A NOVEL IMAGE/VIDEO CODING METHOD BASED ON COMPRESSED SENSING THEORY

Yifu Zhang, Shunliang Mei

Department of Electronic Engineering,
Tsinghua University, Beijing.

Quqing Chen, Zhibo Chen

Thomson Corporate Research,
Beijing.

ABSTRACT

Compressed Sensing (CS) has been recently proposed for more efficient signal compression and recovery at theoretical level. This paper proposes a new image/video coding approach combining the CS theory into the traditional discrete cosine transform (DCT) based coding method to achieve better compression efficiency for spatially sparse signal. Furthermore, this new approach is integrated into JPEG and H.264/AVC coding framework as a new coding mode. Rate-distortion optimization is employed for adaptive selection between the new coding mode and the conventional coding modes. Experimental results demonstrated remarkable coding gain for different kinds of natural image/videos by the proposed method.

Index Terms— compressed sensing, discrete cosine transform, rate-distortion optimization, image/video coding

1. INTRODUCTION

Compressed sensing (CS) theory [1] is a newly proposed approach used in signal processing. By employing linear programming or other mathematical programming methods, it is able to recover originally spatially sparse signal from its incomplete frequency coefficients. Since a large quantity of image and video frames fit this criterion, increasing amount of research work has been done to incorporate the CS theory into this field.

Existing image/video coding technologies can be classified into two categories based on different transform types: wavelet based image/video compression and Discrete Cosine Transform (DCT) based hybrid image/video coding. For example, JPEG2000 is the representative standard using wavelet based structure, while JPEG, MPEG-1, MPEG-2, MPEG-4, H.261, H.263, H.264/AVC, and VC-1 all employ the DCT based hybrid image/video coding framework, which is more widely used.

However, all image/video coding methods mentioned above have the following limitation: if several frequency domain coefficients of the signal are removed, the reconstructed signals after inverse quantization and inverse transformation will never contain the exact information in the original signals, and distortions are introduced thereby. This is quite a normal rule in the existing compression theory.

Another problem, especially for DCT coding, is the poor performance at edges of objects. A single sharp edge in a block can be rendered with significant distortion pattern after coding. Fortunately, the CS recovery criterion is based on minimum sum of absolute pixel value criterion, and blocks containing simple edges can be better reconstructed in this approach.

Using CS theory in image coding has been studied theoretically in [2], [3] and [4]. Various experimental results of the methods above have also been given in [5]. For natural image coding, [6] has given a block based image compression method combining CS and wavelet transform. A comprehensive list of related works can be found at [7].

In our research, we focused on incorporating the CS theory with popular existing image/video coding schemes that employ DCT transform. Due to the characteristics of the DCT transform, the CS method can be embedded into the traditional coding and decoding system seamlessly, and much better reconstruction for blocks containing "edges" can also be achieved.

The rest of this paper is organized as follows. Section 2 is an overview of the CS theory. Based on this, our new coding method is proposed and illustrated in Section 3. Corresponding experimental results are given in Section 4. Finally, Section 5 concludes this topic.

2. COMPRESSED SENSING THEORY OVERVIEW

The Shannon/Nyquist sampling theorem specifies that to avoid losing information while sampling a signal, at least twice the signal bandwidth should be covered. However, the compressed sensing method, which is also called compressive sampling by some other authors, can tolerate sampling and representing spatially sparse signals at a rate significantly below the Nyquist rate. It employs non-adaptive linear projections that preserve the structure of the signal; the signal is then reconstructed from these projections using an optimization programming process [1].

For a piece of finite-length, real-valued 1-D discrete signal $\boldsymbol{x}$, its projection can be expressed as:

$$\boldsymbol{x} = \sum_{i=1}^{N} \boldsymbol{\psi}_i s_i = \boldsymbol{\Psi} \boldsymbol{s}, \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{s}$ are $N \times 1$ column vectors, and $\boldsymbol{\Psi}$ is an $N \times N$ basis matrix, with the vectors $\{\boldsymbol{\psi}_i\}$ $(i = 1, 2, \cdots, N)$ as columns. Clearly, if $\boldsymbol{\Psi}$ is full ranked, $\boldsymbol{x}$ and $\boldsymbol{s}$ are equivalent representations of the signal, $\boldsymbol{s}$ in the, for example, time or space domain and $\boldsymbol{x}$ in the $\boldsymbol{\Psi}$ domain. The signal $\boldsymbol{x}$ has a sparse representation if it is a linear combination of only K basis vectors. That is, only K coefficients of $\{s_i\}$ $(i = 1, 2, \cdots, N)$ in (1) are nonzero and the rest $(N - K)$ ones are zero. The case of interest is $K \ll N$. The signal $\boldsymbol{x}$ is compressible if the representation (1) has just a few large coefficients and many small coefficients.

Take M ($M < N$) linear, non-adaptive measurement of $\boldsymbol{x}$ through a linear invertible transform $\boldsymbol{\Psi}$, namely,

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{s} = \boldsymbol{\Theta}\boldsymbol{s}, \tag{2}$$

where $\boldsymbol{\Phi}$ is an $M \times N$ matrix, $M < N$, and its M rows can each be considered as a basis vector, usually orthogonal. $\boldsymbol{x}$ is thus transformed, or down sampled, to an $M \times 1$ vector $\boldsymbol{y}$.

Since $M < N$, the task of recovering $\boldsymbol{s}$ from $\boldsymbol{y}$ seems ill-conditioned. However, the additional assumption of the sparsity of $\boldsymbol{s}$ makes it possible and practical. The compressed sensing theory gives that when taking $M \geqslant cK \log N$ random measurements, where c is a small positive number affecting the probability of recovery, the signal $\boldsymbol{s}$ satisfying $\boldsymbol{y} = \boldsymbol{\Theta}\boldsymbol{s}$ can be exactly recovered under the minimum ℓ_1-norm reconstruction with high probability [8], i.e.:

$$\hat{\boldsymbol{s}} = \arg\min \|\boldsymbol{s}\|_1, \ \text{ s.t. } \boldsymbol{\Theta}\boldsymbol{s} = \boldsymbol{y}. \tag{3}$$

For real valued signal $\boldsymbol{s}$, the ℓ_1-norm means the sum of absolute value of its non-zero components. This is a convex optimization problem that can be conveniently reduced to a linear program known as basis pursuit [4].

3. PROPOSED IMAGE/VIDEO CODING METHOD

3.1. 2-D Signal Representation

For image and video coding, the signal is 2-D instead and few blocks can be defined sparse by the above criterion. However, blocks containing a sharp edge or several edges are more common. For these blocks, the ℓ_1-norm criterion can not be directly applied since a large percentage of pixels are not zero-valued. Instead, these blocks can be defined as gradient sparse, i.e. significant pixel value variations only occur at a few pixels. For an $n \times n$ block, the gradient can be defined as [5]:

$$D_{ij}\boldsymbol{s} = \begin{pmatrix} D_{h,ij}\boldsymbol{s} \\ D_{v,ij}\boldsymbol{s} \end{pmatrix}. \tag{4}$$

Its horizontal components is

$$D_{h,ij}\boldsymbol{s} = \begin{cases} s_{i+1,j} - s_{ij} & i < n \\ 0 & i = n \end{cases}, \tag{5}$$

and its vertical component is

$$D_{v,ij}\boldsymbol{s} = \begin{cases} s_{i,j+1} - s_{ij} & j < n \\ 0 & j = n \end{cases}. \tag{6}$$

Thus, the total variations of $\boldsymbol{s}$ is simply the sum of the magnitudes of this discrete gradient at every point:

$$\mathrm{TV}(\boldsymbol{s}) = \sum_{i,j} \sqrt{(D_{h,ij}\boldsymbol{s})^2 + (D_{v,ij}\boldsymbol{s})^2} = \sum_{i,j} \|D_{ij}\boldsymbol{s}\|_2 \tag{7}$$

With these definitions, the image recovery can be recast as a Second-Order Cone Programming (SOCP) problem [5]:

$$\hat{\boldsymbol{s}} = \arg\min \mathrm{TV}(\boldsymbol{s}), \ \text{ s.t. } \boldsymbol{\Theta}\boldsymbol{s} = \boldsymbol{y}. \tag{8}$$

Usually in popular image/video coding schemes, the transform is usually 2-D Discrete Cosine Transform, the 2-D representation of 1-D DCT, to accommodate the characteristics of 2-D signals. However, for the purpose of simplicity and accordance with equation (1), this paper proposes to employ 1-D DCT for the line-by-line scanned vector of the block rather than 2-D transform. In this case, $\boldsymbol{\Psi}$ is the 1-D DCT basis matrix and $\boldsymbol{\Phi}$ is an $M \times N$ matrix to reduce the signal's dimension.

3.2. The Quantization and Sampling Process

Different from the theoretical analysis and experiment of 2-D image CS application in [5], all practical image and video coding schemes inevitably adopt block coding, quantization, rounding, and other methods to reduce coding length and computational complexity. To make CS cooperate with existing coding schemes, corresponding modifications are needed.

For problems in the form of (8), the reconstruction performance increases as the dimension grows. However, the computational complexity of this problem is $O(n^3)$ [1]. Thus, in practical implementation of image and video coding, there exists a trade-off between complexity and coding performance. Dividing image into blocks is necessary, and the block size becomes an important parameter. In our practical implementation, the 8×8 block type is used for this compromise.

Another problem in practical coding is quantization. The coding process in theoretical demonstration uses double precision format data in every procedure, which ensures the complete recovery of the sparse signal. However, quantization and rounding are indispensable in compression and incomplete recovery is expected. Therefore, in the existence of quantization and rounding, the equality constraints $\boldsymbol{\Theta}\boldsymbol{s} = \boldsymbol{y}$ in (8) is no longer effective, and another SOCP problem with quadratic constraints is used:

$$\hat{\boldsymbol{s}} = \arg\min \mathrm{TV}(\boldsymbol{s}), \ \text{ s.t. } \|\boldsymbol{\Theta}\boldsymbol{s} - \boldsymbol{y}\| \leqslant \epsilon, \tag{9}$$

where ϵ is a constraint to allow a certain extent of distortion.

As stated above, for blocks with relatively sparse gradient, we can cut off unimportant frequency component representing the entire block, and get reconstruction with equal or even better quality. However, the authors in [1] proved that as long as the positions of maintained frequency components do not form a subgroup or coset of all frequency component positions (e.g., for discrete transforms, the set of all even number positions form a subgroup of the set of all positions, while the set of all odd number positions form a coset of the set of all positions), successful recovery is ensured. Therefore, with enough sampling rate, even random selection has almost no possibility of failure. In our proposal, to keep the most significant frequency components, we simply keep the first M ones out of a total of N coefficients and the rest are truncated.

With all these specifications, we have our new proposal of coding and reconstructing scheme, as shown in Fig. 1.

3.3. Rate-Distortion Optimization (RDO)

Though theoretically lossless under ideal condition, the above proposed CS based coding method still results in distortion due to quantization, rounding, coefficient truncation. For natural images, a great proportion of blocks are recorded with dense gradient distribution, rather than the sparse conditions. In this case, the CS method might not work very well. Consequently, with truncated number of frequency components, the CS method may frequently lose its rate-

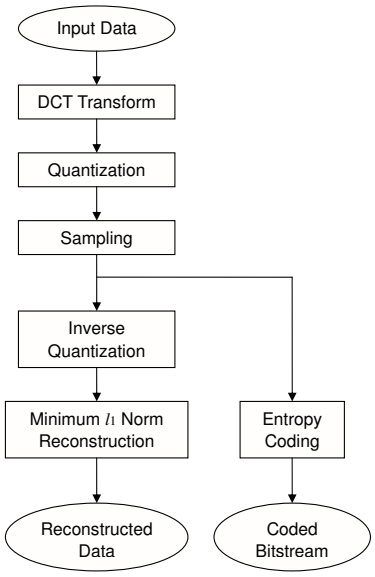


Fig. 1. The flow chart of the proposed coding scheme based on compressed sensing.

distortion (RD) performance to the DCT scheme, although the truncation (or sampling) process ensures a lower bit rate than that of DCT for the same block with the same quantization step.

Generally, the CS coding method is more suitable for the block with sparse gradients, while the DCT method fits complicated blocks better.

To take the advantage of both methods, this paper proposes to combine the above proposed CS method and the existing DCT based method together, by defining the CS based method as a new coding mode. Each 8×8 block is encoded and decoded in normal DCT coding mode or the proposed CS coding mode. RDO strategy [9] is introduced to adaptively select the mode with better RD performance. A flag bit is inserted into the bitstream of each 8×8 block to indicate the selected mode if this block contains any non-zero AC coefficients. Both modes employ the same existing coefficient entropy coding method in H.264/AVC.

3.4. Truncation Rate M

As proposed in 3.1 and 3.2, one block is first rearranged line by line to form a vector, and this vector is 1-D DCT transformed and truncated to maintain the first M coefficients. Since 1-D DCT coefficients are expected bigger at lower frequency and smaller at higher frequency in the sense of probability, as quantized more heavily, more coefficients turn zero from higher frequency downward. To keep the efficiency of the truncation process, we generally lower the value of M as the quantization step gets bigger.

The choice of truncation rate M for certain quantization step is still being investigated.

4. EXPERIMENTAL RESULTS

Our algorithm is integrated into the JM12.2 H.264/AVC codec for intra frame coding. MATLAB is used as computational engine, seamlessly combined with JM12.2, to solve the SOCP problem to reconstruct blocks. The open-source MATLAB code in [5] is used in our experiment. For simplicity, currently only direct current (DC) prediction is enabled and only fixed transform block size of 8×8 is used, since, as stated in 3.2, the reconstruction performance deteriorates as block size decreases, and 4×4 blocks are too small for linear programming to perform normally. Therefore, the 1-D DCT employed for the CS method is 64×1 dimensional, while the corresponding 2-D DCT is 8×8 dimensional.

Table 1. 4 pairs of parameters corresponding to QP values in our experiment

QP	22	27	32	37
Truncation rate M	40	32	26	20
Quantization step q	6	12	25	45

Table 2. Coding efficiency comparison result (average results of 4 QPs: 22, 27, 32, 37 according to the calculation method in [10])

Image or sequence	Bitrate reduction (%)	PSNR gain (dB)
rush_cif_1	-5.20	0.511
rush_cif_2	-2.44	0.139
camera_128x128	-5.96	0.427
tiger_720x480	-2.95	0.254
carphone_cif	-2.21	0.145
foreman_cif	-2.76	0.161
Average	**-3.59**	**0.272**

The proposed method is tested at four different quantization steps: for 2-D DCT coefficients QP 22, 27, 32, 37 are used, and the corresponding 1-D DCT coefficients for CS are quantized at four states listed in Table 1.

The relationship between QP and quantization step q in H.264/AVC can be approximately described as:

$$q = \frac{5}{2} \times 2^{\frac{\mathrm{QP}-12}{6}}. \tag{10}$$

As shown in the table above, we slightly decreased the quantization step for each state since the 1-D DCT coefficients still need to be truncated. Making them larger can keep the RDO method more balanced.

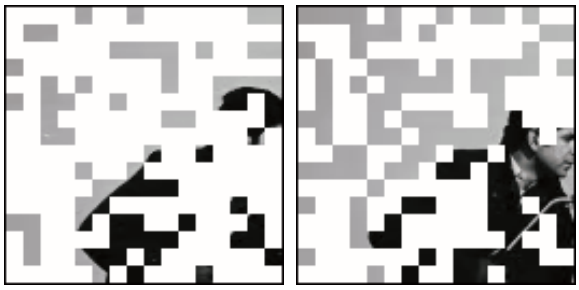
In our experiment, we tested on image frames, intra frames for different video sequences, with traditional DCT method and our proposed compressive sensing method, respectively. The bitrate reduction and PSNR gain for each frame, calculated according to the criteria in [10], are shown in Table 2.

The first 4 rows in Table 2 correspond to the 4 frames in Fig. 2. For images and video frames containing fierce impulses, e.g. many stars in a night sky background, as shown in Fig. 2(a) and 2(b), or intense lines in relatively smooth background, e.g. some photo images (Fig. 2(c)) and most cartoon video sequences, as shown in Fig. 2(d), this RDO method can generally have notable coding gain. Common sequences like *carphone* and *foreman* have lower coding gain for relatively lacking this characteristic.

(a) *rush* 1 (b) *rush* 2

(c) *camera* (part) (d) *tiger*

Fig. 2. Frames containing impulses (2(a) and 2(b), from sequence *rush*) or intense edges (2(c)), and a cartoon sequence frame (2(d)).



(a) Blocks using CS method (b) Blocks using DCT method

Fig. 3. Blocks employed compressive sensing (3(a)) vs. blocks employed traditional DCT (3(b)) in the image *camera* in Fig. 2(c) compressed with RDO method.

An example of the effectiveness of the proposed RDO method is shown as follows.

Taking part of image *camera* in Fig. 2 for example, Fig. 3 is an illustration of this improvement. Non-white blocks in Fig. 3(a) are coded with our new CS method in the reconstruction image using RDO, while non-white blocks in Fig. 3(b) employ traditional DCT method. It should be noticed that if a white block in Fig. 3(a) is also white in Fig. 3(b) in the same position, this block is coded in "bypass" mode, i.e., this block only contains DC value, and neither CS nor DCT is employed for this block. Clearly, simple blocks with relatively sparse gradient distribution, especially those with intense edges, e.g. the block containing the brighter point in the background, and blocks containing edges of the photographer's upper arm, shoulder, and hair, tend to be assigned with compressive sensing coding, while complicated blocks, e.g. the photographer's face and neck, have a tendency to enable traditional DCT coding.

In our experiment, only images and intra predicted frames are tested and analyzed. Further experiments using more intra-prediction modes, inter predicted frame coding, and variable block sizes will be carried out in the future. Adaptive quantization step q and truncation rate M selection will also be studied.

Finally, though block-based method is used to balance the efficiency, the encoding complexity remains high for the SOCP process in reconstruction. Fast mode selection method between normal DCT coding mode and the CS mode will be studied to reduce encoding complexity in the next step. Also, fast solution to the optimization problems in this paper can be studied to speed up the reconstruction process.

5. CONCLUSION

This paper investigated into the CS theory in signal processing and incorporated it into the existing image/video coding framework. The new approach has been tested both theoretically and on the practical H.264 coding platform, and significant coding gain has been achieved. The CS theory itself can be used in many fields other than image/video compression, and the proposed adaptation method between CS coding and normal coding can be generalized to other signal processing application, e.g., audio coding.

6. REFERENCES

[1] D. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[2] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. on Information Theory*, December 2005.

[3] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. on Information Theory*, vol. 52, no. 12, pp. 5406–5425, December 2006.

[4] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Information Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[5] E. Candès and J. Romberg, "L1-MAGIC: Recovery of sparse signals via convex programming," 2005.

[6] Lu Gan, "Block compressed sensing of natural images," *Proc. Int. Conf. on Digital Signal Processing*, July 2007.

[7] "http://www.dsp.ece.rice.edu/cs/," .

[8] R. G. Baranniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, July 2007.

[9] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 115, no. 6, pp. 74–90, November 1998.

[10] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," *ITU-T Q.6/SG16 VCEG, VCEG-M33*, April 2001.