{1}{4}} = c|s_i|^{-\frac{1}{2}},\ i = 1, 2, ..., N. \tag{9}$$

The general model can be considered as

$$y_i = c|s_i|^{-\frac{1}{2}} s_i + n_i,\ i = 1, 2, ..., N. \tag{10}$$

This is a set of independent scalar nonlinear estimation problems. We can estimate the amplitude of s_i via a ZF estimator

$$|\hat{s}_i| = (y_i/c)^2, \tag{11}$$

and use the sign of the received signal to approximate that of s_i, i.e., $\mathrm{sign}(\hat{s}_i) \approx \mathrm{sign}(y_i)$. We obtain an estimation of s_i as

$$\begin{aligned} \hat{s}_i &= |\hat{s}_i| \cdot \mathrm{sign}(\hat{s}_i) \\ &\approx (y_i/c)^2 \cdot \mathrm{sign}(y_i). \end{aligned} \tag{12}$$

Eqn. (12) shows that the amplitude of the DCT signal is proportional to the squared amplitude of the received signal. For each DCT coefficient in one GoP, the decoder only needs to know the value of constant c. In practical communication systems, this scalar value along with the total power P can be transmitted to the decoder at a negligible cost per GoP. That is, this proposed signal recovery method is almost *metadata-free* and works with nearly no overhead.

The expected theoretical MSE of the proposed blind detection algorithm is

$$\mathrm{MSE}_s = \left(4/\rho + 3/\rho^2\right) \left(\mathbb{E}\left[|s_i|\right]\right)^2. \tag{13}$$

where ρ is the SNR and $\mathrm{E}[|s_i|]$ is the average of the absolute value of the DCT coefficients in one GOP. The proof is provided in Appendix A. It can be seen that the MSE depends on both the channel condition (i.e., ρ) and the video content (i.e., s_i). We can jointly consider the channel characteristics and video content in a cross-layer design.

IV. Performance Evaluation

A. Simulation Setup

1) Performance Metric: We evaluate the video quality in terms of *peak signal-to-noise ratio* (PSNR) and Structural Similarity Index (SSIM). PSNR is a widely used metric for video delivery, defined (in dB) as

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{(2^L-1)^2}{\mathrm{MSE}_X}\right) \tag{14a}$$

$$\mathrm{MSE}_X = \mathbb{E}\left[(X_{ij} - \hat{X}_{ij})^2\right], \tag{14b}$$

where L is the number of bits used to encode pixel luminance (usually L = 8 bits), and X_{ij} and X'_{ij} are the transmitted and reconstructed luminance pixels of a picture, respectively. In Appendix B, we prove that the MSE in the pixel domain is equal to that defined in the frequency domain, namely,

$$\mathrm{MSE}_X = \mathrm{MSE}_s. \tag{15}$$

By substituting (8) and (15) into (14), we obtain the theoretical PSNR of the soft video delivery scheme. Generally, improvements of PSNR of magnitude larger than 0.5dB are noticeable visually, and a PSNR below 20dB is considered not acceptable. SSIM is a perceptual metric that quantifies image quality degradation: a value closer to 1 indicates higher perceptual similarity between the original and decoded images.

2) Test Video: We use monochrome video sequences at a frame rate of 20 frames per second (fps) in the tests. The standard video sequences used include *akiyo*, *mother-daughter*, *foreman*, and *highway* in the CIF format with a resolution of 352×288 [9]. The sample format is YUV420. The standard test picture *Lena* is also used in our simulation.

3) Video Encoder Setting: We set the GoP size for all reference schemes to four. In the conventional chunk division based soft video delivery scheme, each picture is divided into 16×16, 8×8, 4×4, and 2×2 chunks (of increasing chunk size). For digital schemes, we use the HEVC software for video encoder and decoder, where the *encoder_lowdelay_main.cfg* configuration file is used to operate the encoder with both intra encoding and motion compensation.

4) Wireless Setting: The received symbols are delivered through an AWGN channel. Since the video sequence used here is in a relatively low resolution (i.e., in the CIF format), the symbol rate is set to 25K. Then we adopt convolution channel coding for error correction with a rate 1/2 and constraint length 7. The digital modulation formats are either BPSK, QPSK, or 16QAM. For a fair comparision, we use the same average power for both digital transmission and soft video transmission (as verified in the results).

B. Test on Images

Figure 1 presents the PSNR performance under the AWGN channel for delivering the test picture *Lena* with a resolution of 512×512. Essentially, by optimally scaling the DCT coefficients and letting the receiver know the exact value of the power scaling factors, we can obtain the "optimal performance" as shown in the figure.

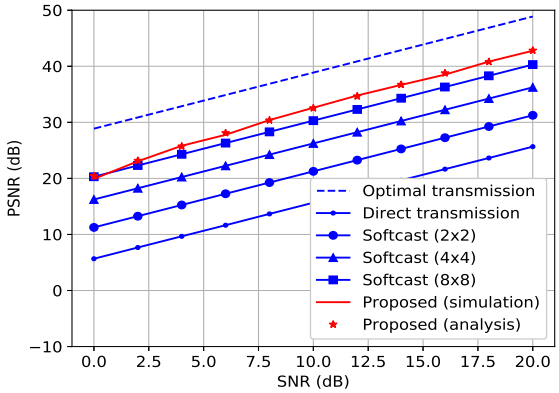

Fig. 1. PSNRs achieved by the proposed and baseline schemes for *Lena* (512×512, gray).

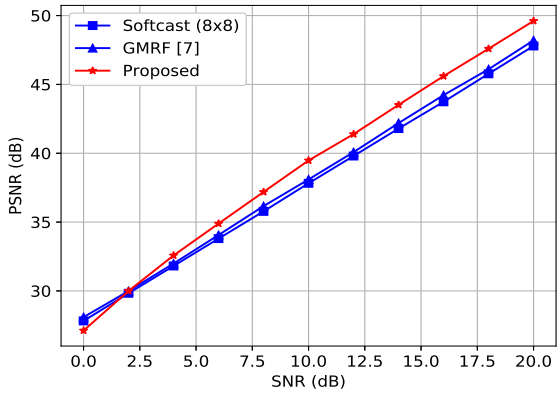

Fig. 2. PSNR performance comparison for video sequence *akiyo*.

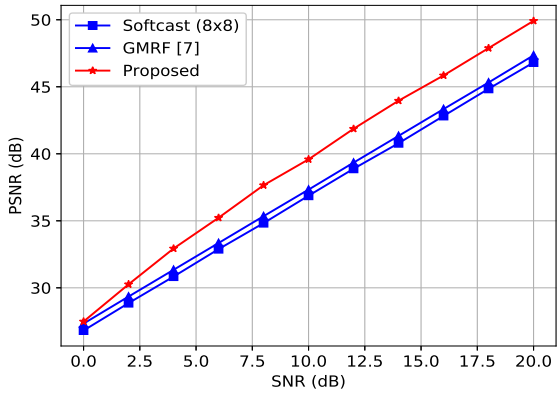

Fig. 3. PSNR performance comparison for video sequence *mother-daughter*.

As a comparison, when the receiver knows nothing about the value of power scaling factors, or in some circumstances the packet which contains the value of g_i is lost, the receiver simply decodes the signals by treating all g_i as a constant. The results in this case is termed "direct transmission" in Fig. 1. As can be seen, there exists a nearly 20dB gap between the two, hence it is of great importance for the receiver to have correct information about the power scaling factors. The conventional chunk based Softcast scheme divides each picture into chunks. With the increase of the number of chunks, the PSNR performance will continue to improve, eventually converging to the optimal performance. However, this comes at a cost of an increased metadata overhead. Compared with a 8×8 chunk division scheme, where most of the existing works assume, our proposed blind data detection method achieves a nearly 2dB PSNR improvement with almost no metadata overhead. Moreover, the theoretical results obtained via (13) match well with our simulations, which can be used for system optimization in our future work.

C. Overhead Reduction for Video Delivery

Figures 2 and 3 shows the PSNR performance of delivering video *akiyo.yuv* and *mother-daughter.yuv*, respectively. It can be seen that the GMRF method [7] has a similar performance as conventional Softcast, which divides each picture into $8 \times 8 = 64$ chunks. The proposed blind detection method generally achieves a 2dB gain in PSNR, especially in the high SNR regime. This is because for blind data detection, we recover the coefficients from the amplitude of received signals. In the high SNR regime, the impact of noise is generally lower. Hence we can recover the coefficients more accurately.

In the simulation, for both GMRF and the conventional Softcast method, we assume that the metadata can be transmitted accurately with additional power and channel resources. The metadata overhead for the conventional chunk based Softcast (8×8) and GMRF method is 256 and 5 symbols (or variables) per chunk, respectively, when the GOP size is 4. While for the proposed method, the metadata overhead is almost zero.

Figures 5 and 6 compare the visual quality of the proposed method and the chunk-based Softcast scheme for video sequences *foreman* and *highway*, respectively. The video sequence is transmitted at an SNR of 10dB. For *foreman*, the SSIM achieved by Softcast that divides each picture into 16×16, 8×8, 4×4, and 2×2 chunks, are 0.92196, 0.87724, 0.73161, and 0.46255, respectively, whereas the SSIM achieved by the proposed scheme is 0.87745. From the experiment, we can see that a finer chunk division generally leads to an improved visual quality. However, this also comes at a price of increased metadata overhead. Compared with conventional chunk based Softcast, the proposed scheme generally achieves a 2dB PSNR improvement while being almost metadata-free. The reason is that the proposed scheme optimally scales the DCT coefficient at the transmitter while the conventional Softcast suffers an extra distortion due to the improper SED modeling with rectangular chunks.

D. Impact of GOP Size

In Fig. 4, we examine the impact of GOP size under different SNR settings. The test video sequence is *mother-daughter.yuv*. It can be seen that with the increase of GOP size, the PSNR value also tends to increase. This performance improvement is brought by the inter-frame correlation. Within a certain range, a slight increase of GOP size may help remove the temporal redundancy more effectively, and hence a better PSNR can be achieved. However, an extremely large GOP will bring a huge computational overhead. Usually, we set GOP to

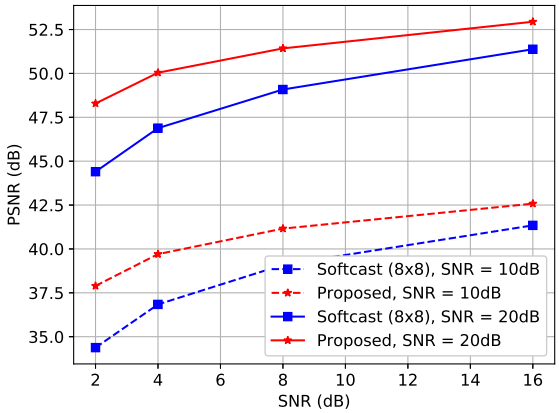

Fig. 4. The impact of GOP size on chunk-based Softcast and the proposed scheme.

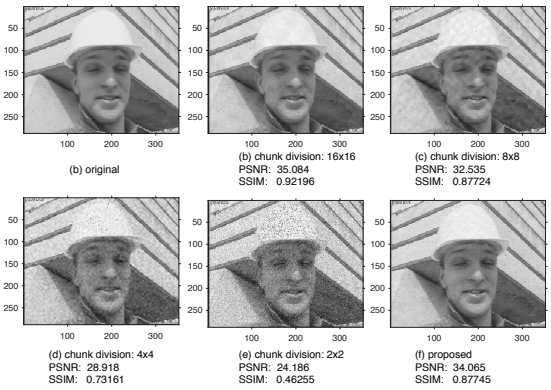

Fig. 5. Snapshot of video sequence *foreman* (frame #1) delivered by each scheme at an SNR of 10 dB.

be 4 to 20. In our simulation, we set the GOP size to 4. As shown in [10], performance optimization through tuning the GoP size is unnecessary. Finally, we observe that regardless of the GOP size, the proposed method always outperforms the chunk-based Softcast scheme.

E. Comparison with Digital Solutions

In Fig. 7, we compare the PSNR performance of the proposed algorithm with that of digital benchmarks. The figure confirms the characteristics of cliff effect of current digital video wireless transmission systems. To be specific, there exists a critical SNR below which the video quality drops sharply; and conversely, above the critical SNR, the video quality does not improve with increased SNR. The maximum video quality is limited by the lossy encoding process at the transmitter. The video quality will not improve anymore even if the channel quality keeps improving. Moreover, when the channel quality changes dynamically, complex rate-distortion control has to be performed to guarantee a satisfactory performance. On the contrary, the proposed soft video delivery schemes improves the video quality gracefully according to the channel quality. Note that, in digital video transmission, coding rate corresponds to video quality and video size. In general, a higher coding rate will ensure a higher video quality. By adjusting the coding rate, the HEVC transmission scheme may perform better or worse than the soft video transmission scheme, where a source does not need to select a coding rate. In this experiment, we show a general trend. for a more fair comparison between soft video delivery and digital video transmission, we refer readers to [3], [4].

V. Related Research

Wireless video transmission has been an important problem attracting considerable interests [11]–[16]. Our study is closely related to the existing works on soft video delivery. Since the pioneering work first proposed in [3], [4], there have been a variety of follow-ups. Unlike SoftCast, which builds an analog code that adjusts the compression-protection tradeoff with power allocations, FlexCast [17] achieves a similar goal with bit allocation. DCast [18] aims to remove the inter frame redundancy by incorporating motion prediction and compensation. LayerCast [19] considers a broadcast framework that can simultaneously accommodate heterogeneous users with diverse SNRs and diverse bandwidths. A hybrid analog and digital coding scheme is proposed in [20] to leverage the advantages of digital and analog coding. In [21], a design is proposed to perform optimal channel and power allocation for a fast-fading channel.

ParCast [22] is the first work to extend SoftCast to a multiple antenna system. It decomposes the MIMO channel into parallel sub-channels by MIMO precoding. By assigning high priority DCT coefficients to higher quality sub channels, the reconstructed video quality can be optimized. However, the proposed framework only works for a single receiver. In [23], a modeling method is proposed for joint power and bandwidth allocation in soft video delivery, while the work in [24] extends Softcast to a wireless video multicast scenario with receiver antenna heterogeneity. The designed multicast system allows receivers to achieve a reconstructed video quality that improves with the number of equipped antennas.

VI. Conclusion

In this paper, we considered the metadata reduction problem in soft video delivery. We proposed a blind data detection algorithm that recovers the signal from squared amplitude of the received signals. This method can prevent the cliff-effect in conventional digital video coding. In terms of PSNR performance, it outperformed the conventional Softcast by approximately 2dB and is almost metadata-free. Although the results were obtained for AWGN channels, the proposed scheme can be easily adapted for general fading channels.

Appendix A
Proof of Equation (13)

Suppose that in this linear AWGN system, the sign information of the true DCT signal $\mathrm{sign}(\hat{s}_i)$ can be well approximated by that of the received signal, i.e.,

$$\mathrm{sign}(\hat{s}_i) \approx \mathrm{sign}(y_i). \tag{16}$$

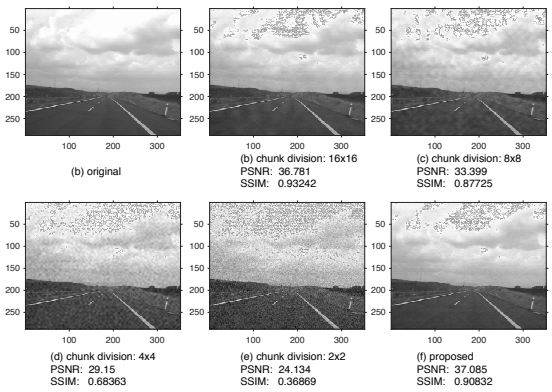

Fig. 6. Snapshot of video sequence *highway* (frame #1) delivered by each scheme at an SNR of 10 dB.

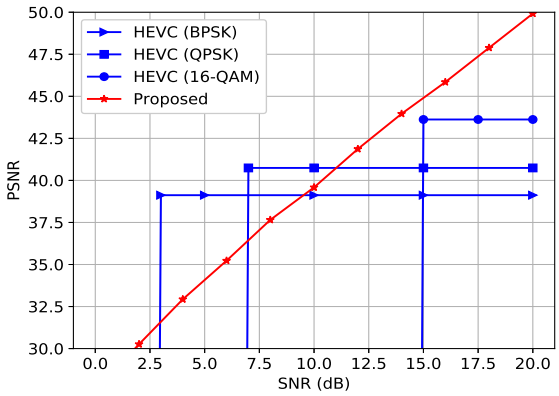

Fig. 7. PSNR performance comparison with digital solutions for video sequence *mother-daughter*.

Then (12) can be rewritten as

$$\begin{aligned}\hat{s}_i &= |\hat{s}_i| \cdot \mathrm{sign}(\hat{s}_i) \\ &\approx \left(\frac{y_i}{c}\right)^2 \cdot \mathrm{sign}(y_i) \\ &\approx \frac{1}{c^2}(g_i s_i + n_i)^2 \cdot \mathrm{sign}(s_i) \\ &= \frac{1}{c^2}(c|s_i|^{-\frac{1}{2}} s_i + n_i)^2 \cdot \mathrm{sign}(s_i) \\ &= s_i + \left(\frac{2}{c}|s_i|^{-\frac{1}{2}} n_i + \frac{n_i^2}{c^2}\right) \cdot \mathrm{sign}(s_i). \end{aligned} \tag{17}$$

Therefore, the MSE of a GOP can be derived as

$$\begin{aligned}\mathrm{MSE}_s &= \mathbb{E}[(s_i - \hat{s}_i)^2] \\ &= \mathbb{E}\left[\left(\frac{2}{c}|s_i|^{-\frac{1}{2}} n_i + \frac{n_i^2}{c^4}\right)^2\right] \qquad (18) \\ &= \frac{4\sigma_n^2}{c^2 \mathbb{E}[|s_i|]} + \frac{3\sigma_n^2}{c^4}, \qquad (19)\end{aligned}$$

where the expectation involved above is taken w.r.t. all the DCT coefficients in the GOP.

From (18) to (19), we have assumed that the transmitted signal is uncorrelated to the channel noise, which is quite general in practice. Furthermore, in calculating the high order moments of Gaussian random variable n_i, we used the property that if a normal random variable $x \sim \mathcal{N}(0, \sigma^2)$, then we have [25, p. 148]

$$\mathbb{E}[x^n] = \begin{cases} 0, & \text{if } n \text{ is odd} \\ 1 \cdot 3 \cdots (n-1)\sigma^n, & \text{if } n \text{ is even.} \end{cases}$$

When the chunk size is 1, we have

$$\begin{aligned} c &= \sqrt{\frac{P_s}{\sum_i |s_i|}} \\ &= \sqrt{\frac{P_s}{N\mathbb{E}[|s_i|]}}, \end{aligned} \tag{20}$$

and $\rho = \frac{P}{N\sigma_n^2}$. Then (19) becomes

$$\mathrm{MSE}_s = \left(\frac{4}{\rho} + \frac{3}{\rho^2}\right) \cdot \mathbb{E}[|s_i|]^2. \tag{21}$$

Appendix B
Proof of Equation (15)

First of all, we consider the simplest one-dimensional case. Suppose $\mathbf{T} \in \mathbb{R}^{N \times N}$ is a DCT matrix, by multiplying the DCT matrix to a vector $\mathbf{b} \in \mathbb{R}^{N \times 1}$, we obtain the corresponding DCT coefficients $\mathbf{a} = \mathbf{T}\mathbf{b} \in \mathbb{R}^{N \times 1}$. Let $\hat{\mathbf{a}}$ be a noisy estimate of $\mathbf{a}$ and $\hat{\mathbf{b}} = \mathbf{T}^T\hat{\mathbf{a}}$ be the inverse DCT transform of $\hat{\mathbf{a}}$. Then we have

$$\begin{aligned}\mathrm{MSE}_a &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(a_i - \hat{a}_i)^2\right] \\ &= \frac{1}{N}\mathbb{E}\left[(\mathbf{a} - \hat{\mathbf{a}})^T(\mathbf{a} - \hat{\mathbf{a}})\right] \\ &= \frac{1}{N}\mathbb{E}\left[(\mathbf{T}\mathbf{b} - \mathbf{T}\hat{\mathbf{b}})^T(\mathbf{T}\mathbf{b} - \mathbf{T}\hat{\mathbf{b}})\right] \\ &= \frac{1}{N}\mathbb{E}\left[(\mathbf{b} - \hat{\mathbf{b}})^T\mathbf{T}^T\mathbf{T}(\mathbf{b} - \hat{\mathbf{b}})\right] \\ &= \frac{1}{N}\mathbb{E}\left[(\mathbf{b} - \hat{\mathbf{b}})^T(\mathbf{b} - \hat{\mathbf{b}})\right] \\ &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(b_i - \hat{b}_i)^2\right] \\ &= \mathrm{MSE}_b. \end{aligned} \tag{22}$$

In the above derivation, we have applied the fact that the DCT is an orthogonal transform and thus $\mathbf{T}^T\mathbf{T} = \mathbf{I}$. Therefore, the MSE does not change after 1D-DCT.

Similarly, a 2D-DCT can be decomposed into a succession of two 1D-DCTs, while a 3D-DCT can be decomposed into a succession of a 2D-DCT and a 1D-DCT [26]. For soft image and video delivery, the orthogonality nature ensures that the MSE does not change after the 3D-DCT transformation. This property provides us an easy way to measure the PSNR qualitatively.

Note that, our results only show that the average MSE for a GOP does not change after the 3D-DCT transformation. However, the MSE for different pictures in a GOP may still be different. Despite that, we can always use the GOP's average MSE_s to approximate the MSE_x for each picture in the GOP, because usually the PSNR for consecutive pictures in a GOP does not fluctuate dramatically.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, July 2003.

[2] S. Kokalj-Filipovic, E. Soljanin, and Y. Gao, "Cliff effect suppression through multiple-descriptions with split personality," in *Proc. IEEE ISIT 2011*, St. Petersburg, Russia, Aug. 2011, pp. 948–952.

[3] S. Jakubczak and D. Katabi, "SoftCast: One-size-fits-all wireless video," *ACM SIGCOMM Comput. Commu. Rev.*, vol. 40, no. 4, pp. 449–450, Oct. 2010.

[4] ——, "A cross-layer design for scalable mobile video," in *Proc. ACM MobiCom'11*, Las Vegas, NV, Sept. 2011, pp. 289–300.

[5] H. Cui, Z. Song, Z. Yang, C. Luo, R. Xiong, and F. Wu, "Cactus: A hybrid digital-analog wireless video communication system," in *Proc. ACM MSWiM'13,*, Barcelona, Spain, Nov. 2013, pp. 273–278.

[6] R. Xiong, J. Zhang, F. Wu, J. Xu, and W. Gao, "Power distortion optimization for uncoded linear transformed transmission of images and videos," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 222–236, Jan. 2017.

[7] T. Fujihashi, T. Koike-Akino, T. Watanabe, and P. V. Orlik, "High-quality soft video delivery with GMRF-based overhead reduction," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 473–483, Feb. 2018.

[8] K.-H. Lee and D. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun*, vol. 24, no. 12, pp. 1283–1290, Dec. 1976.

[9] "Xiph.org video test media [derf's collection]." [Online]. Available: https://media.xiph.org/video/derf/

[10] D. Yang, Y. Bi, Z. Si, Z. He, and K. Niu, "Performance evaluation and parameter optimization of SoftCast wireless video broadcast," in *Proc. EAI MobiMedia'15*, Chengdu, China, May 2015, pp. 79–84.

[11] S. Mao, S. Lin, S. S. Panwar, Y. Wang, and E. Celebi, "Video transport over ad hoc networks: Multistream coding with multipath transport," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1721–1737, Dec. 2003.

[12] M. Chen, V. C. M. Leung, S. Mao, and M. Li, "Cross-layer and priority path scheduling for real-time video communications over wireless sensor networks," in *Proc. IEEE VTC 2008-Spring*, Marina Bay, Singapore, May 2008, pp. 2873–2877.

[13] Y. Xu and S. Mao, "A survey of mobile cloud computing for rich media applications," *IEEE Wireless Communications Magazine*, vol. 20, no. 3, pp. 46–53, June 2013.

[14] Z. He, S. Mao, and T. Jiang, "A survey of QoE driven video streaming over cognitive radio networks," *IEEE Network Magazine*, vol. 29, no. 6, pp. 20–25, Nov./Dec. 2015.

[15] M. Amjad, M. Rehmani, and S. Mao, "Wireless multimedia cognitive radio networks: A comprehensive survey," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 2, pp. 1056–1103, SEcond Quarter 2018.

[16] Z. Wang, S. Mao, L. Yang, and P. Tang, "A survey of multimedia big data," *IEEE/CIC China Communications*, vol. 15, no. 1, pp. 155–176, Jan. 2018.

[17] S. Aditya and S. Katti, "FlexCast: Graceful wireless video streaming," in *Proc. ACM MobiCom'11*, Sept., Las Vegas, Nevada 2011, pp. 277–288.

[18] X. Fan, F. Wu, D. Zhao, O. C. Au, and W. Gao, "Distributed soft video broadcast (DCAST) with explicit motion," in *Proc. IEEE Data Compression Conf. 2012*, Snowbird, UT, Apr. 2012, pp. 199–208.

[19] X. Fan, R. Xiong, D. Zhao, and F. Wu, "Layered soft video broadcast for heterogeneous receivers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1801–1814, Nov. 2015.

[20] L. Yu, H. Li, and W. Li, "Wireless scalable video coding using a hybrid digital-analog scheme," *IEEE Trans. Circuits Syst. Video Technol*, vol. 24, no. 2, pp. 331–345, 2014.

[21] H. Cui, C. Luo, C. W. Chen, and F. Wu, "Robust uncoded video transmission over wireless fast fading channel," in *Proc. IEEE INFOCOM'14*, Toronto, Canada, Apr./May 2014, pp. 73–81.

[22] X. L. Liu, W. Hu, C. Luo, Q. Pu, F. Wu, and Y. Zhang, "ParCast+: Parallel video unicast in MIMO-OFDM WLANs," *IEEE Trans. Multimedia*, vol. 16, no. 7, pp. 2038–2051, Nov. 2014.

[23] D. Liu, J. Wu, H. Cui, D. Zhang, C. Luo, and F. Wu, "Cost-distortion optimization and resource control in pseudo-analog visual communications," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3097–3110, Nov. 2018.

[24] H. Cui, C. Luo, C. W. Chen, and F. Wu, "Scalable video multicast for MU-MIMO systems with antenna heterogeneity," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 26, no. 5, pp. 992–1003, May 2016.

[25] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York, NY: Tata McGraw-Hill Education, 2002.

[26] X. Li, A. Dick, C. Shen, A. Van Den Hengel, and H. Wang, "Incremental learning of 3D-DCT compact representations for robust visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 35, no. 4, pp. 863–881, 2013.

Ticao Zhang received the B.E. degree in 2014 and the M.S. degree in 2017 from School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. He is currently pursuing a Ph.D. degree in Electrical and Computer Engineering at Auburn university. His research interests include video coding and communications, and optimization and design of wireless multimedia networks.

Shiwen Mao [S'99-M'04-SM'09-F'19] received his Ph.D. in electrical and computer engineering from Polytechnic University, Brooklyn, NY (now New York University Tandon School of Engineering). Currently, he is the Samuel Ginn Professor in the Department of Electrical and Computer Engineering, and Director of the Wireless Engineering Research and Education Center (WEREC) at Auburn University, Auburn, AL.

His research interests include wireless networks, multimedia communications, and smart grid. He is a Distinguished Speaker of the IEEE Vehicular Technology Society. He is on the Editorial Board of IEEE Transactions on Network Science and Engineering, IEEE Transactions on Mobile Computing, IEEE Transactions on Multimedia, IEEE Internet of Things Journal, IEEE Multimedia, IEEE Networking Letters, and ACM GetMobile, among others. He received the Auburn University Creative Research & Scholarship Award in 2018 and NSF CAREER Award in 2010, as well as the 2017 IEEE ComSoc ITC Outstanding Service Award, the 2015 IEEE ComSoc TC-CSR Distinguished Service Award, and the 2013 IEEE ComSoc MMTC Outstanding Leadership Award. He is a co-recipient of the IEEE ComSoc MMTC Best Conference Paper Award in 2018, the Best Demo Award from IEEE SECON 2017, the Best Paper Awards from IEEE GLOBECOM 2016 & 2015, IEEE WCNC 2015, and IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a Fellow of the IEEE.