Distributed soft video broadcast (DCAST) with explicit motion

Xiaopeng Fan[1], Feng Wu[2], Debin Zhao[1], Oscar C. Au[3], and Wen Gao[4]
[1] School of Computer Science, Harbin Institute of Technology, China
[2] Microsoft Research Asia, Beijing, China
[3] ECE department, Hong Kong University of Sci. and Tech., Hong Kong
[4] School of EECS, Peking University, Beijing, China *†

Abstract: Video broadcasting is a popular application of wireless network. However, the existing layered approaches can hardly accommodate users with diverse channel conditions as analog communication can do. The newly emerged 'softcast' approach, utilizing soft broadcast, provides smooth multicast performance but is not very efficient in inter frame compression. In this work, we propose a motion-aligned wireless video multicast scheme DCAST. Instead of using conventional close loop prediction (CLP), DCAST is based on distributed source coding (DSC) theory. This helps DCAST to avoid error propagation but still achieve high compression efficiency in inter frame coding. DCAST outperforms softcast 5dB in video PSNR while maintaining the similar graceful degradation feature as softcast.

1 Introduction

The main challenge of wireless video broadcast is to accommodate different users with different channel conditions and provides the video quality corresponding to their channel conditions respectively. Typical wireless video broadcast schemes based on the DVB-T standard[1] combine a layered transmission scheme[2][3] and scalable video coding (SVC) scheme [4][5]. SVC encodes the video signal into one base layer (BL) and multiple enhancement layers (EL). In transmission, the hierarchical modulation (HM)[6] superimposes the multiple layer bits in one wireless symbol and allow the user to decode different numbers of layers according to their own channel condition. With SVC and HM, low SNR users can receive rough video signal while high SNR

*Dr. Xiaopeng Fan is an associate professor in Harbin Institute of Technology. This work was done in Microsoft Research Asia (MSRA) where he was a visiting researcher in 2011.

†This work was supported in part by the Program for New Century Excellent Talents in University (NCET), the National Science Foundation (NSFC) of China under Grants 61100095 and 60736043, and the Major State Basic Research Development Program of China (973 Program 2009CB320905).

1068-0314/12 $26.00 © 2012 IEEE
DOI 10.1109/DCC.2012.27

IEEE computer society

users can receive high quality video signal. However, the layered schemes reduces both the compression efficiency and the transmission efficiency. Also, such scheme only provides limited choices of BL and EL rates, e.g. DVB-T standard specifies 3 BL rates and 5 EL rates. This creates cliff effects in video quality as opposed to continuously changing channel condition.

In contrast to the digital transmission, analog transmission naturally supports broadcasting to users with different SNR, since the channel noise is directly transformed into reconstruction noise of the video. However, transmitting video signal directly in analog form without compression is inefficient and of low quality. Recently, a novel wireless video broadcasting approach called Softcast [7] has been proposed based on soft compression and soft transmission. Softcast transmits the linear transform of the video signal directly in analog channel without quantization, FEC and modulation. However, Softcast exploits intra frame redundancy only and thus is not very efficient in the aspect of video signal compression. In a recent improved version of Softcast, the utilization of 3D-DCT partially enables inter frame compression[8]. However, without motion compensation the inter frame redundancy is still not fully exploited in 3D-DCT based softcast.

In this paper, we propose a new wireless video multicast approach called DCAST. DCAST utilizes soft broadcast as softcast does. Additionally, we apply motion estimation in DCAST. To apply traditional inter frame coding in soft compression is inefficient due to the inter frame error propagation. Therefore, we utilize the distributed source coding (DSC)[9] technique in DCAST to achieve high efficient inter frame compression and meanwhile avoid error drifting. Instead of transmitting (the linear transform of) the video signal itself, DCAST transmits the coset code [10] of the video signal by raw OFDM. This significantly reduces the magnitude of the signal but the receiver can still decode the signal with the help of the inter frame prediction. Furthermore, to improve the performance, the ME process and MV transmission are optimized. In experiments, the proposed approach achieves significant gain over Softcast. Moreover, the proposed approach has no frame delay and is applicable in realtime applications.

The rest of the paper is organized as follows: Section 2 introduce the background of this paper. Section 3 presents the proposed DCAST. Section 4 gives experimental results and Section 5 concludes the paper.

2 Background

2.1 Softcast

Softcast is a simple but joint design covering the functionality of video compression, channel coding and PHY layer transmission in one scheme. Softcast consists of three steps: transform, power allocation and whitening. Transform removes the spatial redundancy of a video frame. Power allocation minimizes the total distortion by optimally scaling the transform coefficients. Whitening step transforms the coefficients by Hadamard matrix to create packets with equal average power and equal impor-

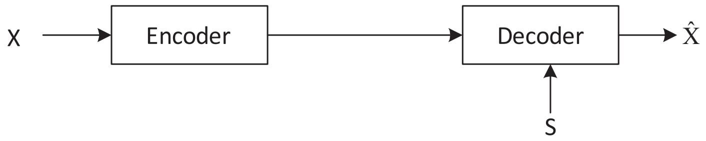


Figure 1: Compression of X when its side information S is available at the decoder

tance. All the steps are linear operations thus the channel noise is directly transformed into reconstruction noise of the video. Therefore, Softcast can accommodate multiple user with different channel SNR. In addition, by skipping low importance coefficients Softcast can also efficiently broadcast video in narrow band channels.

More importantly, softcast accommodates multiple users without scarifying the performance of any single users. According to the experimental result in [8], **softcast, as a broadcast approach, can give each individual client the corresponding video quality that a best conventional unicast approach (H.264+channel coding+QAM) can provide**.

2.2 Distributed source coding

The main difficulty to enable inter frame coding in softcast is the error propagation problem. Typical inter frame coding schemes utilize close loop prediction (CLP), i.e. encode the motion compensated difference between a video frame and its previous frame. However, in Softcast, the noise of each frame will add to its following frames if we only transmit their difference. This will successively reduce the reconstruction quality frame after frame. The main reason to cause this problem is that the encoder cannot exactly know the decoder reconstruction frame.

To compress a source with its prediction only available at decoder is a typical problem in distributed source coding(DSC). As shown in Fig. 1, X is the source representing the current video frame, S is its side information representing the predicted frame. The theoretical foundations of DSC, the Slepian-Wolf theorem[11] and the Wyner-Ziv theorem[12], presents an important conclusion that, a source X can be efficiently compressed with its predictor S only available at the decoder.

Practically, DSC employs coset coding [13] or syndrome coding[14]. Accompanied by the advances of the practical solutions, DSC has found considerable usage in video compression[15][13].

3 Proposed DCAST approach

The proposed DCAST approach is a wireless video multicast system based on soft broadcast. It utilizes linear transform and distributed source coding to remove both intra frame redundancy and inter frame redundancy.

Fig. 2 depicts the server side of DCAST. DCAST first transforms the original image into DCT domain. Meanwhile, DCAST performs ME and MC on the original video sequence to get the encoder predictions and MVs. Then DCAST applies coset

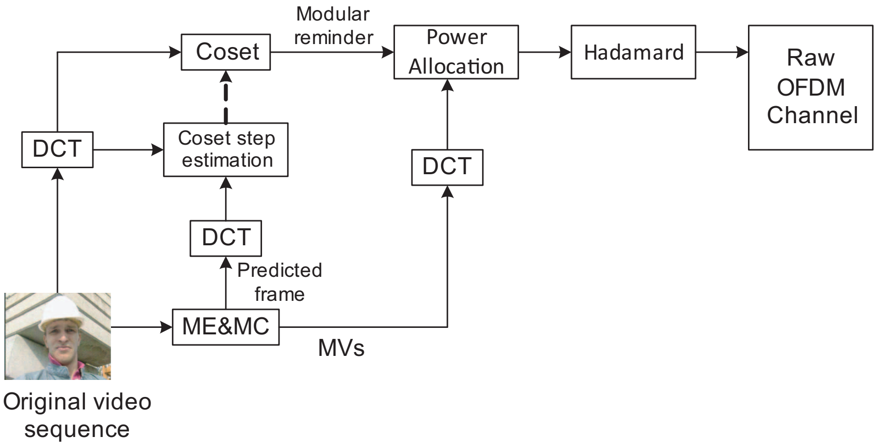


Figure 2: DCAST server

coding on the transform coefficients of the original image to get, for each DCT coefficients, the modular reminder. The step of the coset coding is estimated by using the information of the encoder prediction. The MVs of the current frame, in the form of a matrix, is also transformed by DCT. The modular reminders and the transformed coefficients of the MVs are then scaled for optimal power allocation. Then, before soft transmission, Hadamard transform is applied on the signal to whiten the noise. At last, the resulting signal is directly transmitted over the raw OFDM channel as if it was analog signal.

The client side of DCAST is depicted in Fig. 3. The signal received from the raw OFDM channel is raw signal plus channel noise. After inverse Hadamard transform, the DCT coefficients of the coset values and the MVs are estimated by MMSE. The MVs are transformed back to spatial domain by inverse DCT. Then the MC module generates the predicted frame by the MVs and the reference frame. The predicted frame is transformed into frequency domain by DCT. Then with the coset values and the predictors, the coset decoding module recovers the DCT coefficients of the current frame. At last, the signals are transformed back to spatial domain, and are linearly combined with the predicted signals by MMSE to generate the final reconstruction.

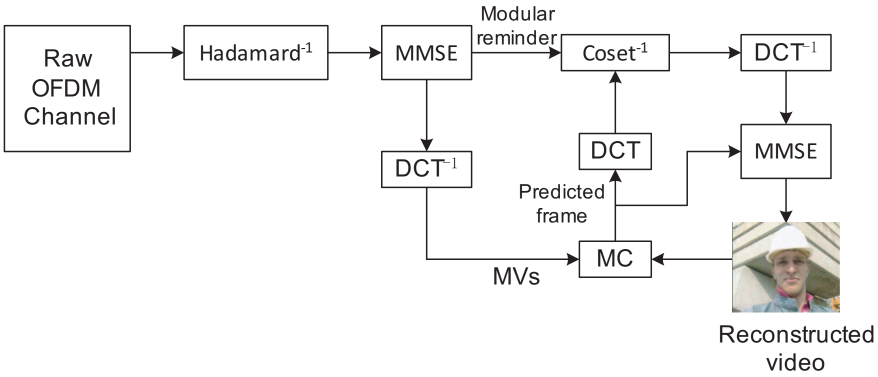


Figure 3: DCAST client

3.1 Coset coding

Coset coding is a typical technique used in DSC. It partitions the set of possible input source values into several cosets and transmits the coset index to the decoder. With the coset index and the predictor, the decoder can recover the source value by choosing the one in the coset closest to the predictor. Coset coding achieves compression because the coset index has typically lower entropy than the source value.

The proposed DCAST uses a special coset code with real value input and real value output. The proposed approach divides each transform coefficient X by a step q and get the remainder L as follows.

$$L = X - \lfloor \frac{X}{q} + \frac{1}{2} \rfloor q \tag{1}$$

L is the coset index although it is real value. This is actually throwing away the main part of X. In some sense L represents the detail of X.

At the user side, with the received coset value $\hat{L}$ and the side information S (i.e. the predicted DCT coefficients), the receiver reconstructs the DCT coefficients by coset decoding. Given the coset value $\hat{L}$, there are multiple possible reconstructions of X forming a coset $\mathcal{C}$.

$$\mathcal{C} = \{\hat{L}, \hat{L} \pm q, \hat{L} \pm 2q, \hat{L} \pm 3q, ...\} \tag{2}$$

DCAST selects in $\mathcal{C}$ the one nearest to the side information S as the reconstruction of the DCT coefficient.

$$\hat{X} = \arg\min_{c \in \mathcal{C}} |c - S| \tag{3}$$

The value of q is calculated by estimating the noise of the decoder prediction as shown in [16].

3.2 Channel Coding: Power allocation and Whitening

The channel coding of DCAST is similar to softcast[8]. The coset data, i.e. the L value of each DCT coefficients are encoded using several linear operations. First the DCT coefficients are divided into 64 subbands and for each subband i we calculate the variance $\sigma_L^2(i)$ of the L values. Then all L are scaled for optimal power allocation between different subbands. Let P be the total power, and g_i be the gain (scaling factor) of L_i, the optimal power allocation in terms of minimizing the distortion is

$$\tilde{L}_i = g_i L_i, \quad g_i \quad = \quad \left(\frac{\sigma_L^{-1}(i) P}{\sum \sigma_L(i)} \right)^{1/2} \tag{4}$$

After this optimal scaling, the variances of the $\tilde{L}$ values of each subbands are still different. To redistribute energy, the $\tilde{L}$ values from different subbands are combined together to form vectors and the new vectors are transformed by Hadamard matrix. This creates packets with equal energy and equal importance. In PHY layer, the packets are directly mapped into transmitted signal. This direct mapping is by modifying the existing 802.11 PHY layer to allow raw data to bypass the FEC and QAM. The gain g of each subband are compressed and transmitted to receiver side by standard 802.11 PHY layer with FEC and BPSK modulation.

3.3 Optimal MV coding

The communication of motion information are also through soft broadcast. Similar to the transmission of the pixel value, DCAST performs 2D DCT transform on the motion vector field $\mathbf{I_{MV}}$ to obtain the transform coefficients C_{MV}.

$$\mathbf{C_{MV}} = \mathbf{A}\mathbf{I_{MV}}\mathbf{A}^T \tag{5}$$

However, optimal power allocation of the C_{MV} is different from the one of the pixel value. The size of the motion field is only 1/64 of the image resolution for 8x8 block size. Thus it is inefficient to divide the C_{MV} into subbands and transmit the gain g of each subbands.

In DCAST, the gain g of C_{MV} are derived as follows. We model the motion vector field I_{MV} as random Markov field with auto correlation function $r(l,m) = \sigma_{MV}^2 \rho^{|l|}\rho^{|m|}$ where l and m are spatial distance between two MVs and ρ is the correlation coefficient between the MVs of two immediate neighbor blocks. According to [17], the transform domain variance and the spatial domain variance has following linear relation:

$$\mathbb{E}(\mathbf{C_{MV}^2}) = \sigma_{MV}^2 \mathbf{V_{MV}} \tag{6}$$

where $\mathbf{C_{MV}^2}$ is the element-wise square of matrix $\mathbf{C_{MV}}$, and

$$\mathbf{V_{MV}} \quad = \quad \mathsf{diag}(\mathbf{A}\mathbf{R_{MV}}\mathbf{A}^T)\mathsf{diag}(\mathbf{A}\mathbf{R_{MV}}\mathbf{A}^T)^T \tag{7}$$

is a scaling matrix with

$$\mathbf{R_{MV}} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}. \tag{8}$$

Then with the $\mathbb{E}(\mathbf{C_{MV}^2})$ in 6, DCAST calculates the optimal gain of each C_{MV} at both encoder and decoder. The calculation of the optimal gain g_{MV} is similar to the calculation of the g in Eq.4. In DCAST, the ρ is estimated at the encoder and is sent to the decoder side together with all the g_i in the Eq.4. Let $\mathbf{G_{MV}}$ be the matrix formed by optimal gain g_{MV}. The optimal coding of the DCT coefficients of MVs is

$$\tilde{\mathbf{C}}_{\mathbf{MV}} = \mathbf{G_{MV}} \odot \mathbf{C_{MV}}. \tag{9}$$

where $\odot$ denotes element-wise multiplication.

3.4 Optimized motion estimation

DCAST performs block based motion estimation. The target of the ME in DCAST is to get, at decoder side, the best prediction in terms of MSE. The prediction noise at the decoder consists of three components: the encoder prediction noise, the noise propagated from reference pixel, and the noise caused by MV error. In this work, we assume these three components are independent, and thus we have:

$$N_{pred}^{dec} = N_{pred}^{enc} + N_{ref} + N_{MV} \tag{10}$$

where N_{pred}^{dec} is the MSE of decoder prediction, N_{pred}^{enc} is the MSE of encoder prediction, N_{ref} is the noise variance of reference pixels, N_{MV} is the additional noise variance caused by MV error.

Let $s^*(l)$ be the encoder prediction and $s(l)$ be the decoder prediction. With the analysis of the power density function [18], the additional prediction error is:

$$N_{MV} \quad = \quad \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} 2\Phi_{ss}(\boldsymbol{\omega})(1-\mathbb{E}\{cos(\boldsymbol{\omega}^T\boldsymbol{\Delta})\})d\boldsymbol{\omega} \tag{11}$$

where $\boldsymbol{\Delta}$ is the MV error. In this work, we assume the MV error $\boldsymbol{\Delta}$ satisfies Gaussian distribution with zero mean and covariance $\frac{1}{2}\sigma_{\boldsymbol{\Delta}}^2\mathbf{I}_{2\times 2}$. For small $\sigma_{\boldsymbol{\Delta}}^2$,

$$N_{MV} \quad \approx \quad \frac{1}{8\pi^2}\sigma_{\boldsymbol{\Delta}}^2\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \Phi_{ss}(\boldsymbol{\omega})\boldsymbol{\omega}^T\boldsymbol{\omega}d\boldsymbol{\omega} \tag{12}$$

Furthermore, the variance $\sigma_{\boldsymbol{\Delta}}^2$ can be calculated by using the signal power of the MV and the SNR of the channel. Therefore, N_{MV} can be estimated by

$$N_{MV} \approx \gamma\sigma_{\boldsymbol{\Delta}}^2 = \gamma\sigma_{MV}^2\left(\frac{Es}{N_0}\right)^{-1} = \gamma\left(\frac{Es}{N_0}\right)^{-1}\mathbb{E}(I_{MV}^2). \tag{13}$$

where the $\frac{Es}{N_0}$ is the channel SNR and $\gamma = \frac{1}{8\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\Phi_{ss}(\boldsymbol{\omega})\boldsymbol{\omega}^T\boldsymbol{\omega}d\boldsymbol{\omega}$.

In 10, the noise variance of reference pixels are constant to the ME process of the current frame. Therefore the minimization of 10 is equivalent to the minimization of following cost function:

$$J = MSE + \lambda I_{MV}^2 \tag{14}$$

where the $\lambda = \gamma\left(\frac{Es}{N_0}\right)^{-1}$. The value of $\frac{Es}{N_0}$ is the SNR of the user we would like to optimize for.

3.5 MMSE at decoder

The proposed approach contains two Minimum Mean Square Estimator (MMSE), operating in transform domain and spatial domain respectively.

The first MMSE is to reconstruct the coset value L in transform domain with minimum distortion. The received signal can be written as:

$$Y = HGL + N \tag{15}$$

where H is Hadamard matrix, G is the diagonal matrix for power allocation, and N is the channel noise. The MMSE reconstruction $\hat{L}^*$ is

$$\begin{aligned} \hat{L}^*(i) &= \frac{\sigma^2_{L(i)}}{\sigma^2_{L(i)} + \sigma^2_N / g_i^2} \tilde{L}(i) && (16) \\ \hat{L} &= (HG)^{-1} Y && (17) \end{aligned}$$

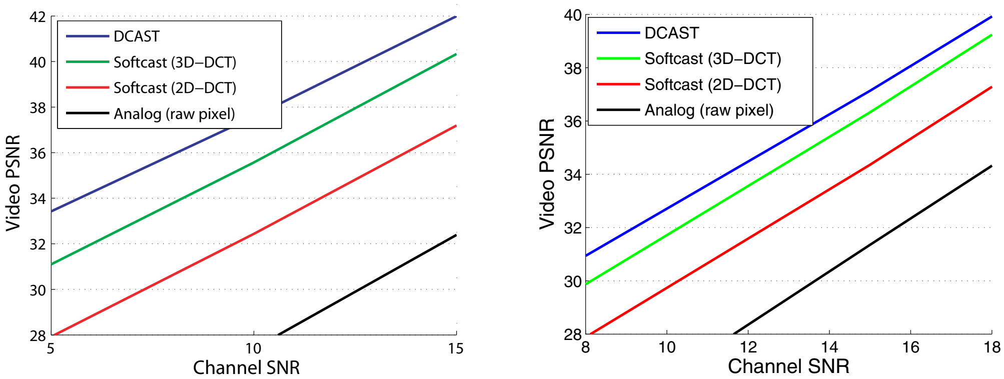
where $\sigma^2_{L(i)}$ and σ^2_N are the variance of $L(i)$ and N respectively.

The second MMSE is to reconstruct the pixel value x in spatial domain with minimum distortion. Considering the coset output $\hat{x}$ as the first noisy observation and the predicted pixel S as the second observation, the current pixel x is reconstructed by following MMSE:

$$\begin{aligned} x^* &= \alpha s + (1 - \alpha)\hat{x} && (18) \\ \alpha &= \frac{\sigma^2_{\hat{x}-x}}{\sigma^2_{s-x} + \sigma^2_{\hat{x}-x}} && (19) \end{aligned}$$

where $\sigma^2_{\hat{x}-x}$ is the variance of the original reconstruction noise, and σ^2_{s-x} is the variance of the prediction noise. In DCAST, the prediction noise variance is estimated at block level. Since $\hat{x}$ is close to x, the prediction noise variance is estimated by calculating $\sigma^2_{s-\hat{x}}$. The reconstruction noise variance is estimated in frame level. The estimation is based on the power of the received signal (the coset value) and the SNR of the channel.

4 Experiments



(a) Foreman_qcif (b) Bus_qcif

Figure 4: Performance comparison in AWGN channel

In experiments, we evaluate the performance of the proposed DCAST in video multicast. We compare DCAST with Softcast[7][8] and analog transmission. Analog transmission means to transmit the video pixels directly over the channel. The test is video multicast to users with diverse SNR.

Both DCAST and Softcast encode the video into packets by soft compression. The video test sequences are 'foreman_qcif.yuv' and 'bus_qcif.yuv'. The video frame rate is 30Hz. The GOP structure is 'IPPP...'. The channel bandwidth is equal to the video bandwidth (i.e. the number of video pixels per second). Note that DCAST have to transmit the MVs to the receiver. In our implementation, the ME is of block size 8x8. Thus the MVs occupies about 2/64 of the bandwidth. In both softcast and DCAST, the number of subbands is also 64. To be fair in bandwidth occupation, we let DCAST to discard for each frame two subbands with minimum prediction error.

After soft compression, DCAST and Softcast communicate the video packets by soft transmission. The video packets are transmitted to OFDM. The OFDM signal is transmitted over AWGN channel. The receiver passes the signal to the OFDM module to perform CFO corrections, channel estimation and correction, and phase tracking. Then it inverts the operations of the transmitter and forwards the soft information to video decoding layer.

The results are given in Fig.4. The 2D-DCT based softcast (softcast2D) [7] is 4-6dB better than analog transmission. The 3D-DCT based softcast (softcast3D) [8] is about 2-3dB better than softcast2D. Our DCAST is 3-6dB better than softcast2D, and is 0.8-2.2dB better than softcast3D. Moreover, DCAST does not introduce frame delays as softcast3D do and is applicable for realtime video multicast like softcast2D. The visual quality comparison is given in Fig.5. The channel SNR is set to be 10dB. It is clear that DCAST has better visual quality than softcast.

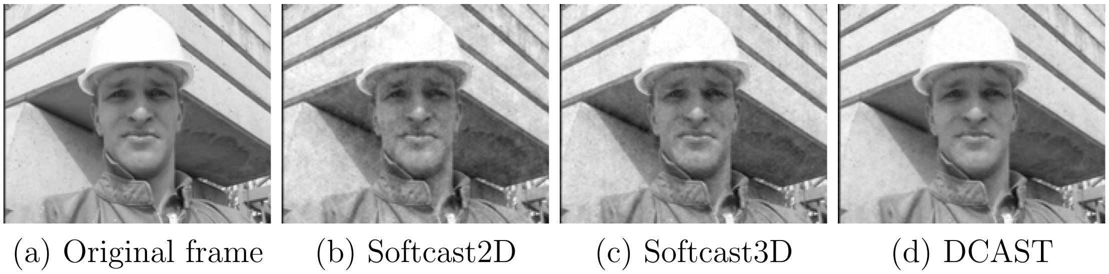

(a) Original frame (b) Softcast2D (c) Softcast3D (d) DCAST

Figure 5: Visual quality comparison, the 5^{th} frame of foreman_qcif.yuv, SNR=10dB

5 Conclusions

In this work we propose a DSC based video multicast approach: DCAST. DCAST is based on soft broadcast, thus performs gracefully in video multicast. DCAST applies DSC principle into video multicast and benefits from motion compensation while avoiding the error propagation. DCAST comprehensively utilizes linear transform, coset coding to achieve better performance than the state-of-art multicast approach Softcast.

References

[1] "Digital Video Broadcasting (DVB)," Website, 2009, http://www.etsi.org/deliver/etsi_en/300700_300799/300744/01.06.01_60/en_300744v010601p.pdf.

[2] N. Shacham, "Multipoint communication by hierarchically encoded data," in *INFOCOM '92. Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE*, may 1992, pp. 2107 –2114 vol.3.

[3] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven layered multicast," in *Conference proceedings on Applications, technologies, architectures, and protocols for computer communications*, ser. SIGCOMM '96. New York, NY, USA: ACM, 1996, pp. 117–130. [Online]. Available: http://doi.acm.org/10.1145/248156.248168

[4] F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 3, pp. 332 –344, mar 2001.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103 –1120, sept. 2007.

[6] K. Ramchandran, A. Ortega, K. Uz, and M. Vetterli, "Multiresolution broadcast for digital hdtv using joint source-channel coding," in *Communications, 1992. ICC '92, Conference record, SUPERCOMM/ICC '92, Discovering a New World of Communications., IEEE International Conference on*, jun 1992, pp. 556 –560 vol.1.

[7] S. Jakubczak, H. Rahul, and D. Katabi, "One-Size-Fits-All Wireless Video," in *Proc. Eighth ACM SIGCOMM HotNets Workshop*, New York City, NY, October 2009.

[8] S. Jakubczak and D. Katabi, "A cross-layer design for scalable mobile video," in *Proceedings of the 17th annual international conference on Mobile computing and networking*, ser. MobiCom '11. New York, NY, USA: ACM, 2011, pp. 289–300. [Online]. Available: http://doi.acm.org/10.1145/2030613.2030646

[9] S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes(DISCUS): design and construction," *IEEE Trans. Inform. Theory*, vol. IT-49, pp. 626–643, 2003.

[10] ——, "Distributed source coding using syndromes (DISCUS): design and construction," in *Proc. IEEE Data Compression Conf.*, 1999, pp. 158–167.

[11] J. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, 1973.

[12] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–10, 1976.

[13] R. Puri, A. Majumdar, and K. Ramchandran, "PRISM: a video coding paradigm with motion estimation at the decoder," *IEEE Trans. Image Processing*, vol. 16, no. 10, pp. 2436–2448, 2007.

[14] J. Garcia-Frias and Y. Zhao, "Compression of correlated binary sources using turbo codes," *IEEE Commun. Lett.*, vol. 5, pp. 417–419, 2001.

[15] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," in *Proc. IEEE*, vol. 93, 2005, pp. 71–83.

[16] X. Fan, F. Wu, and D. Zhao, "D-Cast: DSC based Soft Mobile Video Broadcast," in *ACM International Conference on Mobile and Ubiquitous Multimedia (MUM)*, Beijing, China, December 2011.

[17] A. Jain, *Fundamentals of digital image processing.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1989.

[18] A. Secker and D. Taubman, "Highly scalable video compression with scalable motion coding," *Image Processing, IEEE Transactions on*, vol. 13, no. 8, pp. 1029–1041, 2004.