Weighted Measurement Reuse for Compressive Video Sensing Reconstruction

Ivan Lee
School of Computer and Information Science
The University of South Australia
Adelaide, Australia
Email: Ivan.Lee@unisa.edu.au

***Abstract*—Compressive sensing is a technique to sample signals with sparsity below the Nyquist sampling rate. For video compression, encoding individual frames independently using compressive sensing is inefficient since inter-frame similarities from the temporal domain are not utilised. This paper investigates the performance of weighted compressive video sensing, which brings predicted frame and bi-directional predicted frame from traditional video codecs to compressive video sensing. The weighted compressive video sensing applies only to the decoder to reconstruct video frames, without changing the low complexity characteristic of compressive video sensors. This paper examines weighted measurement reuse on different scenarios with previous frame only, subsequent frame only, and both previous and subsequent frames.**

I. Introduction

In the information age, multimedia communication has become a popular service to the general public. Multimedia enabled devices, such as wireless video sensors, mobile phones, set-top boxes, multimedia gateways, and video on-demand servers, are widely used today. These multimedia devices are interconnected via the Internet to facilitate various applications such as video conferencing, on-demand video streaming, and video surveillance.

Modern video compression algorithms yield a high-performance quality versus bitrate ratio, with a high complexity coder and a low complexity decoder [1], [2]. The simplified decoder design works well with traditional multimedia application such as video storage on a DVD where the coding process is performed once and the decoding process (for playing back video) is performance multiple times. On the other hand, the complicated encoder algorithm is not suitable for multimedia sensors where a limited processing capability and a reduced power supply are usually assumed [3], [4]. This paper investigates compressive video sensing architecture, to address the requirement of video sensors with demands on limited power constraints and processing capacities (power), transmission bandwidth (rate), and quality level (distortion) expectations[5], [6].

The classic Shannon-Nyquist sampling theorem indicates that there is no information loss if the signal is sampled at twice its bandwidth. Compressive sensing (CS) is an emerging research topic which takes advantages of signal sparsity and signal incoherence [3], [7] to reduce the sampling rate. Compressive sensing [3], [8] utilises the sensing operator, which is like a convex lens, to convert a signal into one concentrated sample every time. When sufficient number of samples obtained (which could be below the Shannon-Nyquist sample rate), CS theory believes the signal can be precisely restored. This theory has attracted great interest in image signal processing because of its low encoding complexity, whereas heavy computation offloaded to the decoder. The problem of sparse signal recovery from a small number of incoherent measurements follows uniform uncertainty principle. For example, high resolution image can be generated from low-resolution compressed video [9], [10]. As the number of measurements increases, error decreases at near-optimal rate. This property is ideal for constructing a video codec with scalable quality levels. In addition, compressive sensing is democratic and robust, which means that all measurements are equally as important. This property helps reconstructing original signal when information are decoded out-of-order, which may be a result of today's best effort, multi-path interconnected networks. Compressive sensing offloads computing power from the encoder to the decoder, which means that the posteriori computing power is used to reduce the priori sampling complexity [8]. In this paper, CS-based video codec is designed for high performance video decoders.

If an N-length real and discrete signal can be sparsely represented with less than N non-zero coefficients in a transform domain, it is regarded as a sparse or compressible signal. In other words, if the signal can be converted to a sparse form with only K non-zero coefficients and $K \ll N$, the signal is considered as sparse; if the largest K coefficients in the sparse form can retrieve a perfect approximation of the signal and $K \ll N$, the signal is considered as a compressible signal. CS technique can be applied as long as the signal possesses either of the above features.

Image and video signals are typically dense in spatial domain, but signal processing techniques can be taken to transform image and video signals with sparse representation, such as using discrete cosine transform (DCT) in JPEG [11] or wavelet transform in JPEG-2000 [12]. Using DCT as example, let $X = [x_1, x_2, \ldots, x_N]^T$ denote an N-length real and discrete raster-scanned image signal and Ψ denote the collection of orthogonal basis of 1-D DCT transform, then $X = \Psi X = \sum_{i=1}^{N} s_i \Psi_i$, where $\Psi = (\psi_1|\psi_2|\ldots\ |\psi_N)$ is an n-by-n DCT transformation matrix with ψ_i being a column vector of Ψ. $S = [s_1, s_2, \ldots, s_n]^T$ is a sparse representation of x in DCT domain and the sparsity degree of S (also the number of non-zero DCT coefficients) is hypothetically K ($K \leq N$).

CS sampling process uses an m-by-n measurement matrix Φ to measure x and to obtain an m-dimensional measured vector y. The components in the measured vector are called measurements. To let Φ and Ψ comply with Restricted Isometry Property (RIP), the measurement matrix $\Phi = (\phi_1|\phi_2|\ldots|\phi_m)^T$ is constructed using a Gaussian random matrix with entries are independent and identically distributed (i.i.d) random variables with a zero mean.

The measured vector, y, can be represented as $y = \Phi x = \Phi\Psi S$. In general, m is much smaller than n and comparable with K, which means that the n-dimensional signal is reduced to an m-dimensional measured vector, hence compression is embedded within the the sampling process. If Φ is incoherent with Ψ and follows RIP, the target image signal can be reconstructed with measured vector y, and $m \geq O(cKlog(n/K))$, where c is a universal constant and K denotes the sparsity degree of a sparse representation. Popular algorithms for reconstructing compressed sensing signals include orthogonal matching pursuit [13], gradient projection for sparse reconstruction [14], and subspace pursuit [15].

II. Compressive Video Sensing with Measurement Reuse

Motion compensation is a popular approach for improving the compression ratio for modern video codecs [16]. On the other hand, exhaustive macro-block matching process to choose the optimal motion vector demands extensive computing power, which may not be suitable for low power devices such as video sensors in wireless sensor networks. Therefore, instead of applying motion compensation in the video encoding process, decoder driven CS measurement reuse is investigated in this paper. Using the same underlying assumption regarding similarities between adjacent video frames (x_i and x_{i+1} are similar where i denotes the frame number within the same GOP), it is expected that S_i and S_{i+1} will also share similar pattern of K-sparsity. Therefore, it is expected that the measured vector y_{i-1} is highly correlated to y_i, so that y_{i-1} could be re-used for reconstructing x_i according to [17].

If Pan-Tilt-Zoom (PTZ) features are available on top of video sensors such as single pixel camera [4], it is possible to zoom into a Region of Interest (ROI) for capturing more foreground CS measurements. However, this approach relies on mechanical PTZ movement for selecting target blocks may not be a cost-effective solution. Instead, a masking disk may be applied to perform block-based compressive video sensing.

Figure 1 illustrates the structure of a block-based compressive video sensing camera based on single pixel camera [9]. A image is firstly controlled by a lens to control the incoming image project to a dense mirror array. In between the lens and the dense mirror array, a rotary masking disk is used to control different block regions of the incoming image can be passed for capturing. The dense mirror array dynamically changes mirror reflection angles on a pseudo-random manner controlled by the encoder and decoder, and the reflected image is sent through a convex lens and projected to a photo-diode at the lens focal point. The signal captured by the photo-diode is called a measurement of compressive video sensing.

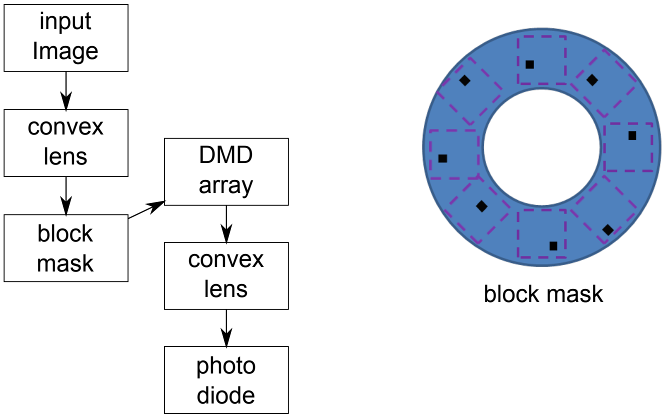


Fig. 1. Block-based single pixel camera

The masking disk can be used to control the order of blocks captured by the single pixel camera. Since the masking pattern is fixed which is known to both the encoder and the decoder, location of the block can be determined from the order of the sensing measurements. The encoder can terminate the transmission of sensing measurement to selected blocks, hence to reduce the data rate for communication. The decoder can also ignore selected blocks for video reconstruction to reduce the computational demand. In other words, the masking disk facilitates coding of region-of-interest which can be managed both at the encoder and the decoder.

This paper examines the performance of weighted measurement reuse in compressive video sensing. All blocks are used at the encoder and the decoder; in other words, the entire video frame is considered as the region-of-interest.

III. Weighted Measurement Reuse in Compressive Video Sampling

As indicated in Section II, due to similarities between adjacent video frames, measurements can be reused to improve the reconstructed video quality. The concept of CS measurement reuse is similar to interpolation in the temporal domain; however, instead of applying interpolation techniques, measurement in CS-domain is used at the decoder. As illustrated in Figure III, when frame i is reconstructed at the decoder, its previous frame $i-1$ and its subsequent frame $i+1$ can contribute to improve the signal recovery. The number of measurement used for frame $i-1$, i, and $i+1$ are denoted as w_{i-1}, w_i, and w_{i+1}, respectively.

When compressive video sensing is applied in the encoder, measurements are captured sequentially. Reconstructed video frames without measurement reuse are like I-frames in traditional video codec, such that no drift-error will be encountered. Reconstructed video frames with measurements

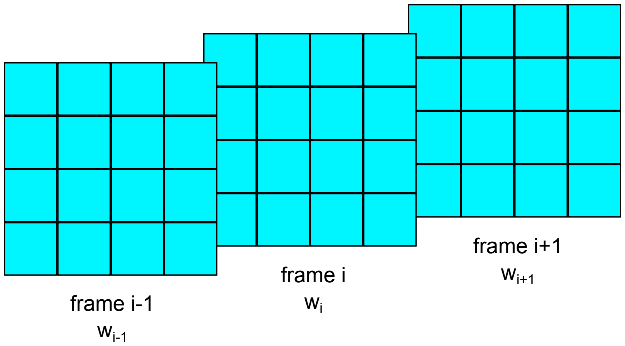


Fig. 2. Weighted measurement reuse

reused from previous-frame are like P-frames, whereas the ones with measurements reused from both previous and subsequent frames are like bi-directional predicted frames. Given that the decoding process is assumed on a device with powerful computing facility (whereas the encoder is simplified targeting for wireless sensor networks, for example,) re-using all measurements from previous frame is a feasible option providing that a sufficient storage is available to caching previous measurements and a sufficient processing power is equipped for additional computing. On the other hand, measurement reuse from subsequent frames will introduce additional delay, this is because, like B-frames, measurements from a subsequent frames need to be available to reconstruct a current frame.

It is worth noting that compressive sensing at its current status does not outperform traditional transformation-based image/video codecs. However, it poses a promising alternative such that a simplified encoder is required and the excessive computation is offloaded to the decoder. Such characteristics represent a unique niche for wireless sensor networks or remote sensing applications.

IV. Experimental Results

The luminance component (Y) on the first viewpoint of standard Ballroom video sequence is used in the experiment to examine the performance of weighted measurement reuse. To lower the simulation complexity, a reduced number of video frames is used in the experiment. Each video frame is partitioned into 16x16 blocks for encoding, and the number of sensing measurement is chosen to be up to 5 percent of each block's pixel count (i.e. 13 sensing measurements per block.) The reconstruction algorithm applied in this paper is reconstruction using gradient projection for sparse reconstruction[14].

In the experiment, a reconstructed video without measurement reuse has yielded a peak-signal-to-noise (PSNR) value at 21.67dB. Table I shows the output of different experimental setup: measurement reuse from previous frame (previous-frame-only), measurement reuse from subsequent frame (subsequent-frame-only), and measurement reuse from both previous and subsequent frames (previous-subsequent-frames). The first setup shows that an improved video quality is reconstructed with an increase in measurement reuse, and the same characteristics are observed for the second and the third setup. Such observation concludes that measurement reuse helps improving the reconstructed video quality in compressive sensing by taking inter-frame similarity from the temporal domain.

Comparing the first two setups, subsequent-frame-only yields a slightly better video quality; however, the difference is marginal, with PSNR gain up to 0.27dB. The third setup with measurement reuse from both previous and subsequent frames yields a consistent quality gain, with up to 1.73dB gain in PSNR over the second setup.

Sample video output is illustrated in Figure 3, taken from frame 3, viewpoint 1 of the Ballroom standard sequence. The extreme scenarios, reconstructed video frame without measurement reuse (I-frame alike) 3(a), and the one with full measurement reuse from previous and subsequent frames (B-frame alike) 3(f), demonstrate a 2.64dB gain in video quality.

Similar experiment has been conducted with the luminance component of standard News sequence. With the same block size (16x16), up to 24 measurements (9.4% of the number of pixels) are used. The reconstructed video qualities, using PSNR and structural similarity (SSIM) index [18], are obtained from the mean of 300 video frames in the News sequence. The average video quality under different scenarios are shown in Table II. Similar to the observations on the Ballroom sequence, the video quality in terms of both PSNR and SSIM improves with an increase of measurement reuse from either previous frame, or subsequent frame, or both frames. Example reconstructed video frames (frame 30 of the News sequence) are shown in 4. The extreme scenarios in 4(a) and 4(f) represent 4.96dB improvement in PSNR and 0.1611 improvement in SSIM.

V. Conclusion

This paper investigated weighted measurement reuse which bringing inter-frame analysis into compressive video sensing, and examined its impact on the decoded video quality. It was observed that increasing the number of reused sensing measurements, the reconstructed video quality is consistently improved for all scenarios of previous-frame-only, subsequent-frame-only, and previous-subsequent-frames setups. We concluded that higher the weighting the measurement reuse from adjacent frames leads to an improved reconstructed video quality. Assuming that the decoder is equipped with sufficient memory storage and extended processing power, all measurement from previous frame should be reused. Measurement reuse from subsequent frame improves the reconstructed video quality at the cost of additional delay, thus the trade-off decision is available to the user.

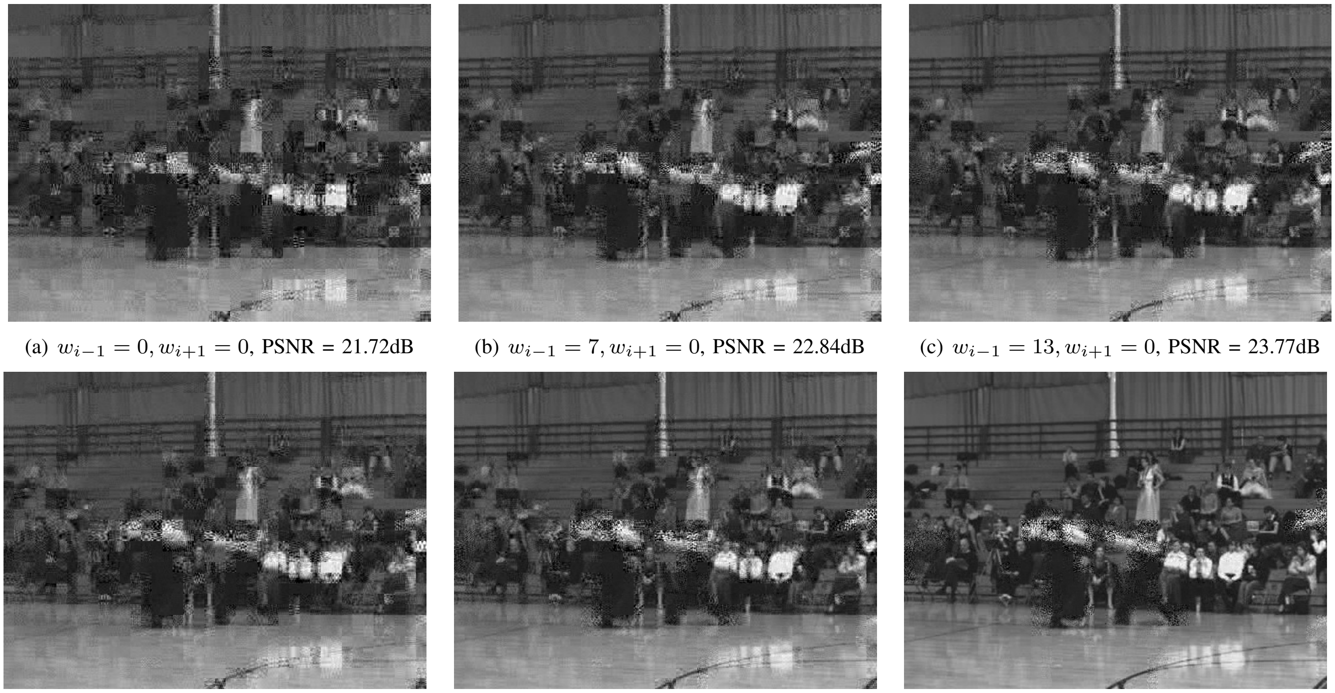
References

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/AVC video coding standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] T. Sikora, "Trends and perspectives in image and video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 6–17, 2005.

TABLE I
RECONSTRUCTED VIDEO QUALITY IN PSNR USING VIEWPOINT 1 OF THE BALLROOM SEQUENCE

Measurements	1	3	5	7	9	11	13
Previous-frame-only	21.82	22.20	22.52	22.82	23.13	23.39	23.59
Subsequent-frame-only	21.94	22.36	22.79	23.09	23.39	23.58	23.78
Previous-Subsequent-frames	23.67	23.80	23.88	24.00	24.10	24.11	24.45



(a) $w_{i-1} = 0, w_{i+1} = 0$, PSNR = 21.72dB

(b) $w_{i-1} = 7, w_{i+1} = 0$, PSNR = 22.84dB

(c) $w_{i-1} = 13, w_{i+1} = 0$, PSNR = 23.77dB

(d) $w_{i-1} = 0, w_{i+1} = 7$, PSNR = 23.08dB

(e) $w_{i-1} = 13, w_{i+1} = 7$, PSNR=24.16dB

(f) $w_{i-1} = 13, w_{i+1} = 13$, PSNR = 24.36dB

Fig. 3. Sample video frames of reconstructed Ballroom sequence viewpoint 1 (frame 3), with 13 measurements of current frame

TABLE II
RECONSTRUCTED VIDEO QUALITY IN PSNR & SSIM USING THE NEWS SEQUENCE

Measurements		3	6	9	12	15	18	21	24
Previous-frame-only	PSNR (dB)	22.09	22.45	22.78	23.09	23.40	23.81	24.10	24.44
	SSIM	0.6968	0.7135	0.7285	0.7415	0.7537	0.7662	0.7768	0.7866
Subsequent-frame-only	PSNR (dB)	22.08	22.44	22.78	23.13	23.47	23.77	24.10	24.41
	SSIM	0.6963	0.7132	0.7280	0.7413	0.7540	0.7658	0.7760	0.7861
Previous-Subsequent-frames	PSNR (dB)	24.67	24.88	25.09	25.26	25.48	25.68	25.84	26.04
	SSIM	0.7922	0.7980	0.8040	0.8091	0.8146	0.8199	0.8246	0.8294

[3] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, p. 12891306, 2006.

[4] R. G. Baraniuk, “Compressive sensing,” *IEEE Signal Processing Magazine*, vol. 24, no. 4, p. 118, 2007.

[5] G. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[6] Z. He, Y. Liang, L. Chen, I. Ahmad, and D. Wu, “Power-rate-distortion analysis for wireless video communication under energy constraints,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 5, pp. 645–658, 2005.

[7] E. Candes and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?” *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[8] M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk, “An architecture for compressive imaging,” in *IEEE International Conference on Image Processing*, 2007, pp. 1273–1276.

[9] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, “Single-pixel imaging via compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, 2008.

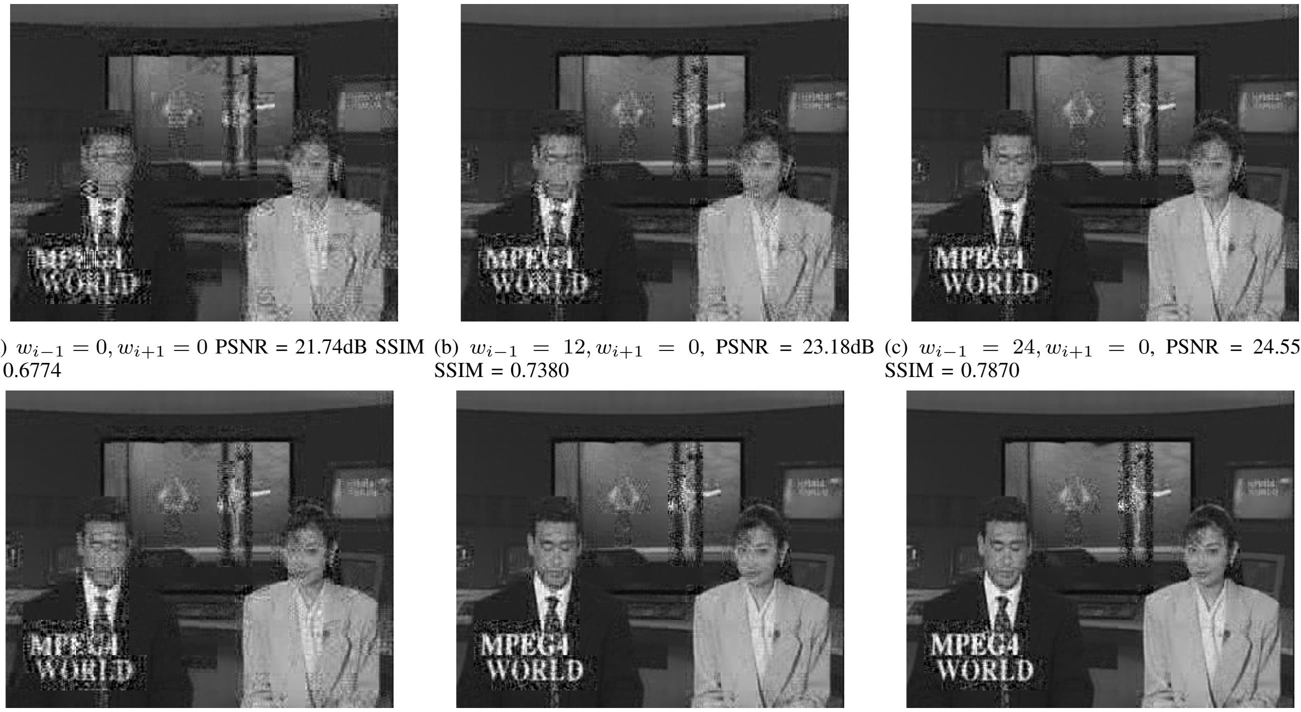
[10] C. A. Segall, R. Molina, and A. K. Katsaggelos, “High-resolution images from low-resolution compressed video,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 37–48, 2003.

[11] G. Wallace, “The jpeg still picture compression standard,” *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[12] B. Usevitch, “A tutorial on modern lossy wavelet image compression: foundations of jpeg 2000,” *Signal Processing Magazine, IEEE*, vol. 18, no. 5, pp. 22–35, 2001.

[13] Z. Wang and I. Lee, “Sorted random matrix for orthogonal matching pursuit,” in *International Conference on Digital Image Computing: Techniques and Application (DICTA)*, 2010, pp. 116–120.

[14] M. Figueiredo, R. Nowak, and S. Wright, “Gradient projection for sparse



(a) $w_{i-1} = 0, w_{i+1} = 0$ PSNR = 21.74dB SSIM = 0.6774

(b) $w_{i-1} = 12, w_{i+1} = 0$, PSNR = 23.18dB SSIM = 0.7380

(c) $w_{i-1} = 24, w_{i+1} = 0$, PSNR = 24.55dB SSIM = 0.7870

(d) $w_{i-1} = 0, w_{i+1} = 12$, PSNR = 23.36dB SSIM = 0.7461

(e) $w_{i-1} = 24, w_{i+1} = 12$, PSNR = 25.67dB SSIM = 0.8145

(f) $w_{i-1} = 24, w_{i+1} = 24$, PSNR = 26.70dB SSIM = 0.8385

Fig. 4. Sample video frames of reconstructed News sequence (frame 30), with 24 measurements of current frame

reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[15] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[16] D. Marpe, T. Wiegand, and G. Sullivan, "The h. 264/mpeg4 advanced video coding standard and its applications," *IEEE Communications Magazine*, vol. 44, no. 8, pp. 134–143, 2006.

[17] Z. Wang and I. Lee, "A study of video coding by reusing compressive sensing measurements," in *International Workshop on Ubiquitous Multimedia Computing and Communication*, 2010, pp. 64–69.

[18] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.