Scalable Video Coding with Compressive Sensing for Wireless Videocast

Siyuan Xiang and Lin Cai
Electrical and Computer Engineering, University of Victoria

***Abstract*—Channel coding such as Reed-Solomon (RS) and convolutional codes has been widely used to protect video transmission in wireless networks. However, this type of channel coding can effectively correct error bits only if the error rate is smaller than a given threshold; when the bit error rate is underestimated, the effectiveness of channel coding drops dramatically and so does the decoded video quality. In this paper, we propose a low-complex, scalable video coding architecture based on compressive sensing (SVCCS) for wireless unicast and multicast transmissions. SVCCS achieves good scalability, error resilience and coding efficiency. SVCCS encoded bitstream is divided into base and enhancement layers. The layered structure provides quality and temporal scalability. While in the enhancement layer, the CS measurements provide fine granular quality scalability. In addition, we incorporate the state-of-the-art technologies to improve the compressive sensing coding efficiency. Experimental results show that SVCCS is more effective and efficient for wireless videocast than the existing solutions.**

I. Introduction

Video transmission over wireless networks is a challenging task. Compared to other types of data transmission, video applications have stringent Quality of Service (QoS) requirements. On the other hand, due to the inherent impairments of wireless channels, channel error, erasure and variation may degrade video fidelity or even make the video bitstream undecodable.

In order to reduce the error and erasure of wireless channels, error correction coding such as Reed-Solomon (RS) code and convolutional code has been widely used. However, this type of channel coding is not flexible. It can correct the bit errors only if the error rate is smaller than a given threshold. Therefore, it is hard to find a single channel code suitable for unknown or varying wireless channels.

For unicast applications, retransmission in the link layer or the transport layer can help recover the errors at the cost of delay. When we utilize the broadcast nature of wireless medium to multicast video, due to the independence of different receivers' channels, the data needed to retransmit are different for different receivers, which makes retransmission difficult and expensive.

Can we find a flexible channel coding for wireless unicast and multicast? That is, for a wide range of channel error rate, the effectiveness of channel coding degrades gracefully when the channel condition becomes worse. In addition, for multicast applications, without the feedback from individual receivers, the sender can retransmit data that are helpful to all the receivers. These requirements are indeed difficult and challenging for traditional channel coding design.

Thanks to the recent advance in signal processing, the newly developed compressive sensing (CS) technologies can help to achieve the above goals. Compressive sensing or compressive sampling [8][3] has been proposed as a new data acquisition framework which can sample and compress sparse or compressible signals in a single operation. Besides, in the research community, people become more and more interested in the characteristics of the acquired measurements. With CS, random linear projection of signals not only makes the encoder very simple but also makes the acquired measurements *democratic* [11], i.e., they are equally important. For compressible signals such as image and video, compressive sensed measurements demonstrate good scalability, i.e., the more measurements are received, the better user-perceived image and video quality is.

If we only treat compressive sensing as an image compression method, there is a huge gap in terms of coding efficiency between compressive sensing and conventional coding methods [10]. Although compressive sensing has the advantage of being a joint source and channel coding, its coding efficiency needs to be improved, since minimizing bandwidth consumption is one of the most important goals in codec design, particularly for wireless transmissions.

The main contributions of this paper are twofold. First, we propose a low-complex, scalable video coding architecture based on compressive sensing (SVCCS) for wireless unicast and multicast transmissions. SVCCS achieves good scalability, error resilience and coding efficiency. A SVCCS encoded bitstream is divided into a base and an enhancement layer. The layered structure provides quality and temporal scalability. The base layer is composed of a small portion of discrete cosine transform (DCT) coefficients. The enhancement layer consists of compressive sensed measurements. While in the enhancement layer, the CS measurements provide fine granular quality scalability. In addition, we incorporate the state-of-the-art technologies of compressive sensing (e.g., analysis-based ℓ_1 minimization and dictionary) to improve the coding efficiency. Second, we study the performance of SVCCS and the contribution of each component of the codec. We also compare the performance of SVCCS and MJPEG in wireless video multicast.

The rest of the paper is organized as follows. Section II gives a brief introduction of compressive sensing and the related work. Section III describes the architecture of SVCCS. The performance of SVCCS is studied in section IV, followed by concluding remarks in section V.

II. Background and Related Work

A. Background of Compressive Sensing

Compressive sensing or sampling [8][3] was proposed as a new acquisition framework which can sample and compress sparse or compressible signals in a single operation.

Suppose that a signal $x \in R^n$ can be transformed to a coefficient vector θ with some basis Ψ, i.e., $x = \Psi\theta$. Ψ can be any representing basis such as DCT or wavelet. The measurements of compressive sensing, $y \in R^m$, are obtained by multiplying signal x with a measurement matrix $\Phi \in R^{m \times n}$, i.e., $y = \Phi x$. Since $m < n$, (1) is an underdetermined system with infinite solutions. Using the reverse operation in (1) to recover x is infeasible.

$$y = \Phi\tilde{x} \tag{1}$$

However, [8] and [3] have shown that ℓ_1 minimization may recover the original signal with high probability, which can be formulated as

$$\begin{aligned} \text{min} \quad & \|\tilde{\theta}\|_{\ell_1} \\ \text{subject to} \quad & \|y - A\tilde{\theta}\|_{\ell_2} \leq \epsilon, \end{aligned} \tag{2}$$

where $A = \Phi\Psi$ and ϵ is the noise energy in the measurements. Then the recovered signal $\hat{x} = \Psi^*\hat{\theta}$, where Ψ^* is the adjoint of Ψ and $\hat{\theta}$ is the solution to (2).

In order to make recovery stable and accurate, sensing matrix A must satisfy the restricted isometry property (RIP). Reference [3] showed the methods of generating sensing matrix holding RIP. One of them is to randomly select m rows from the Fourier matrix. When condition

$$m \geq CS(\log\, n)^4 \tag{3}$$

is satisfied, the sensing matrix A obeys RIP with overwhelming probability, where C is a constant.

For compressible signals, [3] also showed that if there are $O(S\, \log\, n)$ measurements, the recovery is as good as knowing x and selecting the largest S coefficients from θ.

There are four important observations which are exploited in this paper. 1) The sufficient condition (3) for RIP only cares about the number of rows of Fourier matrix instead of which row is selected. In other words, *CS measurements are equally important*. The property is also called democracy [11]. 2) *Compressive sensing is scalable*. The more measurements, the smaller recovery error is. 3) Given a fixed number of measurements, the faster the signal decays, the smaller the recovery error is. 4) The recovery error is proportional to noise energy ϵ. The noise may include quantization and transmission errors, which should be carefully managed.

B. Related Work

There are a few related work using compressive sensing as joint source and channel coding framework. [6] used compressive sensing to protect sparse signals over erasure channels. The proposed method compensates the lost measurements during transmission by sending more measurements.

With respect to image and video transmission, [12] designed a video encoder based on CS and a streaming protocol for wireless video transmission. To exploit temporal redundancy, the difference frame of the I frame and the target frame is compressive sensed in [12]. This means the original video frame is used as the reference frame in the encoder side, while the decoder uses the recovered image as the reference frame. The discrepancy will degrade decoded video quality. When there is transmission error in the reference frame, the error will affect the frames depending on it.

In [15], the hardware and algorithm to acquire image and video signals were proposed, which exploited the correlation of adjacent frames. They used the joint measurement matrix and 3D wavelets as the representing basis, encoding several frames or even the whole video sequence and then recovering the frames together. However, this method increases computational complexity and also increases both the encoding and decoding delay.

Therefore, in this paper, we propose a low-complex, scalable video coding architecture based on compressive sensing, which utilizes temporal redundancy and the state-of-the-art technologies to improve the compressive sensing coding efficiency and makes it resilient to transmission errors.

III. Proposed Video Coding Architecture

Previously, coding efficiency of compressive sensing cannot compete with conventional codec such as JPEG or MPEG4 when dealing with already acquired image or video signals with high resolution and quality. Compressive sensing only shows its advantage when it acquires and compresses image at the same time. Our goals of designing SVCCS are two-fold: 1) improve its coding efficiency, and 2) make it error resilient. In this section, we describe how these goals are achieved.

A. Layered Structure Design

Figs. 1 and 2 illustrate the proposed video encoder and decoder architecture, respectively. As shown in Fig. 1, video frames are divided into two categories, i.e., I frames and P frames. I frames are DCT transformed and d coefficients are extracted in a zigzag order, then uniformly quantized and entropy encoded. Although the number of these coefficients is small, they contain the majority energy of the image. Therefore, after these coefficients are inversely quantized and inversely DCT transformed, the resultant image provides moderate image quality and can be used as a reference frame. Then the difference between the reference frame and the I or P frame, called the difference frame, is fed into the compressive sensing block.

The concept of the reference frame comes from conventional video codecs, where temporal redundancy of adjacent frames is exploited to improve coding efficiency. For conventional video codecs, the difference of adjacent images contains less energy thus needs less bits to represent. For compressive sensing, the difference of adjacent images is more sparse and compressible.

Motion compensation is not adopted although it can improve the coding efficiency greatly, since it is the most computational complex operation in the traditional encoder. In addition, motion compensation requires to transmit the motion vectors error-free. If motion vectors are corrupted during transmission, degradation of received video quality is inevitable.

In order to minimize transmission errors, a small portion of DCT coefficients are chosen as the reference frame instead of the whole I frame, since it is easier and less costly to protect the small amount of the DCT coefficients than the whole I frame. As any part in the reference frame is corrupted, the error will propagate among the entire group of pictures (GOP). We can see the similarity between the proposed video codec and the scalable video coding (SVC) [14]. The reference frame constitutes the base layer and the compressive sensed difference frame composes the enhancement layer. The layered structure of the proposed video coding architecture thus preserves the quality and temporal scalability.

Fig. 3 shows the layered structure of frames. The GOP size is four. The arrowed dashed lines indicate the dependence between frames. From the figure, we can see that P frames are dependent on the closest I frame(s)' reference frame(s). The P frame in the middle is dependent on the average of the two reference frames to better exploit temporal redundancy.

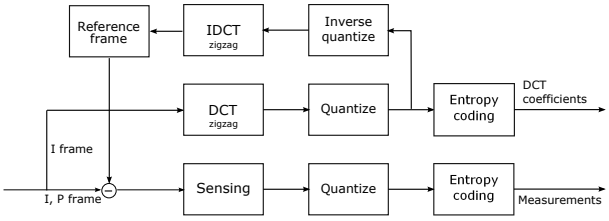

Fig. 1. encoder

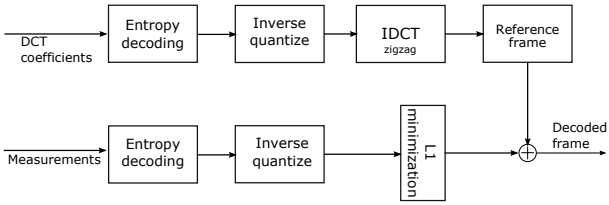

Fig. 2. decoder

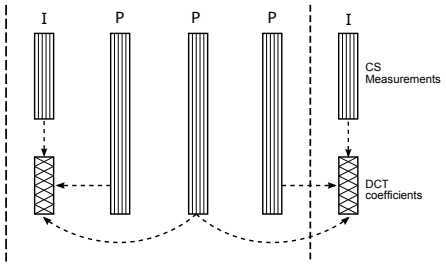

Fig. 3. GOP Structure. Vertical dashed lines contain a group of pictures of size four.

B. Components of Compressive Sensing

The structurally random matrix [7] is chosen as the measurement matrix which can be described as

$$\Phi = QWP, \tag{4}$$

where matrix $P \in R^{n\times n}$ is a global randomizer which randomly permutes matrix W. Matrix $Q \in R^{m\times n}$ randomly selects m rows of WP. $W \in R^{n\times n}$ is a block diagonal matrix.

In addition to the synthesis-based ℓ_1 minimization, an alternative analysis-based ℓ_1 minimization can be formulated as

$$\begin{aligned} \min \quad & \|\Psi^*\tilde{x}\|_{\ell_1} \\ \text{subject to} \quad & \|y-\Phi\tilde{x}\|_{\ell_2} \le \epsilon \end{aligned} \tag{5}$$

According to [5], when Ψ is an orthonormal basis, the above two problems are equivalent. But when Ψ is a redundant dictionary, analysis-based ℓ_1 minimization involves less unknowns so the recovery performance is superior.

Undecimated wavelet transform (UWT) is chosen as the basis Ψ. UWT has been found to outperform the orthogonal wavelet transform in image denoising. It is expected to enhance the sparsity of image.

C. Quantization

DCT coefficients and measurements are uniformly quantized. Let l be the number of bits to represent a measurement. $V_{\max} = \max(|y_i|)$. Range $[-V_{\max}, V_{\max}]$ is divided into 2^l bins. Then the measurement can be represented by the bin number which contains it. The quantization step size $q = 2V_{\max}/2^l$. DCT coefficients quantization follows the same way. The quantization error can be estimated by

$$\Delta = \sqrt{\frac{q^2 m}{12}} = \frac{V_{\max}\sqrt{m}}{\sqrt{3}\cdot 2^l}. \tag{6}$$

When there is no other noise except quantization error, $\epsilon = \Delta$.

Last, we adopt the simple entropy coding technology, the Huffman code, to further improve the code efficiency.

IV. Performance Study

In this section, we first investigate the reasoning behind the codec design and its performance. Then we compare the performance of the proposed SVCCS and the traditional solutions in supporting wireless multicast applications.

A. Performance of SVCCS

We use "Foreman" as the test video sequence to investigate the SVCCS performance. The resolution is QCIF(176×144), and frame rate is 15 frames per second. We use NESTA [1] which is one of a few routines that can solve analysis-based ℓ_1 minimization.

Figure 4 shows the compressibility of an original frame and a difference frame. A difference frame is obtained by subtracting two consecutive original frames. We use db4 UWT with four levels. Since UWT is redundant, the number of resultant transform coefficients is 16 times larger than the frame size. Then, we sort the magnitude of the coefficients in descending order. From the figure, we can see that not only the difference frame's energy is reduced, but also the coefficients of the difference frame decay faster and the portion of small coefficients is lager than that of the original frame.

Figure 5 shows the distribution of DCT coefficients and measurements. Since the DCT coefficients contain the majority

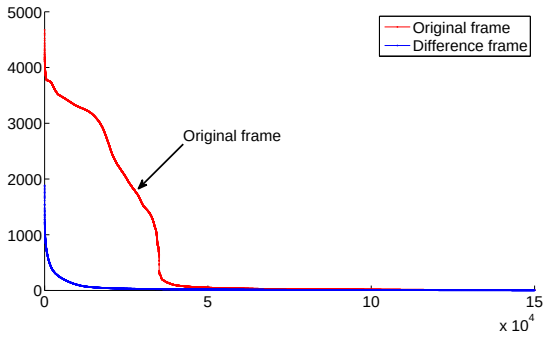

Fig. 4. Compressibility of an original and difference frame.

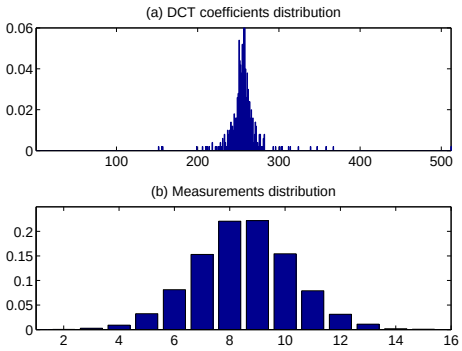

Fig. 5. Distribution of DCT coefficients and measurements

of the energy of an image and are more important, we use more bits (9 bits) to represent a DCT coefficient than that for a CS measurement (4 bits). There are 500 DCT coefficients and 12000 measurements. From the figure, we can see the distribution is skewed and close to a Gaussian distribution. Therefore, the coding efficiency can benefit from the Huffman coding. The average bits of DCT coefficients and measurement is reduced to 5.5 and 2.9 bits, respectively.

In order to study the contribution of each component of the video codec, we turn off the component or use an alternative to see the effect of the component. We compare analysis-based ℓ_1 minimization with total variation(TV) [13]; $9/7$ wavelet transform with UWT; intra-coding with inter-coding; and entropy coding enabled with entropy coding disabled. In this study, we set the number of DCT coefficients to be 500 with 9 bits for each coefficient before Huffman coding. We change the coding rate by changing the number of measurements but with fixed 4 bits for each measurement. We use the same GOP structure depicted in Fig. 3 for inter-coding.

From Fig. 6, we can see that inter-coding is better than intra-coding for both analysis-based ℓ_1 and TV minimization. PSNR is increased by 12% (3.43 dB) on average for analysis-based ℓ_1 with UWT case and inter-coding. If analysis-based ℓ_1 and inter-coding are enabled, the UWT is 11% (3.2 dB) better than biorthorgonal $9/7$ wavelet transform on average. Analysis-based ℓ_1 with inter-coding is 5% (1.6 dB) better than TV with inter-coding. When we use entropy coding, the bitrate is further reduced by 25% on average.

Next, we study the scalability of SVCCS. The effect of measurement loss is the same as changing the size of the measurement matrix. When the measurements are lost, we just

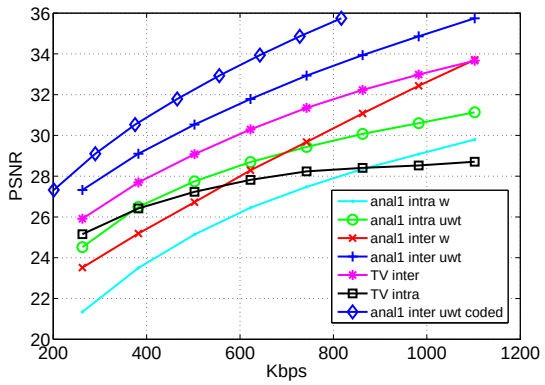

Fig. 6. Comparison of components. anal1 denotes analysis-based ℓ_1 minimization; inter denotes inter-coding with GOP structure IPPP; intra denotes all frames are intra-coded; u denotes biorthorgonal $9/7$ wavelet transform and uwt denotes UWT.

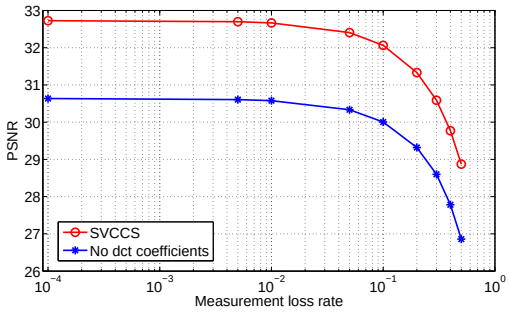

Fig. 7. PSNR vs. measurement loss rate

eliminate the corresponding rows in the measurement matrix and estimated the quantization error Δ again in the decoder side. For SVCCS, we compare our proposed GOP structure with that described in [12], i.e., the whole compressive sensed I frame is served as a reference frame. Assume that the DCT coefficients are received correctly. Fig. 7 shows the comparison. Since the reference frame in [12] cannot be recovered exactly the same as in the encoder, the decoded video quality is thus degraded. If the measurements of the reference frame are lost, the degradation is more apparent. Therefore, as shown in Fig. 7, the proposed GOP structure of SVCCS is more desirable.

There are several parameters that control the rate distortion performance. The number of DCT coefficients and its quantization level control the quality of the base layer, and the number of measurements and quantization level control the video quality of the enhancement layer. An optimization problem can be formulated to maximize the PSNR subject to the bit budget. For example, for each I frame, if we use more DCT coefficients, then we have to allocate less bits (with less quantization level) for each coefficient before entropy encoding. As shown in Table I, there is no optimal quantization level for all bit budget. In addition, we also need to determine how to optimize the allocation of bits between the base layer and the enhancement layer. These optimization problems beckon for further research.

bits(60 Kbit)	6 × 10000	5 × 12000	4 × 15000	3 × 20000
PSNR	33.75	**34.82**	34.78	33.29
bits(48 Kbit)	6 × 8000	5 × 9600	4 × 12000	3 × 16000
PSNR	31.90	32.96	**33.23**	32.19

TABLE I
MEASUREMENT AND BITS ALLOCATION

	Avg. frame size (Kbits)	Avg. PSNR (dB)
SVCCS	36.99	32.74
MJPEG	20.93	31.85

TABLE II
MEASUREMENT AND BITS ALLOCATION

B. Multicast with SVCCS

We study the performance of wireless multicast with SVCCS. We compare the convolutional code protected MJPEG bitstream and SVCCS. 50 frames are encoded with MJPEG and SVCCS, respectively. Table II lists the average frame size and average PSNR of all frames. The SVCCS encoded bitstream is obtained from the joint source and channel coding approach, so we do not apply channel coding. For MJPEG coded bitstream, we apply convolutional code (code rate is $1/2$). After channel coding, the doubled average frame size of MJPEG is even larger than that of SVCCS; thus, SVCCS can take less channel bandwidth.

We assume that the communication channel is AWGN and modulation scheme is DBPSK. Assume that the base layer of SVCCS can be correctly received. This assumption is acceptable as the base layer only counts for 2% of the coded bitstream, which can be protected with very low cost. The average PSNR of the base layer is 21.45 dB.

Fig. 8 shows the advantage of SVCCS which is strongly adaptive to channel conditions. Suppose that the channel quality of users varies from 7 to 9.2 dB. In the simulation, we set the packet size of the SVCCS coded bitstream to 250 bytes. If one bit is transmitted in error, the whole packet and the contained measurements are lost. For MJPEG coded bitstream, we do not drop any packet, as dropping the whole packet makes the decoded quality even worse. Admittedly, this is not the best way of error concealment, but the figure is sufficient to show the better scalability of SVCCS than FEC protected MJPEG.

When we evaluate the video quality at a particular receiver whose SNR is 7.6 dB, some of the frames of MJPEG cannot be decoded and the decoded one shows worse quality than SVCCS, as shown in figure 9.

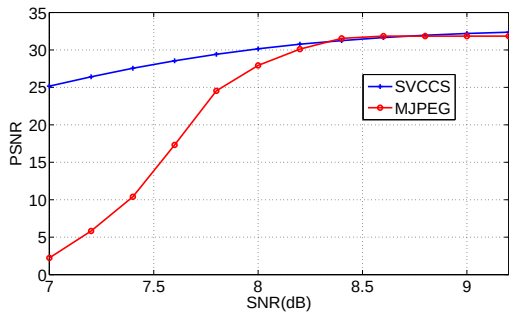

Fig. 8. PSNR vs. SNR

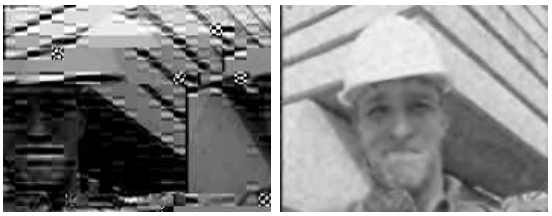
(a) MJPEG (b) SVCCS

Fig. 9. SVC vs. MJPEG

V. CONCLUSION

In this paper, we have proposed a low-complex, scalable video coding architecture based on compressive sensing for wireless unicast and multicast transmissions. As SVCCS can achieve good scalability, error resilience and coding efficiency, it is more effective and efficient than the traditional solution to support wireless videocast. CS based video coding is overall a promising direction with many other open issues that worth further investigation, e.g., for SVCCS, how to optimize the quantization level and the bit allocation for each layer; how to reduce the decoding complexity is an important further research issue; how to enhance network protocols to support SVCCS with even lower cost and better performance.

REFERENCES

[1] S. Becker, J. Bobin, and E.J. Candes. NESTA: A fast and accurate first-order method for sparse recovery. *Submitted. Available from arXiv*, 904, 2009.

[2] E.J. Candes, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[3] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, dec. 2006.

[4] E.J. Candes and M.B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, mar. 2008.

[5] E.J. Candes, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted l 1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[6] Z. Charbiwala, S. Chakraborty, S. Zahedi, Younghun Kim, M.B. Srivastava, Ting He, and C. Bisdikian. Compressive oversampling for robust data transmission in sensor networks. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9, mar. 2010.

[7] T.T. Do, T.D. Tran, and Lu Gan. Fast compressive sampling with structurally random matrices. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3369–3372, mar. 2008.

[8] D.L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, april 2006.

[9] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, Ting Sun, K.F. Kelly, and R.G. Baraniuk. Single-pixel imaging via compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):83 –91, mar. 2008.

[10] V.K. Goyal, A.K. Fletcher, and S. Rangan. Compressive sampling and lossy compression. *Signal Processing Magazine, IEEE*, 25(2):48–56, mar. 2008.

[11] J.N. Laska, P. Boufounos, M.A. Davenport, and R.G. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. *preprint*, 2009.

[12] S. Pudlewski, T. Melodia, and A. Prasanna. C-dmrc: Compressive distortion-minimizing rate control for wireless multimedia sensor networks. In *Sensor Mesh and Ad Hoc Communications and Networks (SECON), 2010 7th Annual IEEE Communications Society Conference on*, pages 1–9, jun. 2010.

[13] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[14] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the H. 264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, 2007.

[15] Michael B. Wakin, Jason N. Laska, Marco F. Duarte, Dror Baron, Shriram Sarvotham, Dharmpal Takhar, Kevin F. Kelly, and Richard G. Baraniuk. Compressive imaging for video representation and coding. In *In Proceedings of Picture Coding Symposium (PCS*, 2006.