Quality Improvement and Overhead Reduction for Soft Video Delivery

Takuya Fujihashi[†*], Toshiaki Koike-Akino[†], Takashi Watanabe[*], and Philip V. Orlik[†]
[†]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
[*]Graduate School of Information and Science, Osaka University, Osaka, Japan

***Abstract*—Soft video delivery, i.e., analog video transmission, has been proposed to provide graceful video quality in unstable wireless channels. However, existing analog schemes need to transmit a significant amount of metadata to a receiver for power allocation and decoding operations. It causes large overheads and quality degradation because of rate and power losses. To reduce the overheads while keeping high video quality, we propose a new analog transmission scheme. Our scheme exploits a Gaussian Markov random field for modeling video sequences to significantly reduce the required amount of metadata, which are obtained by fitting into the Lorentzian function. Our scheme achieves not only reduced overhead but also improved video quality, by using the fitting function and parameters for metadata. Evaluations using several test video sequences demonstrate that our proposed scheme reduces overheads by 97 % with 3.4 dB improvement of video quality compared to the existing analog video transmission scheme.**

I. Introduction

Wireless video delivery has been one of major applications in wireless environment. According to Cisco visual networking index studies, three-fourths of the world's mobile data traffic will be video contents by 2020 [1]. In conventional video streaming, the digital video compression and transmission parts operate separately. For example, the video compression part uses H.264/Advanced Video Coding (AVC) [2] or H.265/High-Efficiency Video Coding (HEVC) [3] standard to generate a compressed bit stream using quantization and entropy coding. The transmission part uses a channel coding and a digital modulation scheme to reliably transmit the encoded bit stream.

However, the conventional scheme has the following problems due to the wireless channel unreliability. First, the encoded bit stream is highly vulnerable for bit errors. When the channel signal-to-noise ratio (SNR) falls under a certain threshold and bit errors occur in the bit stream during communications, the video quality drops significantly. This phenomenon is referred to as cliff effect. Second, the video quality does not gracefully improve even when the wireless channel quality is improved. Finally, quantization is a lossy process and its distortion cannot be recovered at the receiver.

To overcome the above-mentioned problems, analog transmission schemes [4]–[8] have been proposed. For example, SoftCast [4], [5] directly transmits linear-transformed video signals over a lossy channel and allocates power for the signals to maximize video quality. In contrast to the conventional scheme, the video quality of SoftCast can be gracefully improved according to the wireless channel quality.

However, the performance of SoftCast is inefficient due to chunk division. In SoftCast, a sender allocates transmission power to the video signals for noise reduction. The magnitude of power allocation is based on the power of each linear-transformed video signal. Hence, the sender needs to transmit the power information of all the video signals without errors to decode the signals at the receiver. The transmission of these metadata causes large overheads, resulting into video quality degradation due to power and rate loss. To reduce the overheads, SoftCast divides the linear-transformed signals into chunks. However, the chunk division can considerably degrade performance due to improper power allocation, in particular for large chunk sizes to reduce metadata.

To improve performance, coset coding [9], [10], subcarrier assignment [11], and rateless coding [12] were adopted for analog schemes. However, they are oblivious of the chunk division. Although the trade-off between chunk size and video quality were discussed in [13], how to effectively reduce the overheads was beyond the scope of the paper.

In this paper, we propose a new analog scheme without chunk division to overcome the issues of conventional analog schemes. To obtain the power values of linear-transformed video signals without transmitting large-overhead metadata, our scheme uses Gaussian Markov random field (GMRF) [14], [15] to model video signals and exploits a Lorentzian-based fitting function at the sender and the receiver. Specifically, the sender finds a few parameters for the fitting function and sends the parameters as metadata to the receiver. The receiver obtains the power values from the fitting function and the received parameters for decoding. Evaluations using test video sequences show that the proposed scheme improves video quality by 3.4 dB with 97 % reduction in the overheads.

Our contribution is two-fold: 1) we verify that the power of linear-transformed video signals are well fit by Lorentzian-based function with eight parameters when the video signals can be modeled by GMRF and 2) we propose fitting-based power allocation and signal reconstruction to achieve improved video quality and reduced overheads simultaneously.

II. Soft Video Delivery

The purposes of our study are 1) to achieve higher video quality with the improvement of the wireless channel quality and 2) to achieve smaller overheads. Fig. 1 shows the

T. Fujihashi conducted this research while he was an intern at MERL.

978-1-4799-6664-6/16/$31.00 ©2016 IEEE

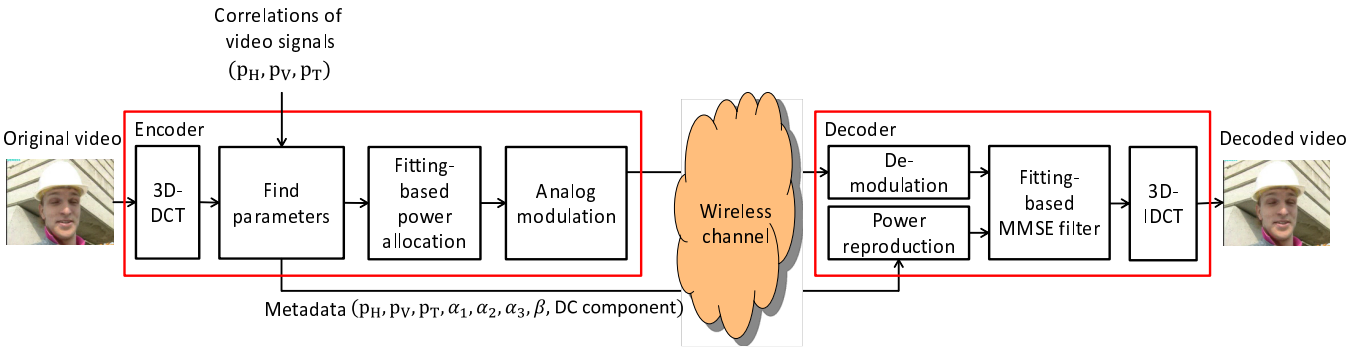


Fig. 1. Proposed analog video transmission scheme.

schematic of our proposed scheme. The encoder first performs 3D-discrete cosine transform (DCT) operation for the original video frames. According to the power of the DCT coefficients, we find the best parameters of a fitting function. The DCT coefficients are then scaled and analog-modulated based on these parameters. Finally, the encoder sends the analog modulated symbols and the fitting parameters to the receiver over a wireless channel with additive white Gaussian noise (AWGN). At the receiver side, the decoder uses minimum mean-square error (MMSE) filter based on the received fitting parameters. This filter is used for the received analog-modulated symbols to obtain the DCT coefficients.

A. Encoder

The encoder first takes 3D-DCT operation for the original sequence to obtain the DCT coefficients. 3D-DCT is used for whole frames in one group of picture (GoP), which is a sequence of successive video frames. The DCT coefficients are mapped to I (in-phase) and Q (quadrature-phase) components after the power allocation.

Let x_i denote the i-th analog-modulated symbol. Each analog-modulated symbol is scaled by g_i for noise reduction:

$$x_i = g_i \cdot s_i. \tag{1}$$

Here, s_i is the i-th DCT coefficient and g_i is the scale factor for the coefficient. After the transmission, each received symbol can be modeled for AWGN channels as follows:

$$y_i = x_i + n_i, \tag{2}$$

where y_i is the i-th received symbol and n_i is an effective noise with a variance of σ^2.

The transmitter performs optimal power controls for g_i to achieve the highest video quality. Specifically, the best g_i is obtained by minimizing the mean-square error (MSE) under the power constraint with total power budget P as follows:

$$\min \quad \text{MSE} = E\left[(x_i - \hat{x}_i)^2\right] = \sum_i^N \frac{\sigma^2 \lambda_i}{g_i^2 \lambda_i + \sigma^2}, \tag{3}$$

$$\text{s.t.} \quad \frac{1}{N}\sum_i^N g_i^2 \lambda_i = P, \tag{4}$$

where $\hat{x}_i$ is an estimate of the DCT coefficient at the receiver, λ_i is the power of i-th DCT coefficient, and N is the number of DCT coefficients. The near-optimal solution is expressed as

$$g_i = \lambda_i^{-1/4} \sqrt{\frac{P}{\sum_j \lambda_j}}. \tag{5}$$

B. Decoder

The receiver extracts DCT coefficients from I and Q components, and reconstructs the coefficients using MMSE filter [4] as follows

$$\hat{s}_i = \frac{g_i \lambda_i^2}{g_i^2 \lambda_i^2 + \sigma^2} \cdot y_i. \tag{6}$$

The decoder then obtains corresponding video sequence by taking the inverse 3D-DCT for the filter output $\hat{s}_i$.

C. Overhead Reduction

In order for the receiver to carry out MMSE filtering in (6), the sender needs to transmit λ_i of all coefficients without errors as metadata. The amount of metadata can be significantly large. For example, when the sender transmits eight video frames with the resolution of 176×144, the sender needs to transmit metadata for all DCT coefficients of $176 \times 144 \times 8 = 202{,}752$ to the receiver. It induces performance degradation due to rate and power losses. To reduce the overheads, conventional methods divide coefficients into chunks and carry out power allocation and MMSE filter for each chunk. However, overheads are still high and the chunk division causes performance degradation due to improper power allocation. When the chunk is a size of 44×36 pixels, 256 metadata are still required every eight frames.

To reduce the overheads, we use a fitting function to approximate the power values λ_i for a variety of video sequences. To this end, our scheme uses GMRF to model video signals. Based on the model, we verify λ_i, except direct current (DC) component, can be fit by a Lorentzian function with seven parameters. The details of the derivation are described in Sec. II-D. Our scheme uses $\hat{\lambda}_i$, which is an estimated power of DCT coefficients obtained from the fitting function, for the power allocation and MMSE filter. To share the same $\hat{\lambda}_i$ at both the sender and the receiver, our scheme transmits eight

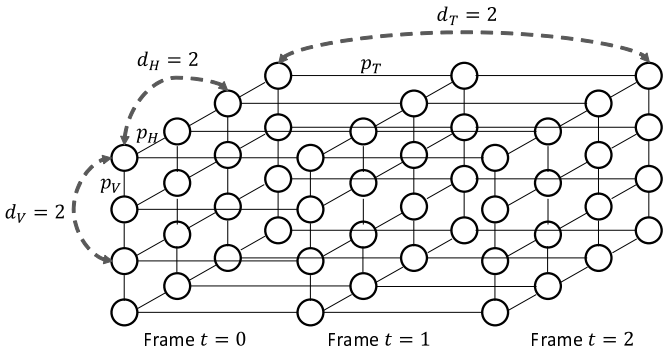


Fig. 2. GMRF model for video signals.

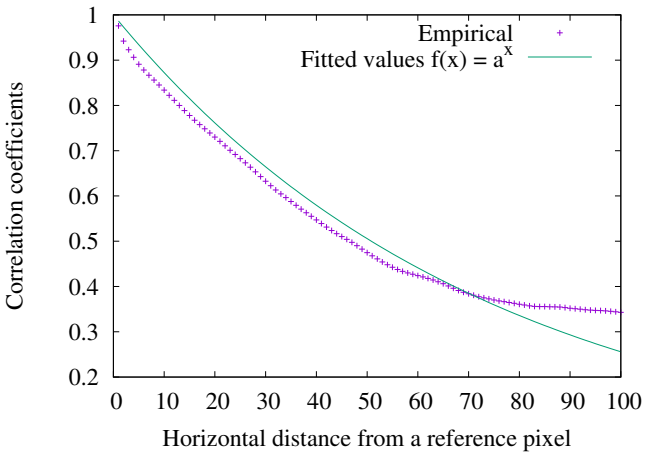


Fig. 3. Fitting horizontal correlation coefficients of *akiyo*.

metadata, which consists of seven fitting parameters and DC component of DCT coefficients. We assume that the encoder uses $1/2$-rate convolutional code and binary phase-shift keying (BPSK) for the metadata transmissions.

D. Fitting Function

We use a simple GMRF to model video signals as shown in Fig. 2. In the video signals, each pixel is connected to three pixels for horizontal, vertical, and time directions. Each direction has different correlations, which are defined as p_{H}, p_{V}, and p_{T}, respectively. In this case, the correlation between two pixels can be described as $p_{\mathrm{H}}^{d_{\mathrm{H}}} \cdot p_{\mathrm{V}}^{d_{\mathrm{V}}} \cdot p_{\mathrm{T}}^{d_{\mathrm{T}}}$, where d_{H}, d_{V}, and d_{T} are horizontal, vertical, and time distances between the pixels, respectively.

The DCT can be regarded as a discrete-time real-valued version of the Fourier transform. After we take the Fourier transform for the signals, the auto-correlation function corresponds to the power spectrum density according to the Wiener–Khintchine theorem. For 3D video signals following the GMRF, the power spectrum density of 3D-DCT coefficients can be asymptotically obtained by the Lorentzian function as follows:

$$F(i,j,k) = \frac{-\beta \log(p_{\mathrm{H}}) \cdot \log(p_{\mathrm{V}}) \cdot \log(p_{\mathrm{T}})}{\{f_1^2(i) + \log(p_{\mathrm{H}})\}\{f_2^2(j) + \log(p_{\mathrm{V}})\}\{f_3^2(k) + \log(p_{\mathrm{T}})\}}, \quad (7)$$

$$f_1(i) = \alpha_1 \frac{\pi i}{N_{\mathrm{H}}}, f_2(j) = \alpha_2 \frac{\pi j}{N_{\mathrm{V}}}, f_3(k) = \alpha_3 \frac{\pi k}{N_{\mathrm{T}}}, \quad (8)$$

where N_{H}, N_{V}, and N_{T} are the number of coefficients in horizontal, vertical, and time domains, respectively. Here, α_k and β are parameters for fitting. Note that above equations express the power spectrum density of alternate current (AC) components in the DCT coefficients. Our scheme ignores the DC component from fitting operation because the DC component cannot be modeled by the Lorentzian function.

E. Correlation Coefficient Estimation

To calculate the fitting function, the encoder estimates the horizontal, vertical, and time correlations of the video sequence, by fitting an empirical auto-correlation function into an exponential function of $f(x) = a^x$ by means of a least-squares method. Fig. 3 shows an example of fitting curves for a video sequence of *akiyo*. In this case, we obtain the estimated parameters p_{H}, p_{V}, and p_{T} of 0.98, 0.98, and 0.99, respectively. From this figure, it is expected that the simple GMRF model depicted in Fig. 2 can capture some useful statistics of real video sequences.

With the estimated correlations, the encoder finds the other fitting parameters based on the empirical power of AC components by least-squares fitting. The encoder then reproduces the power of AC components using the estimated parameters and fitting function. Fig. 4 shows the empirical and fitting power of DCT coefficients within one video frame for the video sequence of *akiyo*. In this case, the estimation error is small; more specifically, the normalized mean-square error (NMSE) between empirical and fitting values is about -25 dB. The proposed scheme can significantly reduce the overheads by transmitting just eight values regardless of the video size.

III. Performance Evaluation

A. Simulation Settings

Metric: We evaluate the performance of reference schemes in terms of the NMSE and peak signal-to-noise ratio (PSNR) defined as follows:

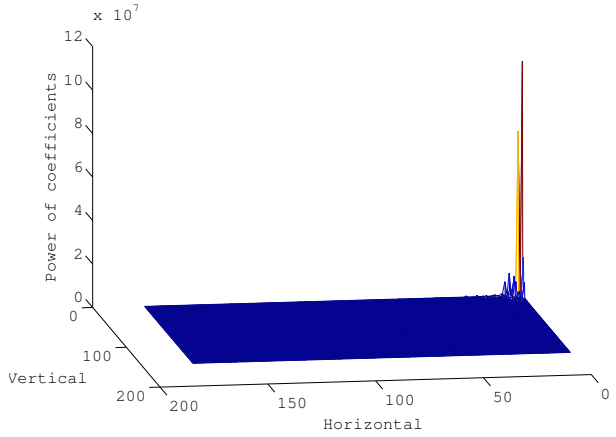
$$\mathrm{NMSE} = 10 \log_{10} \frac{\varepsilon_{\mathrm{MSE}}}{\sum_i^N s_i^2}, \quad (9)$$

$$\mathrm{PSNR} = 10 \log_{10} \frac{(2^L - 1)^2}{\varepsilon_{\mathrm{MSE}}}, \quad (10)$$

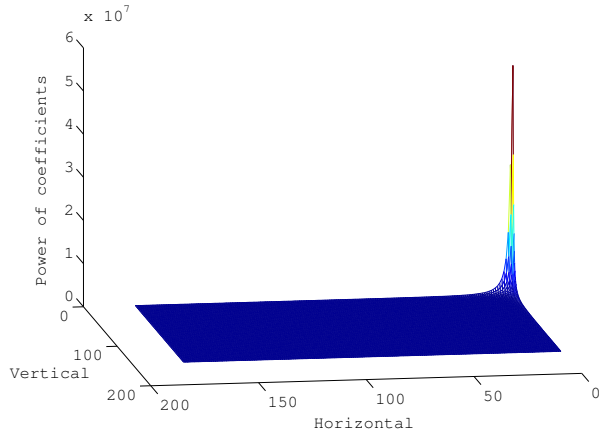
where L is the number of bits used to encode pixel luminance (typically eight bits), and $\varepsilon_{\mathrm{MSE}}$ is the MSE between all pixels of the decoded and the original video. We obtain the average NMSE and PSNR across whole video frames in each video sequence.

Test Video: We use standard reference video, namely, *foreman*, *akiyo*, *mobile*, *coastguard*, and *news* in the QCIF format (176×144 pixels, 30 fps) from the Xiph collection [16].

Wireless Channel Environment: The received symbols are impaired by an AWGN channel.



(a) Empirical



(b) Fitting

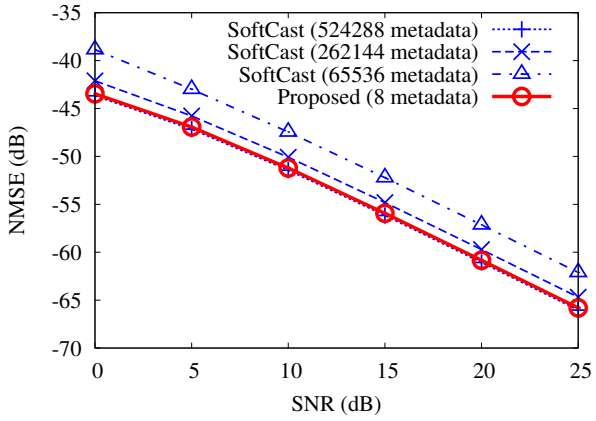
Fig. 4. Power of DCT coefficients: (a) empirical, and (b) fitting.

Amount of Metadata: As we mentioned in Sec. II-C, our proposed scheme sends eight metadata. Conventional analog schemes transmit two metadata (mean and variance) for each chunk.

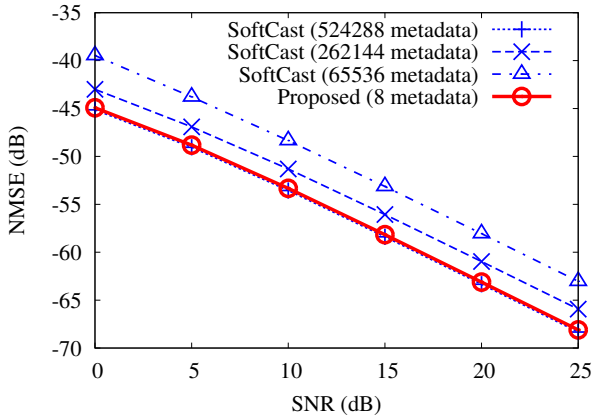
B. Video Signals from GMRF

Before analyzing real video sequences, we first evaluate our proposed scheme for virtual video sequences generated from GMRF model. We assume that the resolution of the signals is $256 \times 256 \times 8$ and the correlations of three domains (horizontal, vertical, and time) are identical. We set the mean and variance of the signals are 128 and 1, respectively. For the comparison, we measure NMSE of the proposed and three SoftCast schemes with different chunk sizes: 1×1, 2×2, and 4×4 pixels. The corresponding number of chunks in SoftCast becomes 524288, 131072, and 32768, respectively. Fig. 5 shows the NMSE with the different correlations: (a) 0.1, (b) 0.5, and (c) 0.9. From these figures, we observe the following two points:

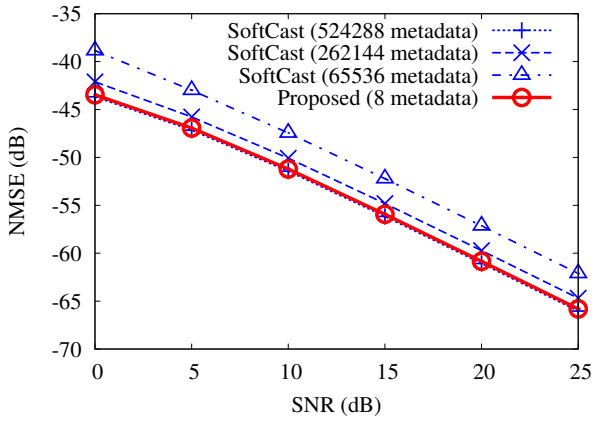
- NMSE of the proposed scheme approaches SoftCast with a minimum chunk size of 1×1, which is an idealistic case. This result means that the estimation error using our fitting function is negligible.
- As the correlations increase, NMSE becomes smaller. When signal correlations are high, most of video information concentrate in lower frequency components. This concentration facilitates protection of analog-modulated



(a) Correlation is 0.1



(b) Correlation is 0.5



(c) Correlation is 0.9

Fig. 5. NMSE vs. SNR for video sequence generated by GMRF: (a) 0.1 correlation, (b) 0.5 correlation, and (c) 0.9 correlation.

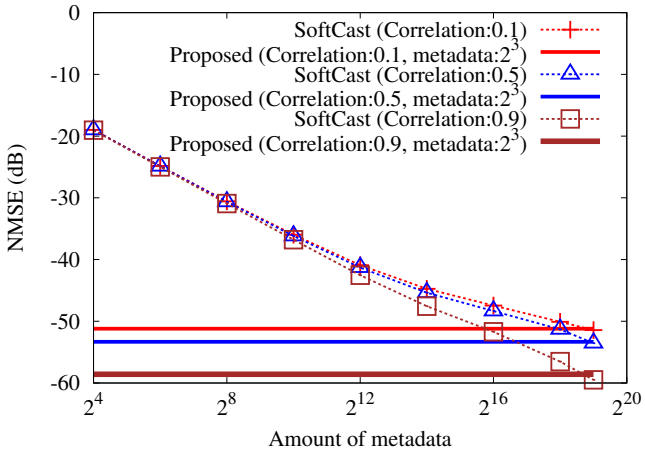


Fig. 6. NMSE vs. amount of metadata at an SNR of 10 dB.

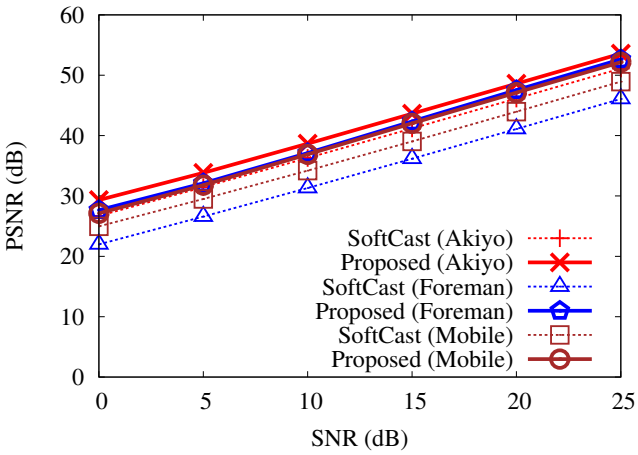


Fig. 7. PSNR vs. SNR with different video sequences.

TABLE I
PARAMETERS FOR FITTING FUNCTION

Video Sequence	p_H	p_V	p_T	α_1	α_2	α_3	β
Akiyo	0.99	0.98	0.99	-0.10	-0.85	-0.48	15419.9
Foreman	0.98	0.96	0.92	-0.36	-0.52	-0.83	24447.8
Mobile	0.98	0.94	0.93	-8.65	-0.89	-0.63	$1.32{\cdot}10^7$
Coastguard	0.99	0.98	0.97	-0.12	-0.89	-0.40	12766.5
News	0.97	0.97	0.99	112.40	-0.61	-0.79	$1.02{\cdot}10^{10}$

symbols from communication noise.

Fig. 6 shows the NMSE with the different chunk sizes at an SNR of 10 dB. Here, we evaluate NMSE of SoftCast with nine chunk sizes: 1×1, 2×2, 4×4, 8×8, 16×16, 32×32, 64×64, 128×128, and 256×256 pixels. The corresponding number of chunks is 524288, 131072, 32768, 8192, 2048, 512, 128, 32, and 8. This figure demonstrates that our proposed scheme significantly outperforms the conventional SoftCast. For example, the proposed scheme improves NMSE approximately by 39.5 dB compared to SoftCast with 2^4 metadata for the correlation of 0.9.

C. Real Video Sequences

Previous evaluations demonstrated that our proposed scheme approaches the performance with the smallest chunk size when video signals are generated from GMRF. However, real video sequences may not follow the model and this model mismatch induces estimation errors. To evaluate the effect on real video sequences, this section uses *foreman*, *akiyo*, *mobile*, *coastguard*, and *news* as the test sequences. Table I lists the values of fitting parameters of each video sequence in the first GoP. We use the chunk size of 44×36 pixels (total 128 chunks) for SoftCast, which is based on [4]. Fig. 7 shows the video quality with the different video sequences. The key results from this figure are summarized as follows:

- The proposed scheme achieves the higher video quality compared to existing SoftCast regardless of test video sequences. For example, the proposed scheme improves video quality approximately by 3.4 dB compared to SoftCast at an SNR of 15 dB for the video sequence of *foreman*.
- The proposed scheme reduces the amount of metadata by 96 % compared to SoftCast. This reduction saves transmission power and leads to additional quality improvement by allocating the saved power for the transmission of analog-modulated symbols.

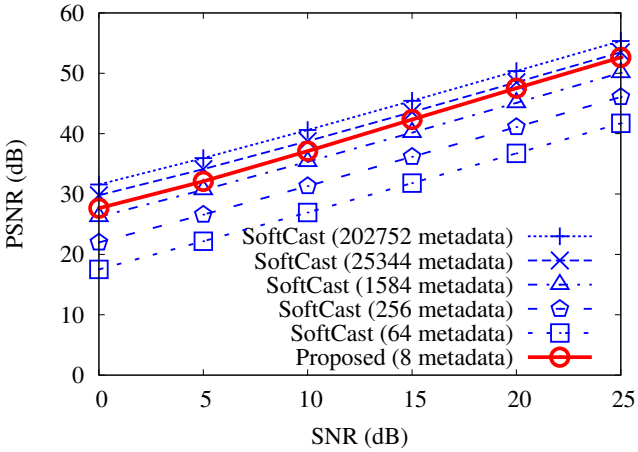


Fig. 8. PSNR vs. SNR of *foreman* with different chunk size.

In addition, we demonstrate that our proposed scheme keeps high video quality approximately by 2.1 and 2.7 dB compared to SoftCast at an SNR of 15 dB for the video sequence of *coastguard* and *news*, respectively.

D. Effect on Amount of Metadata

Previous evaluations revealed that the proposed scheme achieves higher video quality and smaller overheads compared to existing SoftCast with the fixed chunk size. However, the performance of SoftCast can be improved by increasing the overheads. For the detailed discussions, this section compares the different sizes of chunks to demonstrate the impact of our

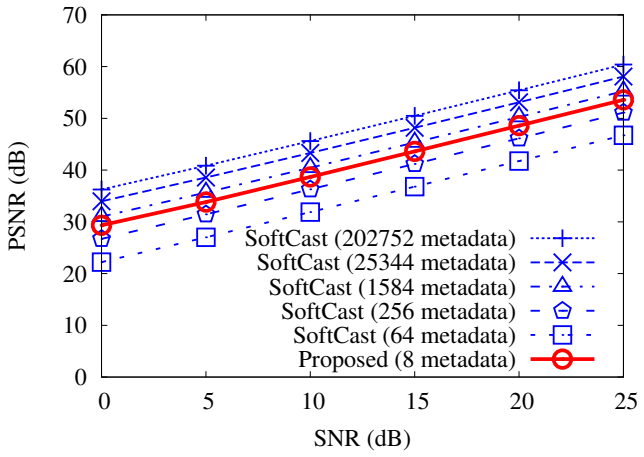


Fig. 9. PSNR vs. SNR of *akiyo* with different chunk size.

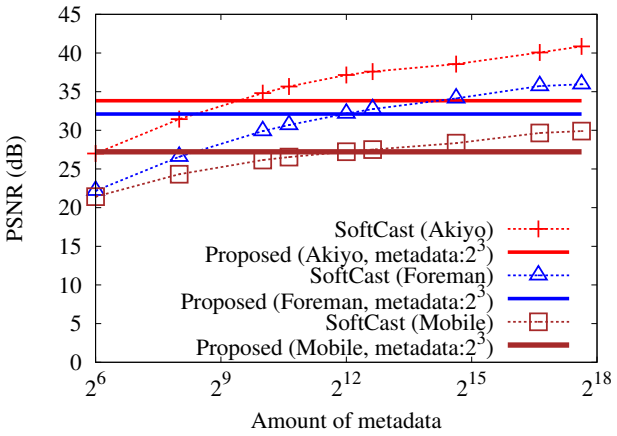


Fig. 10. PSNR vs. amount of metadata at an SNR of 10 dB.

proposed scheme. We evaluate the video quality of proposed scheme and SoftCast schemes with five different sizes of chunk: 1×1, 4×4, 16×16, 44×36, and 88×72 pixels. The corresponding number of chunks becomes 202752, 12672, 792, 128, and 32, respectively. Figs. 8 and 9 show the video quality with the test video sequences of *foreman* and *akiyo*, respectively. Our proposed scheme achieves higher video quality even when the chunk size is 16×16 pixels. Therefore, the proposed scheme can reduce the overheads by at least 99.4 % while achieving better video quality. Note that the estimation errors using fitting function in *foreman* are smaller than *akiyo*.

Fig. 10 shows the video quality with the different amount of metadata in each video sequence. To evaluate the effect of metadata, we plot the performance for nine different sizes of chunk: 1×1, 2×2, 4×4, 8×8, 11×9, 16×16, 22×18, 44×36, and 88×72 pixels. The corresponding number of chunks is 202752, 50688, 12672, 3168, 2048, 794, 512, 128, and 32. Fig. 10 represents that the proposed scheme greatly improves video quality when the overheads of SoftCast are small. Specifically, the improvement of our scheme is 9.9 dB when the amount of metadata of SoftCast is 2^6 for the video sequence of *foreman*.

IV. Conclusion

This paper proposed a new analog transmission scheme based on a simple GMRF model to keep high video quality with the reduction in overhead. The proposed scheme finds parameters for fitting function to obtain the power of DCT coefficients with small overheads. Performance evaluations show that our proposed scheme achieves higher video quality compared to existing analog schemes with the improvement of wireless channel quality. In addition, the proposed scheme significantly reduces the required amount of overheads. This reduction saves transmission power and results in additional quality improvement compared to conventional analog schemes.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2015-2020," Feb. 2016.

[2] W. Thomas, S. G. J, B. Gisle, and L. Ajay, "Overview of the H. 264/AVC video coding standard," *IEEE Transactions of Circuirts And Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[3] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, "Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders," in *IEEE PCS*, 2013, pp. 394–397.

[4] S. Jakubczak, H. Rahui, and D. Katabi, "One-size-fits-all wireless video," in *ACM HotNets*, 2009, pp. 1–6.

[5] S. Jakubczak and D. Katabi, "A cross-layer design for scalable mobile video," in *ACM Annual International Conference on Mobile Computing and Networking*, 2011, pp. 289–300.

[6] X. Lin, Y. Liu, and M. Sun, "Analog channel coding for wireless image/video SoftCast by data division," in *International Conference on Telecommunications*, 2015, pp. 353–357.

[7] S. T. Aditya and S. Katti, "FlexCast: Graceful wireless video streaming," in *ACM MOBICOM*, 2011, pp. 277–288.

[8] X. Lin, Y. Liu, and L. Zhang, "Scalable video SoftCast using magnitude shift," in *IEEE Wireless Communications and Networking Conference*, 2015, pp. 1996–2001.

[9] X. Fan, F. Wu, and D. Zhao, "D-Cast: DSC based soft mobile video broadcast," in *ACM International Conference on Mobile and Ubiquitous Multimedia*, 2011, pp. 226–235.

[10] X. Fan, R. Xiong, D. Zhao, and F. Wu, "Layered soft video broadcast for heterogeneous receivers," *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[11] L. X. Lin, H. Wenjun, P. Qifan, W. Feng, and Z. Yongguang, "Parcast: Soft video delivery in MIMO-OFDM WLANs," in *ACM MOBICOM*, 2012, pp. 233–244.

[12] G. Wang, K. Wu, Q. Zhang, and L. M. Ni, "SimCast: Efficient video delivery in MU-MIMO WLANs," in *IEEE Conference on Computer Communications*, 2014, pp. 2454–2462.

[13] D. Yang, Y. Bi, Z. Si, Z. He, and K. Niu, "Performance evaluation and parameter optimization of SoftCast wireless video broadcast," in *International Conference on Mobile Multimedia Communciations*, 2015, pp. 79–84.

[14] H. Rue and H. Leonhard, *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.

[15] C. Zhang and D. Florencio, "Analyzing the optimality of predictive transform coding using graph-based models," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 106–109, 2013.

[16] Xiph, "Xiph.org media." [Online]. Available: http://media.xiph.org/video/derf/