A COMPRESSIVE SENSING APPROACH TO OBJECT-BASED SURVEILLANCE VIDEO CODING

Divya Venkatraman and Anamitra Makur
School of Electrical and Electronics Engineering
Nanyang Technological University, Singapore

ABSTRACT

This paper studies the feasibility and investigates various choices in the application of compressive sensing (CS) to object-based surveillance video coding. The residual object error of a video frame is a sparse signal and CS, which aims to represent information of a sparse signal by random measurements, is considered for coding of object error. This work proposes several techniques using two approaches- direct CS and transform-based CS. The techniques are studied and analyzed by varying the different trade-off parameters such as the measurement index, quantization levels etc. Finally we recommend an optimal scheme for a range of bitrates. Experimental results with comparative bitrates-vs-PSNR graphs for the different techniques are presented[1]

Index Terms - Surveillance video, Object-based coding, Compressive sensing

1. INTRODUCTION

Compressive sensing is an emerging field which aims to measure sparse and compressible signals at close to their intrinsic information rate than what is considered necessary according to Nyquist criterion. In this work, we have presented an application of compressive sensing for coding of indoor surveillance video. A major aspect of any surveillance system is to efficiently compress the long hours of video to facilitate archival or networking. Identification of objects in motion through segmentation is essential for surveillance and in such a video codec, an arbitrary shaped moving object is described using two features: texture and shape. [1][2] are some previous work on MPEG-4 object based encoding for surveillance video and [6] describes the framework on which this work is based on. The aim of this work is to explore compressive sensing on motion-compensated object error (Eq 5), i.e. random measurements on the residual object error, to represent the texture information for arbitrary-shaped objects. Any signal x in $\Re^N$ can be represented in terms of orthogonal basis vectors $\{\Psi_i\}_{i=1}^N$ of dimensions $N \times 1$ as

$$x = \sum_{i=1}^{N} s_i \Psi_i \text{ or } x = \Psi S,\ s_i = \langle x, \Psi_i \rangle = \Psi_i{}^T x \quad (1)$$

where $\Psi := [\Psi_1 | \Psi_2 | \Psi_3 | ... \Psi_N]$ is obtained by stacking the basic vectors as columns, and S is a column vector of weighting coefficient. x and S are equivalent representations of the same signal in the time domain and Ψ domain respectively. Either the signal x is sparse in time domain or there exists a 'compressible' domain Ψ in which the signal can be approximated by K large coefficients. It is based on the revelation that a small group of non-adaptive linear projections of a sparse signal contains enough information for reconstruction and processing [3].

$$y = \Phi x = \Phi \Psi S \quad (2)$$

Φ is a $M \times N$ measurement matrix where $M < N$. Thus CS requires a *stable measurement matrix* Φ that ensures the important features of the sparse signal are not damaged by the dimensionality reduction from $x \in \Re^N$ to $y \in \Re^M$. The Φ matrix is mostly chosen as a random matrix in which the matrix elements are drawn from independent and identically distributed variables of zero mean [4]. Different *reconstruction algorithms* are developed [3][4] to recover the original signal x from the random measurements y, one of them being the stage-wise orthogonal matching pursuit algorithm [5], which is used in our work, since it is fast with little sacrifice in performance.

The paper is organized as follows: section 2 analyses the characteristics of residual object error, section 3 describes the proposed CS framework with the different schemes in detail with experimental results, section 4 highlights the comparisons between the schemes, and section 5 concludes the paper.

2. ANALYSIS OF RESIDUAL OBJECT ERROR

In our surveillance video coding, shadow-less object segmentation is achieved using frame ratio pixels, edge maps and morphological operation based correction [7]. It is followed by an object-based motion estimation using the sum of squared differences (SSD) [6],

$$u_j^{ssd} = \arg \min_{u \in W} \sum M_j [\tilde{f}_{n-1}(u + x_j) - f_n(x_j)]^2 \quad (3)$$

where, j is the object count in a frame, u_j^{ssd} is the motion vector for the j^{th} object, M_j is the binary object mask obtained from the segmentation, x_j is the location of the object in the n^{th} frame, $\tilde{f}_{n-1}$ is the previously reconstructed frame, f_n is the current frame and W is the search window. The motion vector thus obtained is used to reconstruct the current frame from the previously reconstructed frame by the object-based motion compensation

$$\hat{f}_n = I_{bg} \Pi_j \bar{M}_j + \sum_j M_j [\tilde{f}_{n-1}(u_j^{ssd} + x_j)] \quad (4)$$

$\hat{f}_n$ is the current reconstructed frame and $\bar{M}_j$ is the binary complement of the object mask M_j. The background image I_{bg}, assumed to be known *a priori,* is substituted in the reconstructed image in places other than the object mask. The motion compensated object error for each object j in n^{th} frame

$$\Delta_j = M_j (f_n - \hat{f}_n) \quad (5)$$

[1] This work is supported under I2R-NTU joint R&D (2) Project

needs to be coded along with the motion vector, in order for a complete reconstruction of the frame at the decoder.

The two distinct characteristics of the object error are *inherent sparsity* and *arbitrary shape.* The object error (Eq 5) has sparse representation i.e. percentage of significant values in the array is very less. An accurate motion-compensated object (precise motion vector) will have object error with significant values only along the boundary of the object. Thus the location of the significant values depends on the efficiency of the motion-compensation algorithm.

3. PROPOSED CS BASED VIDEO CODING

In this section we describe the few variations of the proposed CS based video coding framework. A diagrammatic representation of the proposed general framework spanning these variations is shown in Fig 1. The object error (Δ_j) may or may not go through a transform stage. When the transform is present, some of the transform coefficients may be quantized and transmitted separately, giving rise to a hybrid coder. The object error or the (remaining) transform coefficients are vectorized from 2D to a 1D-array for further processing.

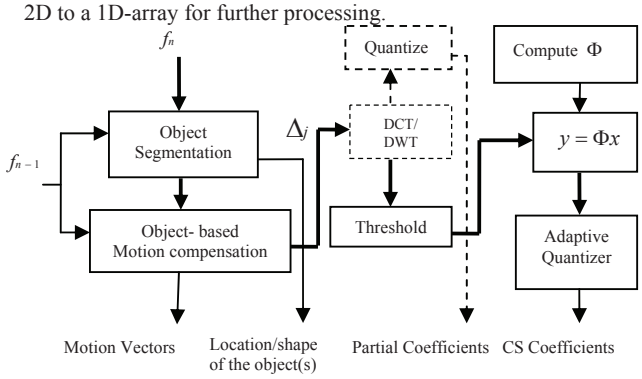

Fig1: Proposed CS framework for surveillance video encoder

It is shown in previous works of Donoho and Tanner [6] that the number of measurements $M \sim 2K\log(N/K)$ is a sharp threshold for successful reconstruction, where N is the length of the signal and K is the number of significant coefficients. It was observed that the sparsity ratio (K/N) of object error array varies across video frames in a sequence and also between objects in a single frame and hence in this work, the number of measurements is calculated as $M = \max(rK, rK\log(N/K))$ where r is termed as *measurement index*. The random measurement vector y is observed to have Gaussian density with the distribution peaking at mean zero because of the sparse nature of the object error signal, and having different variances for varying object sizes within a frame and across the frame sequence.

In a fixed camera indoor gray-level surveillance video the background frame is constant except for occasional illumination changes. Results obtained in this work are using video sequence obtained from CAVIER [12] averaged over 100 frames and with multiple objects in a frame. The knowledge of the background frame is known *a priori* (frame with no objects is used). For robust scenario, the background may be statistically modeled using algorithms such as [9] and shared between the encoder and decoder periodically. Also the dimension of each object is approximately 2% of the total frame area. Since the PSNR from the background area is not relevant in surveillance video compression performance, in this work, we study the *object-PSNR* of the segmented mask of the object instead of the frame PSNR.

The CS measurements may be quantized using uniform or non-uniform quantizer. Table 1 compares the bitrate-vs-object-PSNR for uniform quantizer (Δ =8) and optimum Lloyd-Max non-uniform Gaussian quantizer [10] for quantization level L=8.

r	Uniform Q		Lloyd-Max Q (L=8)	
	Bitrate	PSNR	Bitrate	PSNR
1.5	6590	27.91dB	5503	26.87dB
2	7692	33.27dB	7053	31.23dB
2.5	8858	36.37dB	8617	34.49dB
3	9981	37.91dB	10165	36.64dB

Table 1: Uniform Q vs Lloyd-Max Q for L=8

Arithmetic coding performed after uniform quantization exploits the probability distribution and hence gives better coding performance. But given the non-stationarity of surveillance video, a non-uniform Gaussian adaptive quantizer, adapting to the variance of the distribution, is likely to be more robust, hence it is used in our experiments. The following are the transmitted parameters for each video frame in a CS framework:

1. Motion vectors for different objects in the video frame
2. Location (top left coordinate and size of the bounding box) and shape (chain coded boundary through entropy coding) of the objects
3. Dimensions of the measurement matrix for each object
4. Compressive sensing coefficients (and transform coefficients)
5. Variance of the CS measurements for adaptive quantizer

3.1. Direct CS coding of object error

To the object error, a lower threshold is applied in order to construct a sparse matrix by eliminating insignificant values and converting them to zero. This sparse signal is subjected to CS.

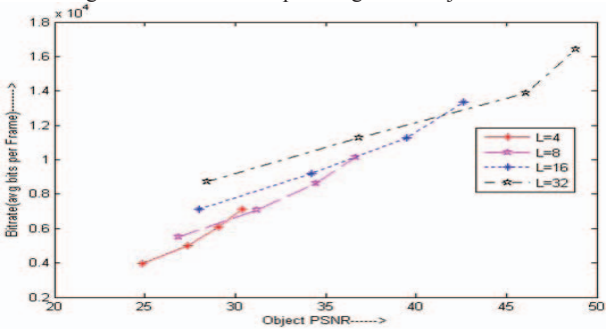

Fig 2. Bitrate vs Object-PSNR for direct CS for different L (different points on each L line correspond to r=1.5,2,2.5,3)

The measurement index r and the numbers of quantization levels L are varied and the bitrate-vs-object-PSNR curve is shown in Fig 2. It can be observed from the graph that the R-D curve for a fixed L and increasing r give a better object PSNR for a corresponding bitrate rather than keeping the measurement index r a constant and increasing just the number of quantization levels. Changing both r and L is necessary to ensure best PSNR over different bitrates, however it can be observed that r around 2.5 gives optimal performance.

3.2. DCT based CS coding of object error

Discrete cosine transform is applied on the object error and the coefficients are thresholded to construct a sparse matrix. The threshold *Th* used is a percentage of the absolute maximum value of the transform coefficients. Two different types of DCT are

applied on the object error. Then CS is applied on the thresholded DCT coefficients.

Whole-object DCT CS: Since the object size in the frame is small, DCT of the whole error block is taken.

Block-wise DCT CS: The object error is divided into 8x8 sub blocks, and smaller sub blocks at the boundary if required and separate DCT is applied for each block.

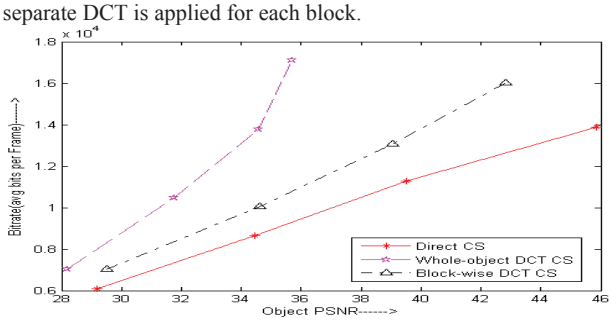

Fig 3: Bitrate vs object-PSNR for direct CS, whole-object DCT CS and block-wise DCT CS for r=2.5, Th=10% (points: L=2,4,16,32)

Fig 3, compares the DCT CS coders with the direct CS coder for r=2.5 (optimal rate mentioned in 3.1) and varying L. We observe that the DCT CS coders peform worse than the direct CS. The block-wise DCT CS is better than the whole-object DCT CS scheme since the spatial correlation of the object error is better exploited in non-boundary subblocks.

3.3. Wavelet based CS coding of object error

In this subsection we investigate a few variations of wavelet based CS coding scheme.

Regular wavelet CS: Discrete wavelet transform is applied on the object error. Daubechies 9/7 biorthogonal wavelet is used in our experiments since it has linear phase and is found to perform well in image compression. Two levels of wavelet decomposition are used. A threshold is then applied on the wavelet transform coefficients to create a single sparse array by merging all bands and CS is applied. The number of CS measurements M is obtained from the N, K values of the object, varying measurement index r.

Multiscale wavelet CS: For better representation of the wavelet coefficients, wavelet bands at different levels are no longer merged but treated separately in a fashion similar to [11]. The scaling coefficients are uniformly quantized using a step size Δ (and entropy coded) and transmitted separately. The number of CS measurements is determined as a ratio α of the size of the object error $M = \alpha N$. These CS measurements are distributed between wavelet coefficients of different levels of decomposition. Since two levels are used, 1/4th and 3/4th of the measurements are used for the second level and first level wavelet coefficients respectively. A threshold is used before CS measurements.

Hybrid wavelet CS: In this technique the scaling coefficients are separately quantized and transmitted as in the multiscale wavelet CS. However, independent assignment of CS measurements at different levels of wavelet decomposition is avoided. This is because, it is better to compress a sparse signal as a whole rather than in parts. Therefore, all wavelet coefficients (except the scaling coefficients) are merged together before thresholding and CS measurements.

Performance of wavelet schemes: Fig 4 shows the results of regular wavelet CS obtained by varying r and L for Th = 10%. It is observed that similar to direct CS, r=2.5 gives optimal R-D characteristics.

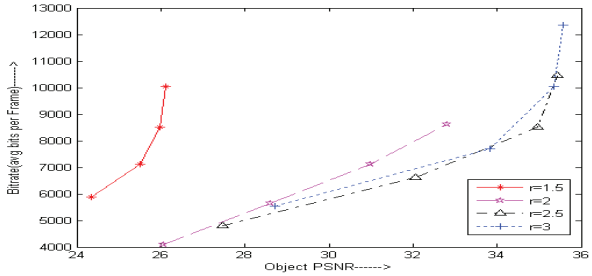

Fig 4: Bitrate vs object-PSNR for regular wavelet CS for varying r, Th=10% (points: L= 4,8,16,32)

Fig 5 shows the R-D performance of the multiscale wavelet CS coder for varying values of measurement ratio α and quantization levels L. It is observed that α =1 is a good choice for the bitrate of 6000 to 15000 bits per frame.

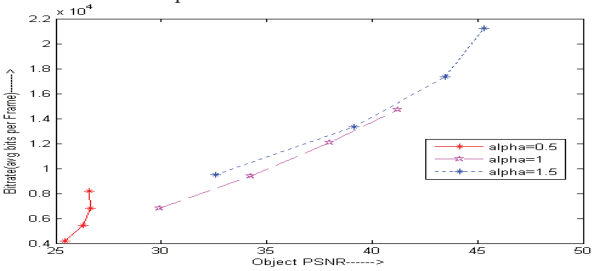

Fig 5: Bitrate vs object-PSNR for multiscale wavelet CS for varying α with Δ =8, Th=10%. (points: L= 4,8,16,32)

The hybrid wavelet CS coder is observed to perform better than the multiscale wavelet CS at all bitrates and the regular wavelet CS at higher bitrates. Consequently, we investigate the hybrid wavelet CS coder in more detail in the rest of this section. For the hybrid wavelet CS, experiments are performed by varying the quantization step size Δ for the scaling coefficients.

L	Δ = 8		Δ = 4		Δ = 2	
	Bitrate	PSNR	BitRate	PSNR	BitRate	PSNR
16	9214	36.48dB	9423	36.42dB	9692	36.57dB
32	11183	37.18dB	11370	37.25dB	11636	37.27dB

Table 2: Bitrate vs object PSNR for varying Δ value in hybrid wavelet CS for r=2.5, Th=10%

It was noticed that Δ =8 is a good choice as seen from Table 2. This is because the object error is high pass and hence the scaling coefficients have less energy and a large Δ may be used.

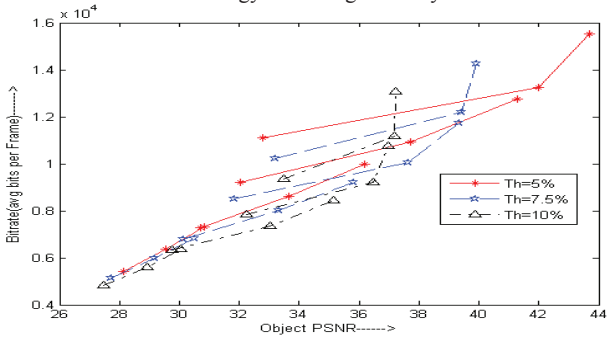

Fig 6: Bitrate vs object-PSNR for hybrid wavelet CS for varying Th with Δ =8 (lines: L=4,8,16,32 , points r=*2,2.5,3*)

Next, R-D performance of the hybrid wavelet CS is experimented by keeping Δ =8 and varying r, L, and threshold Th (=5%,7.5%,

10% of the absolute maximum value). Fig 6 summarises the results. It may be observed that, as before, to obtain the best R-D performance across different bitrates, it is necessary to change all the parameters *r*, *L* and *Th*- smaller *r*(=2,2.5), smaller *L*(=4) and larger *Th*(=10%) are better at low bitrates (≤ 8000 bits per frame), however larger *r*(=3), larger *L*(=16,32) and smaller *Th*(=5%) is preferable at higher bitrates (≥ 10000 bits per frame).

4. COMPARITIVE STUDY

In this section we compare the results of different CS techniques explained in detail previously and recommend an optimal operating technique amongst them. The comparison graph shown in Fig 7 plots the optimal operation curves for the different techniques adapting parameter values for *r*, *L*, Δ and *Th*.

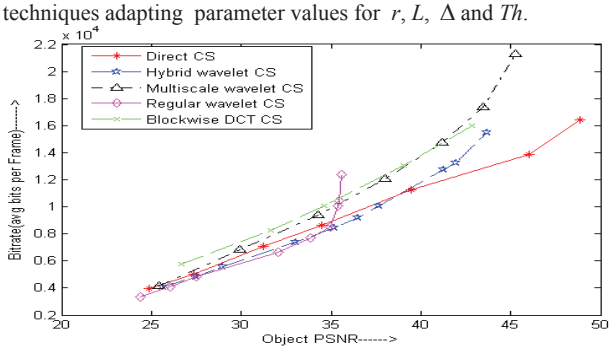

Fig 7: Comparitive bitrate vs PSNR between the techniques

Since the wavelet transform gives better energy compaction than DCT, it is observed that CS works better with DWT. At low bitrates, the hybrid wavelet CS and regular wavelet CS works better than direct CS as the wavelet transform creates a sparser array for CS than the original object error. Regular wavelet CS performs well initially but becomes worse than the hybrid scheme after 35 dB PSNR. This is because for larger CS measurements, the hybrid technique represents the wavelet coefficients (high pass) of object error (which is inherently highpass) better, and scaling coefficients are transmitted separately. At low bitrate, scaling coefficients are also sparse and hence well presented by regular wavelet CS. Hybrid scheme performs better that multiscale wavelet as the CS measurements are applied to all the wavelet bands merged together unlike the multiscale CS. Beyond 10k bits/frame direct CS perfoms better than other schemes. In conclusion, wavelet CS (hybrid scheme, since it is close or better than the regular wavelet) is best at low rates, while direct CS is best at higher rates.

From Fig 8, we observe that the bitrate variation is moderate for both the schemes except after frame 90 (when the two objects of the video merge into one object), where the direct CS has lesser variation. The typical PSNR variation is 29.87dB to 40.74dB for direct CS and 27.4 dB to 39.5dB for hybrid wavelet CS. Fig 9 gives an example of reconstructed frame.

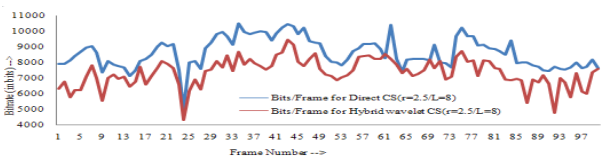

Fig 8: Framewise bitrate for direct CS and hybrid wavelet CS for *r*=2.5, *L*=8, Δ =8, *Th*=10% for 100 frames of "walk" video

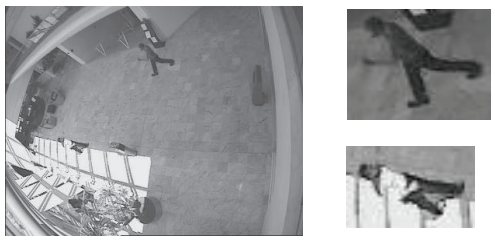

Fig 9: Reconstructed frame and objects using direct CS *r*=2.5, *L*=8

5. CONCLUSIONS

An application of CS for surveillance video coding is discussed using different techniques such as direct CS and DCT/DWT based CS and compared. The hybrid wavelet CS is found to work better at lower bitrates and direct CS for higher bitrates. Different parameters such as measurement index *r*, quantization levels *L*, intervals of Lloyd-Max quantizer (adapted based on variance), step-size of scaling coeffcients Δ, threshold on wavelet coefficients *Th* are varied for robust performance of the system. In this work we target coding only the texture of the object error with shape explicitly coded using chain code. In future, we aim to revive the object shape using implicit shape coding at the decoder.

6. REFERENCES

[1] A. Vetro, T. Haga, K. Sumi, and H. Sun, "Object-based coding for long-term archive of surveillance video," in *IEEE Intl. Conf. on Multimedia & Expo*, July 2003.

[2] C. Kim and J-N. Hwang, "Object-based video abstraction for video surveillance system," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 11281138, Dec. 2002.

[3] R. Baraniuk, "Compressive Sensing", IEEE Signal Processing Magazine, vol 24, no.4, pp.118-120, July 2007.

[4] D.L. Donoho, "Compressed sensing," IEEE Trans. Information Theory, vol. 52, no.4, pp. 1289–1306, Apr. 2006.

[5] D.L. Donoho, Y. Tsaig, I. Drori and J. Starck, "Sparse solution of underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit, Preprint 2007.

[6] R.V. Babu and A. Makur, "Object-based Surveillance Video Compression using Foreground Motion Compensation", 9th Intl Conf Control Automation Robotics Vision, (ICARCV), pp. 458-463, Dec. 2006.

[7] D. Venkatraman and A. Makur, "Shadow-less Segmentation of Moving Humans from Surveillance video", 10th Intl Conf Control Automtn. Robotics Vision (ICARCV), pp. 1317-1322, Dec 2008.

[8] D.L. Donoho and J. Tanner, "Counting faces of randomly-projected polytypes when the projection radically lowers dimension," submitted for publication.

[9] C. Stauffer and W. E. L. Grimson, ''Adaptive background mixture models for real-time tracking'', IEEE Intl. Conf. on Computer Vision and Pattern Recognition, pp. 246-252, 1999.

[10] S.P. Lloyd, "Least Squares Quantization in PCM," IEEE Trans. on Information Theory, pp. 129–137, Vol. IT-28, 1982.

[11] Y. Tsaig and D.L. Donoho, "Extensions of compressed sensing", IEEE Signal Processing Magazine, vol 86, no. 3, pp. 549-571, March 2006.

[12] "Caviar: Context aware vision using image-based active recognition," http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.