Spatiotemporal Compressed Sensing for Video Compression

Tao Xiong*, John Rattray*, Jie Zhang†, Chetan Singh Thakur*, Sang Peter Chin*‡
Trac D. Tran* and Ralph Etienne-Cummings*
*Department of Electrical and Computer Engineering, The Johns Hopkins University
Baltimore, MD, USA, Email: tao.xiong@jhu.edu
†Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
‡Department of Computer Science, Boston University, Boston, MA, USA

***Abstract*—We present a hardware-friendly spatiotemporal compressed sensing framework for video compression. The spatiotemporal compressed sensing incorporates random sampling in both spatial and temporal domain to encode the video scene into a single coded image. During decoding, the video is reconstructed using dictionary learning and sparse recovery. The evaluation results demonstrate the proposed approach can achieve high compression rate $(10 : 1 - 30 : 1)$ and robustness reconstruction quality $(> 20dB)$ on noisy database. Additionally, it also enables power efficient and real-time CMOS implementation $(0.7$ nJ/pixel).**

I. Introduction

As advancements in integrated circuit fabrication steadily continue yielding smaller process sizes and higher circuit densities, new low power methods of data compression are needed in applications where high data rates are ubiquitous. One such application is mobile video transmission where raw data rates, which can range from 2 Mb/s to 2 Gb/s, are too large for conventional transmission methods. Past compression methods that demonstrate high compression factors require complex circuitry and are computationally demanding, neither of which are suitable for low power, resource limited IC design. To satisfy the power and computational requirements for these mobile applications, our approach combines the processes of data acquisition and compression while reducing computational complexity and processing time as compared to methods that treat them as disjoint. Biomedical applications such as functional imaging [1] and calcium imaging [2] can benefit from such a energy efficient and real-time imaging acquisition system.

MPEG standard is one of the most heavily used methods for video compression, which serves as the core of many DVD formats and digital television broadcasts. This lossy form of compression achieves high compression rate (CR) that vary depending on the version of encoding. The high level of compression achieved by MPEG is a result of multiple operations (i.e. motion estimation) taken on the video data in both the spatial and temporal domain, however, the most effective steps in the process have the caveat of also being the most computationally intensive and time dependent, which adds the difficulty of efficient hardware implementation.

One of the computational intensive operations is information reduction through video motion estimation. This operation first computes motion vectors for blocks of pixels between frames. But computing the motion vectors can require an entire frame search for similar blocks which is a time intensive process. Thus the search is typically reduced to a percentage of the frame to increase processing speed. But this may leads to unsuccessful searches. State of the art solutions to computing these motion vectors without an entire frame search have arisen but at the cost of computational complexity. An integral step in compressing the data is taking the discrete cosine transform (DCT) of pixel blocks. Taking advantage of the DCT by removing the insignificant high frequency components of individual frames yields high amounts of compression, however, it is computationally expensive and increases quadratically with the pixel block size [3]. Reducing this size of the pixel blocks has the effect of reducing the time and processing power needed for the DCT but also increases the amount of motion vectors needed to encode the data and increases the compression schemes susceptibility to noise.

Inspired by the theory of Compressed Sensing (CS), a number of temporal CS systems has been proposed to alleviate the intensive computation and enable hardware-friendly implementation for video compression. For example, Llull demonstrated a coded aperture compressive temporal imaging system, which is able to reconstruct 148 frames per coded image [4], [5]. Koller et al. also showed a prototype compressive video camera at 740 fps using CMOS sensors and silicon-dioxide optical coded mask [6]. Tsai extended this technique to compress a multi-spectral, high-speed scene into a monochrome scene using objective lens, coded aperture, piezoelectric stage and monochrome CCD camera [7]. Finally, Liu proposed an efficient space-time sampling approach with pixel-wise coded exposure, which uses a prototype liquid-crystal-on-silicon device to modulate light prior to the image sensor [8], [9].

In order to meet the requirements of high compression rate and power efficient implementation, we proposed a hardware-friendly spatiotemporal compressed sensing framework for video compression with following contributions.

a) Spatiotemporal Compressed Sampling: The spatiotemporal compressed sensing is the key component of our proposed framework, as shown in Fig. 1. Unlike conventional video compression technique, the spatiotemporal compressed sensing requires no motion estimation, compensation and DCT to reduce bits by identifying and eliminating statistical redundancy. In spatiotemporal compressed sensing, the pixels are simply exposed through a random short "single on" exposure

978-1-5090-6389-5/17/$31.00 ©2017 IEEE

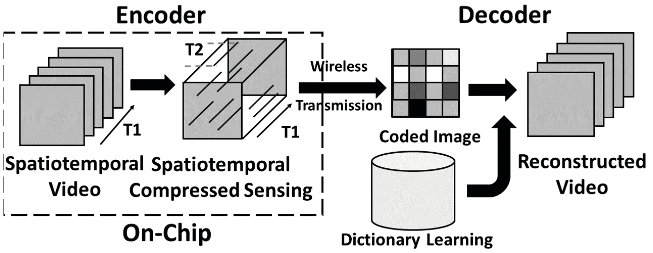

Fig. 1. The block diagram of spatiotemporal compressed sensing framework

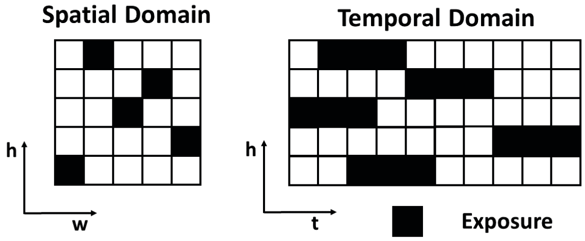

Fig. 2. The demonstration of pixel-wise coded exposure in spatial and temporal domains

of fixed duration, which essentially compresses a spatiotemporal video into a single coded image. Additionally, as shown in Fig. 2 the spatiotemporal compressed sensing is more optimal for video compression because it samples both the spatial and temporal information simultaneously, which sets it apart from the conventional spatial compressed sampling.

b) Hardware-friendly On-Chip Encoding for CMOS Architecture: In spatiotemporal compressed sensing, a sensing cube is adopted to encode the spatiotemporal video into a single coded image. In details, the sensing cube is composed of either 1 or 0, where 1 intuitively indicates exposure is turned on and vice versa as shown in Fig. 2. Therefore, the encoding for video compression can be formulated as a simple addition operation compared to conventional technique, which suffers from the intensive computation. Taking advantage of the simple arithmetic computation, the spatiotemporal compressed sensing is hardware-friendly and enables the real-time and power-efficient implementation on CMOS architecture [10].

Our paper is organized in the following structure: In section II, we introduce compressed sensing theory and demonstrate the spatiotemperal compressed sensing approach with corresponding hardware architecture. In section III, we demonstrate the experiments results and discuss the advantages and limitations of proposed approach for video compression compared to other video compression techniques such as MPEG. Finally, we conclude the paper in section IV.

II. Method

A. Compressed Sensing

Compressed Sensing (CS) theory [11], [12] demonstrates that a S-sparse signal $\mathbf{x} \in \mathbb{R}^N$ is essentially compressed into a measurement $\mathbf{y} \in \mathbb{R}^M$ by a sensing matrix $\mathbf{S} \in \mathbb{R}^{M \times N}$, where normally $S << M < N$. Given the Restricted Isometry Property (RIP) and $M \sim Slog\frac{N}{S}$ satisfied, the signal $\mathbf{x}$ can be exactly reconstructed by solving the optimization problem below.

$$\min_{\mathbf{x}} ||\mathbf{x}||_1 \ s.t. \ \mathbf{y} = \mathbf{S}\mathbf{x}.$$

However, the signal of image or video scene $\mathbf{x}$ is not sparse with respect to time or frequency domain. An over-complete dictionary $\mathbf{D} \in \mathbb{R}^{N \times L}$ needs to be adopted to sparsify the signal $\mathbf{x}$, where $\mathbf{x} = \mathbf{D}\mathbf{a}$ and $\mathbf{a} \in \mathbb{R}^L$ is sparse. Therefore, the reconstruction problem can be formulated as:

$$\min_{\mathbf{a}} ||\mathbf{a}||_1 \ s.t. \ \mathbf{y} = \mathbf{S}\mathbf{D}\mathbf{a}.$$

The reconstruction problem can be solved by using $\ell - 1$ norm optimization and then the signal is recovered as $\hat{\mathbf{x}} = \mathbf{D}\mathbf{a}$ at compression rate $\frac{M}{N}$: 1.

B. Spatiotemporal Compressed Sensing

As shown in Fig.1, the block diagram illustrates the basic framework of the spatiotemporal compressed sensing. In our proposed framework, the spatiotemporal compressed sensing essentially compresses the video scene into a single coded image and wirelessly transmits the coded image to off-chip terminal for reconstruction. Basically, the framework is composed of two stages: encoder and decoder. The encoder, which has been successfully implemented on chip [10], encodes the video scene into the coded image. Upon receiving the coded image, the decoder adopts the dictionary learning and sparse recovery to reconstruct the video.

To illustrate how spatiotempral compressed sensing works, we assume there is a spatiotemporal video signal $\mathbf{X} \in \mathbb{R}^{W \times H \times T}$, where $W \times H$ denotes the size of each frame, T denotes the total number of frames in the video and $\mathbf{X}(w, h, t)$ denotes the intensity value associated with the frame t at position (w, h). A sensing cube $\mathbf{S} \in \mathbb{R}^{W \times H \times T}$ is also given, which stores the spatiotemporal exposure control values for pixel at (w, h, t). In details, the sensing cube value for each pixel is defined as:

$$\mathbf{S}(w, h, t) = \begin{cases} 1 & t \in [t_1, t_2] \\ 0 & otherwise \end{cases}$$

where $0 \leq t_1 < t_2 \leq T$ and $t_2 - t_1$ intuitively indicates the exposure duration. t_1 is randomly chosen for each pixel while the exposure duration is fixed.

Given the video scene $\mathbf{X}$ and sensing cube $\mathbf{S}$, the coded image $\mathbf{Y} \in \mathbf{R}^{M \times N}$ is computed as:

$$\mathbf{Y}(w, h) = \sum_{t=1}^{T} \mathbf{S}(w, h, t) \cdot \mathbf{X}(w, h, t) \ \forall w, h.$$

Therefore, the spatiotemporal compressed sensing encodes the video $\mathbf{X}$ into a coded image $\mathbf{Y}$ at compression rate T : 1

During the reconstruction, we recover the spatiotemporal video, $\hat{\mathbf{X}} \in \mathbb{R}^{W \times H \times T}$, by solving the optimization problem,

$$\min_{\mathbf{a}} \frac{1}{2}||\mathbf{Y} - \mathbf{SDa}||_2^2 + \lambda||\mathbf{a}||_1,$$

where $\mathbf{D} \in \mathbb{R}^{N \times L}$ is an over-complete dictionary learned from the training sample [13] and $\mathbf{a} \in \mathbb{R}^L$ is the sparse coefficient vector. λ is the linear combination coefficient for controlling the sparsity in the recovery. Finally, the reconstructed video is computed as $\hat{\mathbf{X}} = \mathbf{Da}$.

III. Experiments

In this section, we evaluate the spatiotemporal CS approach on the database [14] compared to standard video compression MPEG in terms of reconstruction quality and compression rate (CR). The reconstruction quality is measured in peak signal-to-noise ratio (PSNR). The spatiotemporal video scenes are composed of two categories: "car" with slow motion and "pedestrian" with fast motion. The resolution of each video frame is 480×640 while the number of frame of video scenes is 10, 20 and 30 respectively. We also demonstrate the performance on the noisy video scenes to demonstrate the robustness. Additionally, we compare the spatiotemporal CS with other video compression techniques from the perspective of hardware implementation.

Table I and II demonstrate the performance on the videos without noise. In this experiment, MPEG technique dominates the spatiotemporal CS approach in terms of PSNR because MPEG can take advantage of precise motion detection and estimation to preserve the details of video scene without noise. The spatiotemporal CS approach loses some information in the encoding but the compression rate is more flexible compared to MPEG. Table III and IV demonstrate the performance on the videos with low noise. The proposed approach achieves better performance on the "car" videos and comparable performance on the "pedestrian" videos in terms of reconstruction quality. Furthermore, the proposed approach still achieves high compression rate while MPEG is reduced to around $5:1$. Table V and VI shows the performance on the videos with high noise, which demonstrates the robustness of proposed approach on the videos with high noise. The spatiotemporal CS approach outperforms MPEG in terms of reconstruction quality and compression rate. The robustness of spatiotemporal CS on noisy data benefits from the pixel-wise exposure, which averages the noise during the exposure. Besides, sparse reconstruction using the dictionary that learned from the training sample also further helps improve the robustness. Fig. 3 demonstrates the example of reconstruction video frames on different noisy level.

TABLE I. Comparison of reconstruction performance (in PSNR and CR) of "Car" without noise.

Approach	10 Frames		20 Frames		30 Frames	
	PSNR	CR	PSNR	CR	PSNR	CR
Spatiotemporal	25.08	10:1	25.17	**20:1**	23.43	**30:1**
MPEG	**31.45**	**16:1**	**31.30**	16:1	**31.19**	16:1

TABLE II. Comparison of reconstruction performance (in PSNR and CR) of "Pedestrian" without noise.

Approach	10 Frames		20 Frames		30 Frames	
	PSNR	CR	PSNR	CR	PSNR	CR
Spatiotemporal	24.59	10:1	23.74	**20:1**	23.58	**30:1**
MPEG	**30.03**	**15:1**	**30.46**	15:1	**30.52**	15:1

TABLE III. Comparison of reconstruction performance (in PSNR and CR) of "Car" with low noise.

Approach	10 Frames		20 Frames		30 Frames	
	PSNR	CR	PSNR	CR	PSNR	CR
Spatiotemporal	**23.69**	**10:1**	**23.73**	**20:1**	**22.96**	**30:1**
MPEG	22.62	5:1	22.55	3:1	22.52	5:1

Additionally, we also compare the spatiotemporal compressed sensing with other video compression techniques from the perspective of hardware implementation. In order to enable MPEG technique for mobile applications, the encoders have optimized the processing steps to reduce power and realize real-time encoding. Mochizuki [15] demonstrates a fully developed H.264/MPEG-4 codec on chip in the 90 nm process capable of encoding 30 fps HD-sized video in real-time. This implementation achieves more than a 50% increase in power efficiency as compared to other implementations of other MPEG on chip codecs. Comparably, our work implemented on chip demonstrates a 69% increase in power efficiency over this leading implementation. Other work [16] in MPEG-like compression for mobile applications demonstrate increased power efficiency as compared to our work. These new methods of encoding greatly reduce the computational intensity by approximating the DCT coefficients and reducing the area of search for motion estimation. The work presented by employing a H.265/HVEC encoder in the 28nm process compresses 30 fps HD-sized video with a power efficiency rating of .5 nJ/pixel. This is compared to our work which demonstrates a power efficiency rating of .7174 nJ/pixel and incorporates both video acquisition as well as compression in a process size 84% larger. Our work demonstrates the ability of spatial temporal compressive sampling to outperform conventional compression methods in terms of power efficiency and shows promise for further performance increase in smaller process sizes.

IV. Conclusion

In this paper, we propose a spatiotemporal compressed sensing framework for video compression. Taking advantage of spatiotemporal compressed sampling, the video scene can be encoded efficiently at high compression rate and decoded using sparse recovery. The evaluation results demonstrate its performance in terms of reconstruction quality and compression rate. Compared to conventional MPEG standard, proposed approach achieves better robustness on noisy database and realizes flexible compression rates. Furthermore, the simple arithmetic computation of spatiotemporal compressed sensing is suitable for the power-efficient and real-time biomedical CMOS implementation.

Acknowledgment

The authors are partially supported by NSF under grant DMS-1222567 and AFOSR under grant FA9550-12-1-0136.

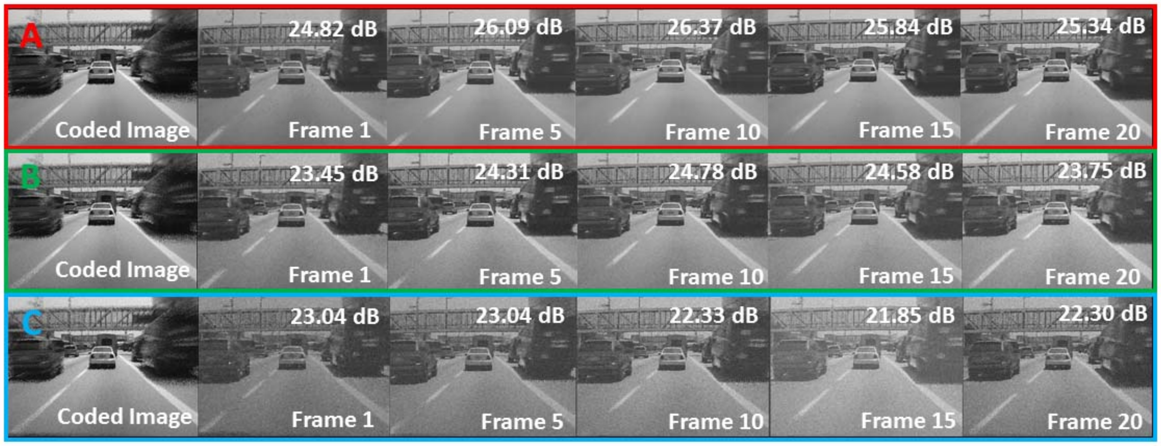

Fig. 3. The demonstration of coded images and reconstructed video frames at compression rate $20:1$. A (red): without noise; B (green): low noise; C (blue): high noise.

TABLE IV. COMPARISON OF RECONSTRUCTION PERFORMANCE (IN PSNR AND CR) OF "PEDESTRIAN" WITH LOW NOISE.

Approach	10 Frames		20 Frames		30 Frames	
	PSNR	CR	PSNR	CR	PSNR	CR
Spatiotemporal	**23.00**	**10:1**	21.45	**20:1**	22.48	**30:1**
MPEG	22.50	4:1	**22.51**	4:1	**22.53**	4:1

TABLE V. COMPARISON OF RECONSTRUCTION PERFORMANCE (IN PSNR AND CR) OF "CAR" WITH HIGH NOISE.

Approach	10 Frames		20 Frames		30 Frames	
	PSNR	CR	PSNR	CR	PSNR	CR
Spatiotemporal	**25.08**	**10:1**	**25.17**	**20:1**	**23.43**	**30:1**
MPEG	19.59	3:1	19.57	3:1	19.55	3:1

TABLE VI. COMPARISON OF RECONSTRUCTION PERFORMANCE (IN PSNR AND CR) OF "PEDESTRIAN" WITH HIGH NOISE.

Approach	10 Frames		20 Frames		30 Frames	
	PSNR	CR	PSNR	CR	PSNR	CR
Spatiotemporal	**22.36**	**10:1**	**22.05**	**20:1**	**22.04**	**30:1**
MPEG	19.68	3:1	19.69	3:1	19.68	3:1

TABLE VII. COMPARISON OF ON-CHIP IMPLEMENTATION (IN TECHNOLOGY (NM) AND POWER (NJ/PIXEL))

Our work [10]		Mochizuki [15]		Ju [16]	
Technology	Power	Technology	Power	Technology	Power
180	0.7	90	2.3	28	0.5

REFERENCES

[1] J. Senarathna, K. Murari, R. Etienne-Cummings, and N. V. Thakor, "A miniaturized platform for laser speckle contrast imaging," *IEEE transactions on biomedical circuits and systems*, vol. 6, no. 5, pp. 437–445, 2012.

[2] D. J. Cai, D. Aharoni, T. Shuman, J. Shobe, J. Biane, W. Song, B. Wei, M. Veshkini, M. La-Vu, J. Lou *et al.*, "A shared neural ensemble links distinct contextual memories encoded close in time," *Nature*, vol. 534, no. 7605, pp. 115–118, 2016.

[3] D. Mitrovic, "Video compression," *University of Edinburgh*, 2012.

[4] P. Llull, X. Yuan, X. Liao, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Temporal compressive sensing for video," in *Compressed Sensing and its Applications*. Springer, 2015, pp. 41–74.

[5] P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," *Optics express*, vol. 21, no. 9, pp. 10 526–10 545, 2013.

[6] R. Koller, L. Schmid, N. Matsuda, T. Niederberger, L. Spinoulas, O. Cossairt, G. Schuster, and A. K. Katsaggelos, "High spatio-temporal resolution video with compressed sensing," *Optics express*, vol. 23, no. 12, pp. 15 992–16 007, 2015.

[7] T.-H. Tsai, P. Llull, X. Yuan, L. Carin, and D. J. Brady, "Spectral-temporal compressive imaging," *Optics letters*, vol. 40, no. 17, pp. 4054–4057, 2015.

[8] D. Liu, J. Gu, Y. Hitomi, M. Gupta, T. Mitsunaga, and S. K. Nayar, "Efficient space-time sampling with pixel-wise coded exposure for high-speed imaging," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 248–260, 2014.

[9] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar, "Video from a single coded exposure photograph using a learned over-complete dictionary," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 287–294.

[10] J. Zhang, T. Xiong, T. Tran, S. Chin, and R. Etienne-Cummings, "Compact all-cmos spatiotemporal compressive sensing video camera with pixel-wise coded exposure," *Optics express*, vol. 24, no. 8, pp. 9013–9024, 2016.

[11] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

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

[13] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[14] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 304–311.

[15] S. Mochizuki, T. Shibayama, M. Hase, F. Izuhara, K. Akie, M. Nobori, R. Imaoka, H. Ueda, K. Ishikawa, and H. Watanabe, "A 64 mw high picture quality h. 264/mpeg-4 video codec ip for hd mobile applications in 90 nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 11, pp. 2354–2362, 2008.

[16] C.-C. Ju, T.-M. Liu, K.-B. Lee, Y.-C. Chang, H.-L. Chou, C.-M. Wang, T.-H. Wu, H.-M. Lin, Y.-H. Huang, C.-Y. Cheng *et al.*, "A 0.5 nj/pixel 4 k h. 265/hevc codec lsi for multi-format smartphone applications," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 56–67, 2016.