DESIGN OF A COMPRESSIVE REMOTE IMAGING SYSTEM COMPENSATING A HIGHLY LIGHTWEIGHT ENCODING WITH A REFINED DECODING SCHEME

George Tzagkarakis[1], Arnaud Woiselle[2], Panagiotis Tsakalides[3] and Jean-Luc Starck[1]
[1]*CEA/DSM, SEDI-SAp, Service d'Astrophysique, Centre de Saclay, F-91191 Gif-Sur-Yvette cedex, France*
[2]*Sagem Défense Sécurité, 95100 Argenteuil, France*
[3]*Institute of Computer Science (ICS) - Foundation for Research & Technology - Hellas (FORTH), Crete, Greece*
{georgios.tzagkarakis, arnaud.woiselle, jstarck}@cea.fr, tsakalid@ics.forth.gr

Keywords: Compressive Video Sensing:Lightweight Remote Imaging Systems.

Abstract: Lightweight remote imaging systems have been increasingly used in surveillance and reconnaissance. Nevertheless, the limited power, processing and bandwidth resources is a major issue for the existing solutions, not well addressed by the standard video compression techniques. On the one hand, the MPEGx family achieves a balance between the reconstruction quality and the required bit-rate by exploiting potential intra- and inter-frame redundancies at the encoder, but at the cost of increased memory and processing demands. On the other hand, the M-JPEG approach consists of a computationally efficient encoding process, with the drawback of resulting in much higher bit-rates. In this paper, we cope with the growing compression ratios, required for all remote imaging applications, by exploiting the inherent property of compressive sensing (CS), acting simultaneously as a sensing and compression framework. The proposed compressive video sensing (CVS) system incorporates the advantages of a very simple CS-based encoding process, while putting the main computational burden at the decoder combining the efficiency of a motion compensation procedure for the extraction of inter-frame correlations, along with an additional super-resolution step to enhance the quality of reconstructed frames. The experimental results reveal a significant improvement of the reconstruction quality when compared with M-JPEG, at equal or even lower bit-rates.

1 INTRODUCTION

Modern lightweight digital sensing devices with high-resolution signal acquisition, processing, and communication capabilities are largely based on the well-established Shannon and Nyquist theories. Managing ever increasing amounts of data remains a challenging task, especially for practical applications, where devices with limited processing, storage, and bandwidth resources are involved. Moreover, the increasing demand for higher acquisition rates and even improved resolution is placing a significant burden on the existing hardware architectures.

Video acquisition and processing posses a central role in numerous emerging applications, such as surveillance and reconnaissance, both at the civilian and battlegroup levels, robot navigation, remote surgery and entertainment. Most video surveillance systems monitor actively and remotely an area of interest through video streaming, or passively by storing the captured video for future use. In addition, recent technological advances enable the design of low-cost devices that incorporate enhanced multimodal sensing, processing, and communication capabilities. At the same time, the limited resources of the compression hardware is still a major bottleneck for the design of lightweight remote imaging systems, where very low data rates are strongly required to maximize the lifetime of the system (*e.g.,* in the case of a terrestrial sensor network), while preserving an increased performance. To cope with such growing compression ratios existing video compression techniques, such as the MPEGx profiles and the Motion JPEG (M-JPEG) scheme, may result in poor image quality.

The framework of *compressive sensing* (CS) (Candés et al., 2006), acting simultaneously as a sensing and compression protocol, could be exploited in the design of low-complexity onboard remote imaging devices with reduced power and processing requirements. Its simplicity stems from the linearity of the associated non-adaptive incoherent projections, which are employed for the representation and reconstruction of sparse signals. Recently, the framework of compressive video sensing (CVS) was introduced as a natural extension proposing distinct approaches for video acquisition using a reduced amount of data, while maintaining a similar reconstruction performance when compared to standard video compression techniques.

Existing CVS approaches perform separate encoding of each frame, based on a non-overlapping block splitting to reduce the storage and computational costs, by combining full sampling of reference frames with CS applied on non-reference frames. Then, at the decoder, the reconstruction is performed separately (Stanković et al., 2008), or jointly by either considering a joint sparsity model as in (Kang and Lu, 2009) or by designing an adaptive sparsifying basis using neighboring blocks in previously reconstructed frames (Do et al., 2009; Prades-Nebot et al., 2009). The major drawbacks are that, since potential spatio-temporal redundancies are not removed at the encoder, the corresponding CVS methods usually result in increased bit-rates, while also being sensitive to the propagation of reconstruction errors along the sequence in the case of joint decoding.

The efficiency of typical video compression standards, such as the MPEGx, in achieving a good trade-off between the reconstruction quality and the associated bit-rates, is primarily based on the capability of removing potential spatio-temporal redundancies by means of intra-frame transform coding and inter-frame motion prediction. However, an encoder with increased memory and processing resources is required, which may be prohibitive in a lightweight remote imaging system. On the other hand, the use of M-JPEG, which is an intra-frame-only video compression scheme, has the advantage of imposing significantly lower processing and memory requirements on the hardware, but at the cost of increasing significantly the required bit-rate, which is restrictive in the case of limited bandwidth.

In the present work, we address the drawbacks of the previous CVS methods, as well as the limitations of MPEGx and M-JPEG compression techniques, by introducing a CVS scheme which could be integrated in onboard video sensing devices with restricted resources. In particular, the proposed CVS method combines a simplified encoding process by embedding a CS module in an M-JPEG-like encoder, along with a refinement phase based on inter-frame prediction at the decoder (as opposed to MPEGx, where the prediction errors are formed at the encoder). The idea of transferring the computational burden of the motion estimation and compensation processes at the decoder was also appeared in (Jung and Ye, 2010) in the framework of dynamic magnetic resonance imaging, where an auxiliary sequence of residual frames was generated at the decoder recursively using a set of fully-sampled reference frames in conjunction with the low-resolution dynamic frames. Moreover, the required bit-rate of our proposed encoder can be further decreased by downsampling the non-reference frames, followed by an additional super-resolution step at the decoder to restore the reconstructed frames in their original resolution. The use of super-resolution as a tool to resize the frames in their original dimension is motivated by recent works on sparse representation-based image super-resolution via dictionary learning (Freeman et al., 2002; Yang et al., 2008; Wang et al., 2011; Zhang et al., 2011), where it has been shown that a super-resolution method results in images with superior quality when compared to the commonly used 2-dimensional interpolation schemes (*e.g.,* bilinear, bicubic, spline).

The paper is organized as follows: in Section 2, the model for the compressed measurements acquisition is introduced. Section 3 describes in detail the proposed CVS architecture, while a performance evaluation is carried out in Section 4. Finally, conclusions and further extensions are outlined in Section 5.

2 CS MEASUREMENTS MODEL

In the following, we consider for convenience the case of square $N \times N$ frames, although the proposed approach is extended straightforwardly in the general non-square case. The main disadvantage when we deal with a remote imaging device with limited capabilities, as mentioned in Section 1, is the high memory and computational expense when we operate at high resolutions. This drawback can be alleviated by proceeding in a block-wise fashion. More specifically, in the proposed CVS system, each frame is divided into equally sized $B \times B$ non-overlapping blocks. Then, a measurement vector $\mathbf{g}_j$, $j = 1, \ldots, n_B$, is generated for each one of the n_B blocks via a simple linear model as follows,

$$\mathbf{g}_j = \Phi_j \mathbf{x}_j \ , \tag{1}$$

where $\Phi_j \in \mathbb{R}^{M \times B^2}$ is a suitable measurement matrix ($M \ll B^2$) and $\mathbf{x}_j \in \mathbb{R}^{B^2}$ denotes the j-th block of frame $\mathbf{x}$, reshaped as a column vector. Although, in general, a distinct measurement matrix can be assigned to each block, in the following we apply the same measurement matrix on each block for simplicity, that is, $\Phi_j \equiv \Phi$, $j = 1, \ldots, n_B$. Thus, the measurement vector $\mathbf{g}_j$ provides directly a compressed representation of the original space-domain block $\mathbf{x}_j$.

Common choices for the measurement matrix Φ are random matrices with independent and identically distributed (i.i.d.) Gaussian or Bernoulli entries. In a remote imaging system some additional requirements should be posed on the choice of Φ, such as the use of a minimal number of compressed measurements, and the fast and memory efficient computation along with a "hardware-friendly" implementation. A class of matrices satisfying these requirements, the so-called *structurally random matrices*, was introduced recently (Do et al., 2008). The block Walsh-Hadamard (BWHT) operator is a typical member of this class and will be used in the proposed design.

If the j-th block is compressible or has a K-sparse representation in an appropriate sparsifying transform domain, and if the measurement matrix along with the sparsifying dictionary satisfy a sufficient incoherence condition, then $\mathbf{x}_j$ can be recovered from $M \gtrsim O(2K \log B)$ measurements by solving the following optimization problem,

$$\min_{\mathbf{w}_j} \|\mathbf{w}_j\|_1 \quad \text{s.t.} \quad \mathbf{g}_j = \Phi \Psi_s^{-1} \mathbf{w}_j \,, \tag{2}$$

where Ψ_s is an appropriate *sparsifying transformation*, such as an orthonormal basis (*e.g.,* discrete cosine transform (DCT), discrete wavelet transform (DWT)) or an overcomplete dictionary (*e.g.,* undecimated DWT (UDWT)), and $\mathbf{w}_j$ is the transform-domain sparse representation of $\mathbf{x}_j$ in Ψ_s. If Ψ_s is different than the identity, then the reconstruction of the j-th block is first performed in the transform domain, $\hat{\mathbf{w}}_j$, followed by an inversion to obtain the final space-domain estimate, $\hat{\mathbf{x}}_j = \Psi_s^{-1} \hat{\mathbf{w}}_j$, otherwise the space-domain solution is obtained directly, that is, $\hat{\mathbf{x}}_j \equiv \hat{\mathbf{w}}_j$. In the subsequent experimental evaluations, and in order to be consistent with the M-JPEG compression scheme to which we compare, the DCT is used as the sparsifying dictionary.

The next section analyzes in detail the structural components of the proposed CVS architecture, starting with a brief overview of M-JPEG. The main reason for choosing the M-JPEG as a benchmark method for comparison, and not a member of the MPEGx family, stems from the fact that both M-JPEG and our approach are based on a very simplified encoder, working directly in the original frame-domain, without exploiting inter-frame redundancies via motion compensation, as in MPEGx, which is proven to be the step with the highest memory and power consumption. We emphasize again that our goal is to build a CVS system with a lightweight encoder, so as to satisfy the constraints of a remote imaging system. On the other hand, we consider that the reconstruction takes place at a base station, where increased memory and computational resources are available in practice.

3 PROPOSED CVS SYSTEM

In the following, the main building blocks of the proposed CVS system are described in detail. Before proceeding, we start with a brief overview of the basic structure of M-JPEG, to which we compare in the rest of the paper.

3.1 Overview of M-JPEG

M-JPEG is a lossy intra-frame compression scheme. In particular, each frame of the video sequence is transformed in the DCT domain, followed by the quantization of the corresponding DCT coefficients using a perceptual model based loosely on the human visual system. This model discards high-frequency information since its changes are almost unperceivable to the human eye. The quantized coefficients are then encoded losslessly and packed into the output bit-stream. The building blocks of an M-JPEG system are shown in Fig. 1

As a purely intra-frame-only compression scheme, the performance of M-JPEG is related directly to the spatial complexity of each video frame. In particular, frames that contain large smooth transitions or monotone surfaces are compressed well, and are more likely to preserve their original details with few visible compression artifacts. On the contrary, frames containing structures such as complex textures, fine curves and lines are prone to exhibit artifacts such as ringing, smudging, and macroblocking. An advantage of M-JPEG is that it is insensitive to the motion complexity.

Since the video frames are compressed separately, M-JPEG imposes lower processing and memory requirements on the hardware devices, when compared with MPEGx. This justifies its widespread use in digital and IP cameras, in HDTV media players and game consoles. Although the bit-rate of M-JPEG is lower than that of an uncompressed video, however, it is much higher than the bit-rate of a video which is compressed using inter-frame motion compensation, such as the MPEGx does.

Motivated by the above, the proposed CVS encoder is designed with the goal of preserving the simple structure of an M-JPEG-based encoder, with the power of CS to represent the information content of a given signal, characterized by a sparse representation

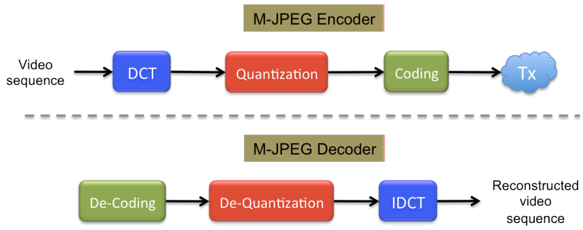


Figure 1: Structure of an M-JPEG system.

in an appropriate transform domain, using a highly reduced set of compressed measurements.

3.2 CVS architecture

3.2.1 Encoder

The need to satisfy the constraints of a lightweight remote imaging system motivated the design of a CVS encoder which combines the simplicity of the M-JPEG approach, with the efficiency of CS to represent the salient information content of a given signal via a low-dimensional set of compressed measurements. Fig. 2(a) shows the building components of the proposed CVS encoder. More specifically, the input video sequence is divided into successive groups of pictures (GOPs) of the form IPP…PI, where I corresponds to a reference frame (I-frame) and P to a non-reference frame (P-frame). The I-frames are fully sampled and compressed using JPEG, that is, the DCT is applied on non-overlapping 8 × 8 blocks of the original frame, followed by quantization and encoding of the transform coefficients.

Without loss of generality, in the following we will focus on the case of grayscale videos. The quantization of the DCT coefficients corresponding to the luminance component is performed as follows,

$$\mathbf{D}_q(m,n) = \text{round}\left(\frac{\mathbf{D}(m,n)}{S\cdot\mathbf{Q}(m,n)}\right) , \qquad (3)$$

where $\mathbf{D}\in\mathbb{R}^{8\times8}$ denotes the matrix of DCT coefficients for an 8 × 8 block of the original frame, $\mathbf{D}_q\in\mathbb{R}^{8\times8}$ is the corresponding matrix of the quantized coefficients, S is a scaling factor varying from 1 (high image quality) to 100 (low image quality), and $\mathbf{Q}\in\mathbb{R}^{8\times8}$ is the following quantization table,

$$\mathbf{Q} = \begin{bmatrix} 8 & 16 & 19 & 22 & 26 & 27 & 29 & 34 \\ 16 & 16 & 22 & 24 & 27 & 29 & 34 & 37 \\ 19 & 22 & 26 & 27 & 29 & 34 & 34 & 38 \\ 22 & 22 & 26 & 27 & 29 & 34 & 37 & 40 \\ 22 & 26 & 27 & 29 & 32 & 35 & 40 & 48 \\ 26 & 27 & 29 & 32 & 35 & 40 & 48 & 58 \\ 26 & 27 & 29 & 34 & 38 & 46 & 56 & 69 \\ 27 & 29 & 35 & 38 & 46 & 56 & 69 & 83 \end{bmatrix} \qquad (4)$$

Then, the quantized DCT coefficients are encoded losslessly using an improved Huffman coding scheme using recursive splitting (Skretting et al., 1999).

In contrast to the full sampling of the I-frames, a CS measurements acquisition process is applied on the P-frames. In particular, a compressed measurement vector is generated for each $B\times B$ block of a P-frame. A satisfactory tradeoff between the computational complexity at the encoder and the reconstruction performance at the decoder is achieved for blocks of size 16 × 16 and 32 × 32. A further reduction of the transmission bit-rate can be attained by downsampling a P-frame prior to the measurements acquisition, that is, the measurement model given by (1) is generalized as follows,

$$\mathbf{g}_j = \Phi\mathcal{D}\{\mathbf{x}_j\} , \qquad (5)$$

where $\mathcal{D}\{\cdot\}$ denotes the downsampling operator. In the subsequent evaluations we will consider that each P-frame is downsampled by a factor of 2. Then a uniform scalar quantization is applied on the CS measurements, followed by Huffman coding.

From the above it can be seen that the parameters which affect the performance of the encoder are: i) the size of a GOP, ii) the scaling factor S (for the I-frames) and the number of quantization levels (for the P-frames), and iii) the sampling rate M/B^2, where M is the number of compressed measurements for a block of size $B\times B$. More specifically, the size of a GOP is related to the motion complexity of a video sequence, and it should decrease (or equivalently, more fully-sampled I-frames must be inserted) for videos with highly varying content. Of course, as the number of the fully-sampled I-frames increases, the required bit-rate also increases. Regarding the quantization of the I-frames, the image quality diminishes for increasing values of S, while the quality of the P-frames improves by increasing the number of quantization levels of the uniform quantizer. Finally, the reconstruction quality is enhanced as the sampling rate grows, but at the cost of higher bit-rates.

3.2.2 Decoder

In contrast to the MPEGx, whose encoding efficiency is primarily based on the inter-frame prediction, the lack of a motion compensation step at the proposed encoder, as well as in M-JPEG, to account for the temporal redundancies, affects the tradeoff between the reconstruction quality and the associated bit-rate. Motivated by this, and also under the assumption that the base station, where the reconstruction takes place, has the necessary memory and computational resources, we transfer the tasks of motion estimation and compensation at the decoder. Doing this, the motion compensation acts as a recursive refinement process which improves the quality of the reconstructed P-frames, combined with a super-resolution module to cope with their potential downsampling at the encoder. Fig. 2(b) shows the block-diagram of the proposed CVS decoder.

Starting with the fully-sampled frames, the reconstruction of a reference I-frame is performed using the inverse JPEG procedure, that is, the received bits are first decoded, followed by inverse quantization and inverse DCT (IDCT), resulting in an approximation $\hat{\mathrm{I}}$ (there will be always some loss of information due to the quantization).

Regarding the non-reference frames, the received bits for the current P-frame are first decoded and dequantized. Then, a reconstruction CS algorithm is applied on the dequantized compressed measurements yielding an estimate of the original frame $\hat{\mathrm{P}}$, by solving (2). Specifically, the iterative hard thresholding (IHT) algorithm (Blumensath and Davies, 2009) is employed for the reconstruction of each block of the current P-frame as follows,

$$\begin{aligned} \tilde{\mathbf{x}}_j^{n+1} &= \hat{\mathbf{x}}_j^n + \Phi^T(\mathbf{g}_j - \Phi\hat{\mathbf{x}}_j^n) && (6) \\ \hat{\mathbf{x}}_j^{n+1} &= \Psi_s^{-1}\big(\mathcal{T}\{\Psi_s(\tilde{\mathbf{x}}_j^{n+1})\}\big) \ , && (7) \end{aligned}$$

where $\mathcal{T}\{\cdot\}$ denotes a hard thresholding operator. The algorithm terminates when a predetermined number of iterations L_{max} has been reached, or if the error between successive iterations falls below a given threshold, $\|\hat{\mathbf{x}}_j^{n+1} - \hat{\mathbf{x}}_j^n\|_2 \leq \varepsilon$.

In the subsequent analysis, the DCT is used as the sparsifying transformation Ψ_s, while the threshold for the hard thresholding is given by (Donoho and Johnstone, 1994)

$$\rho_{Th} = \lambda\sigma\sqrt{2\log(B^2)} \ , \qquad (8)$$

where λ is a scaling factor varying usually between 3 and 5, and σ is the noise standard deviation, which is estimated using the mean absolute deviation of the transform coefficients $\tilde{\mathbf{w}}_j^{n+1} = \Psi_s(\tilde{\mathbf{x}}_j^{n+1})$ as follows,

$$\sigma = \frac{\mathrm{median}(|\tilde{\mathbf{w}}_j^{n+1}|)}{0.6745} \ . \qquad (9)$$

Iterative frame refinement: As it was mentioned in Section 3.2.1, the sampling rate $r = M/B^2$ is one of the key factors, which controls the tradeoff between the bit-rate and the achieved reconstruction quality. In order to satisfy the limitations of a lightweight remote imaging system the CVS encoder must operate at very low sampling rates. However, as the value of r decreases the quality of the reconstructed frames diminishes rapidly, as the subsequent experimental results reveal. This deterioration can be alleviated using an iterative refinement process based on motion estimation and compensation between the reconstructed I- and P-frames at the decoder. An inter-frame compensation method, such as in MPEGx, computes the prediction errors in the original space domain as follows,

$$\mathrm{R} = \mathrm{P} - \mathcal{M}\{\mathrm{I}\} \ , \qquad (10)$$

where R is the residual frame (or equivalently, the prediction error) and $\mathcal{M}\{\cdot\}$ denotes the motion compensation operator.

In our case, the decoder receives the encoded I-frames at full resolution, and the encoded compressed measurements, which are, nevertheless, related with the original (possibly downsampled) P-frames. One of the key properties of CS is that all the components of a CS measurement vector $\mathbf{g}_j$ are equally important. This means that the reconstruction performance can be practically unaffected even if some measurements are lost or disturbed (*e.g.,* due to channel errors). Because of this increased robustness when working directly with the CS measurements, the iterative refinement procedure is implemented in the CS measurements domain. By assuming for convenience that the frames involved in (10) are represented as a single block (reshaped as a column vector), and by substituting $\mathcal{M}\{\mathrm{I}\}$ with I_{MC}, the following expressions hold,

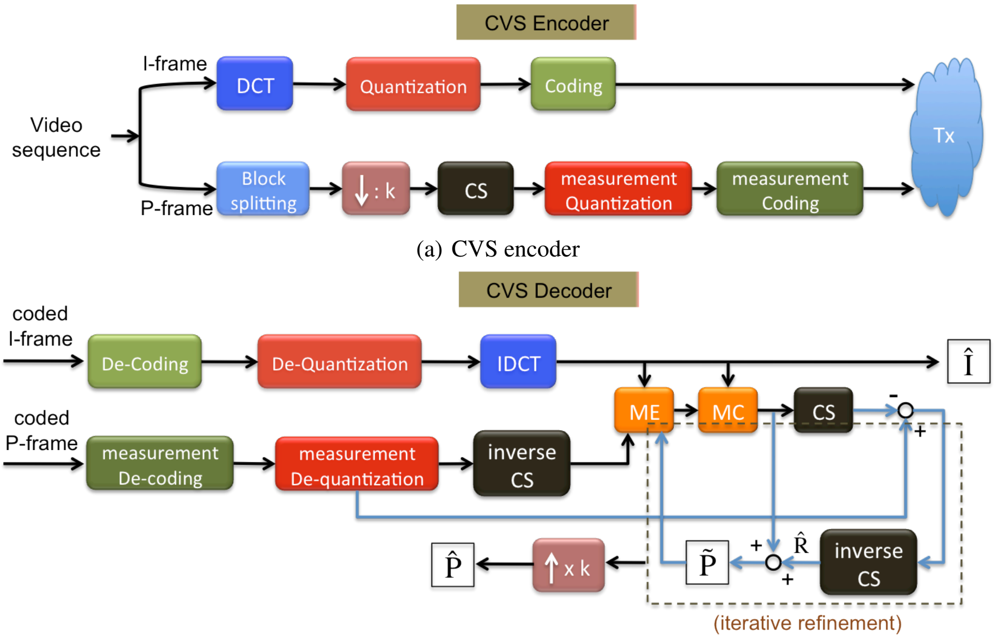
$$\begin{aligned} \hat{\mathrm{R}} &= \hat{\mathrm{P}} - \mathcal{D}\{\hat{\mathrm{I}}_{MC}\} \Rightarrow \\ \Phi\hat{\mathrm{R}} &= \Phi(\hat{\mathrm{P}} - \mathcal{D}\{\hat{\mathrm{I}}_{MC}\}) \Rightarrow \\ \Phi\hat{\mathrm{R}} &= \Phi\hat{\mathrm{P}} - \Phi\mathcal{D}\{\hat{\mathrm{I}}_{MC}\} \overset{(5)}{\Rightarrow} \\ \mathbf{g}_{error} &= \mathbf{g} - \mathbf{g}_{MC} \ . && (11) \end{aligned}$$

Notice that in the general case where the P-frames are downsampled at the encoder, then, the reconstructed I-frames must be also downsampled at the decoder before the CS measurements acquisition ($\mathbf{g}_{MC}$), so as to be consistent with the frame dimensions. This explains the presence of the downsampling operator $\mathcal{D}\{\cdot\}$ in the above equations.

Given the CS measurements $\mathbf{g}$ and constructing $\mathbf{g}_{MC}$, we generate the CS measurements of the residual via (11). Then, by applying the IHT (ref. (6)-(7)) on $\mathbf{g}_{error}$ we obtain an estimate of the residual frame ($\hat{\mathrm{R}}$), which can be used subsequently to refine the estimate of the current P-frame,

$$\hat{\mathrm{P}} = \mathcal{D}\{\hat{\mathrm{I}}_{MC}\} + \hat{\mathrm{R}} \ . \qquad (12)$$

The iterative refinement of the P-frames is summa-



(a) CVS encoder

(b) CVS decoder

Figure 2: Block-diagram of the proposed CVS architecture.

rized as follows:

$$\begin{aligned}
\mathbf{g}^n_{MC} &= \Phi\mathcal{D}\{\hat{\mathrm{I}}^n_{MC}\} \\
\mathbf{g}^n_{error} &= \mathbf{g} - \mathbf{g}^n_{MC} \\
\mathbf{g}^n_{error} &\xrightarrow{\text{IHT}} \hat{\mathrm{R}}^n \\
\hat{\mathrm{P}}^{n+1} &= \mathcal{D}\{\hat{\mathrm{I}}^n_{MC}\} + \hat{\mathrm{R}}^n \,.
\end{aligned}$$

The refinement process terminates when a predetermined number of iterations C_{max} has been reached, where a small number usually suffices to achieve a significant improvement (in the subsequent experiments it is set equal to 10). Notice also the presence of the superscript n in the above expressions associated with the motion compensated frame. This is justified by the fact that the motion estimation and motion compensation are performed after each update of the reconstructed P-frame ($\hat{\mathrm{P}}$) and thus, resulting in a different motion compensated frame ($\hat{\mathrm{I}}_{MC}$). The increased computational resources, which are now available at the decoder, enable us to use a more accurate sub-pixel motion estimation instead of using integer steps. In the following, the motion vectors are estimated with an accuracy of $1/4$ pixel .

A further improvement of the reconstruction quality can be achieved by scanning the current GOP ($\mathrm{IPP}\ldots\mathrm{PI}$) forward and backward, as shown in Fig. 3, since for the P-frames which are placed on the left of the GOP we expect the motion estimation to be more accurate, and thus resulting in sparser residuals, by employing the leftmost I-frame, while for the P-frames which are placed on the right an increased accuracy can be achieved using the rightmost I-frame. In our implementation a single scan in both directions is used.

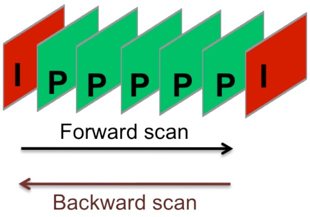


Figure 3: Bi-directional GOP scanning.

Super-resolution for P-frame resizing: As mentioned in Section 3.2.1, an optional downsampling of the P-frames can be used at the proposed CVS encoder to further reduce the transmission workload, when a limited bandwidth is available. In case of downsampling, the reconstructed P-frames must be restored in their original dimension. A straightforward approach would be to apply some of the well-established 2-dimensional interpolation techniques, such as a bilinear or a bicubic one. However, recent works (Freeman et al., 2002; Yang et al., 2008; Wang et al., 2011; Zhang et al., 2011) have shown that a super-resolution method results in images with superior quality. In the proposed CVS decoder we employ a dictionary-based super-resolution approach, as it is described in (Yang et al., 2008). In particular, the super-resolved image is generated from its low-resolution input, whose patches are assumed to have a sparse representation with respect to an overcomplete dictionary. Two coupled dictionaries are trained, with the first one ($\mathbf{D}_{HR}$) corresponding to the high-resolution patches and the second one ($\mathbf{D}_{LR}$) to the low-resolution patches. Then, the sparse representation of a low-resolution patch in the P-frame to be resized, in terms of $\mathbf{D}_{LR}$, will be used to construct the corresponding high-resolution patch from $\mathbf{D}_{HR}$.

In our proposed system, the initial training of $\mathbf{D}_{LR}$ and $\mathbf{D}_{HR}$ is performed using a set of arbitrary images. Notice also that the two dictionaries should be re-trained if the downsampling factor changes. As it was mentioned before, in the following evaluations a factor of 2 will be used. In order to increase the adaptivity of the trained dictionaries on the specific content of a given video sequence, we apply an updating process by incorporating the patches of the reconstructed I-frames, whose quality is superior in comparison to the quality of the reconstructed P-frames, since they are fully sampled at the encoder. However, we note that the updating phase using the current implementation requires an increased amount of time, which can be a drawback in a real-world scenario under potential time limitations. On the other hand, the design of fast and efficient dictionary updating methods is by no means a significant task, which can be the subject of a separate thorough study.

4 EXPERIMENTAL EVALUATION

In this section, the performance of the proposed CVS system is evaluated and compared with M-JPEG. For this purpose, three videos with distinct content are used[1], namely, i) the "Akiyo", which consists of a static background and a slowly moving foreground, ii) the "News", with motion in both the background and the foreground, and iii) the "Coastguard", which is characterized by a highly dynamic scene with complex motion compared with the other two sequences.

For the encoding of P-frames, GOPs of size 6 and blocks of size 16×16 are used, while the BWHT is chosen as a measurement matrix. The number of quantization levels varies from 2^6 to 2^8 (or equivalently, the number of quantization bits varies from 6 to 8) and the sampling rate is fixed at $r = 0.10$ (that is, for each block $M = 0.10 \cdot 16^2 = 26$ CS measurements are acquired). Regarding the refinement process at the decoder, the parameters which control its performance are set as follows: $\lambda = 3$, $L_{max} = 400$, $\varepsilon = 10^{-4}$, and $C_{max} = 10$. In order to achieve similar bit-rates for the M-JPEG, for a fair comparison, the value of the scaling factor S in (3) depends on the input video sequence. More specifically, the correspondence between the number of quantization bits, controlling the bit-rate of CVS, and the value of S for each video sequence are summarized in Table 1.

Table 1: Correspondence between the number of quantization bits (CVS) and the scaling factor S (M-JPEG).

	# Quantization bits			
Video sequence	6	7	8	
Akiyo	40	30	20	S
News	50	40	30	
Coastguard	35	27	20	

In the following, the reconstruction quality is measured in terms of the *structural similarity index* (SSI), which resembles more closely the human visual perception than the commonly used *peak signal-to-noise ratio* (PSNR). For a given image I and its reconstruction $\hat{\mathrm{I}}$ the SSI is defined by,

$$\mathrm{SSI} = \frac{(2\mu_{\mathrm{I}}\mu_{\hat{\mathrm{I}}} + c_1)(2\sigma_{\mathrm{I}\hat{\mathrm{I}}} + c_2)}{(\mu_{\mathrm{I}}^2 + \mu_{\hat{\mathrm{I}}}^2 + c_1)(\sigma_{\mathrm{I}}^2 + \sigma_{\hat{\mathrm{I}}}^2 + c_2)} , \qquad (13)$$

where μ_{I}, σ_{I} are the mean and standard deviation of the luminance of image I (similarly for $\hat{\mathrm{I}}$), $\sigma_{\mathrm{I}\hat{\mathrm{I}}}$ denotes the correlation coefficient of the two images, and c_1, c_2 stabilize the division with a weak denominator. In particular, when SSI is equal to 0 the two images are completely distinct, while when the two images are matched perfectly SSI is equal to 1.

Fig. 4 shows the reconstruction performance averaged over the first 50 frames of each video sequence, as a function of the required bit-rate (in Kbps), for the proposed CVS method and an M-JPEG coding scheme. As it can be seen, for the same bit-rates, our CS-based video compression method achieves a significant improvement of the frame reconstruction quality over the M-JPEG method. Most importantly, this improvement is obtained using an encoder with an even decreased computational complexity when compared with the M-JPEG encoder, since the compression of P-frames is performed via simple matrix-vector products (ref. (5)) instead of using a transform coding approach as the M-JPEG does.

For a visual inspection of the reconstruction quality for each one of the three video sequences, Fig. 5 shows the fifth frame of the original sequence, along with the frame reconstructed using M-JPEG, as well as with the proposed approach before and after applying the iterative refinement step. First, by comparing the third and the fourth columns, presenting the reconstructed frames by applying separate decoding and

[1] `http://media.xiph.org/video/derf/`

Table 2: SSI values between the original and the reconstructed frames shown in Fig. 5.

	M-JPEG	CS w/o MC	CS with MC
Akiyo	0.829	0.584	**0.871**
News	0.733	0.464	**0.768**
Coastguard	0.678	0.516	**0.711**

by employing motion estimation and compensation at the decoder, we can see the significant improvement we are able to achieve via the iterative motion compensation process. Moreover, the M-JPEG method is sensitive to blocking artifacts as the number of quantization levels decreases, as it can be seen in the images of the second column. The corresponding SSI values between the original frames and the M-JPEG reconstruction, along with the proposed approach without and with the use of motion compensation are shown in Table 2.

5 CONCLUSIONS

In the present work we introduced a compressive video sensing method for a lightweight remote imaging system. The limited memory, power, and bandwidth resources of these systems necessitate the design of a very simple encoder, while putting the main computational burden at the decoder. Motivated by the simplicity of M-JPEG, the proposed CVS encoder is able to achieve low bit-rates with a reduced computational complexity, by combining a DCT-based transform coding applied on the reference I-frames with a CS-based compression applied on the (possibly downsampled) non-reference P-frames. At the decoder, the inverse processes are followed to reconstruct the I and P frames. However, the main drawback is the decreased reconstruction performance of the P-frames when we work at very low sampling rates. This problem is alleviated by applying an iterative motion compensation at the decoder, acting as a refinement step. Finally, when the acquired P-frames are downsampled at the encoder, a dictionary-based super-resolution method is employed at the decoder to restore the reconstructed frames in their original dimension. As a general conclusion, the proposed CVS system yielded a superior reconstruction quality compared with M-JPEG, especially at low bit-rates.

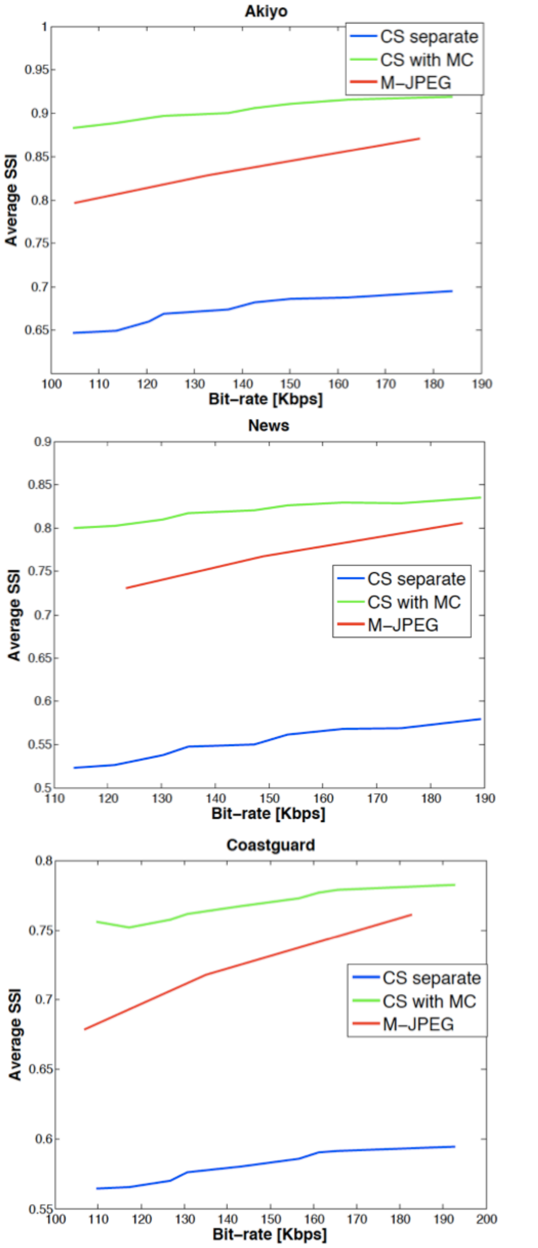


Figure 4: Comparison of reconstruction performance as a function of the bit-rate between CVS and M-JPEG.

Further improvements can be achieved by implementing a more efficient CS reconstruction method, resulting in an increased quality of the initially reconstructed P-frames at very low sampling rates. Moreover, with the current implementation, the dictionary-based super-resolution method cannot be applied efficiently in a real-time scenario, because of the delay in updating the dictionary with previously decoded I-frames. Thus, the design of new techniques for fast dictionary update is necessary so as to enable the use of the proposed method in a real-world application with time limitations at the decoder. When working at very low sampling rates, we do not expect to achieve the superior performance of an MPEGx-like profile, which is based on the increased sparsity of the residual frames exploited at the encoder. How-

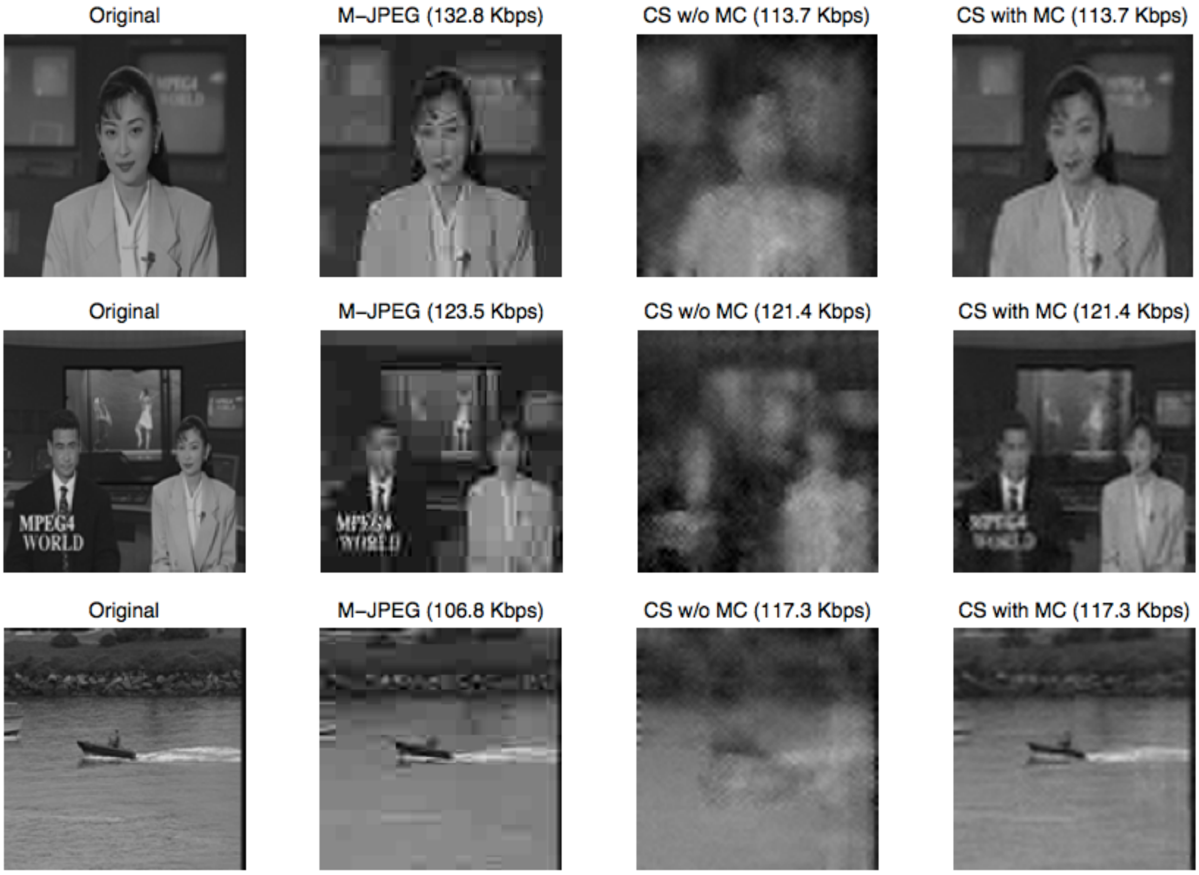


Figure 5: Visual inspection of the reconstruction performance for M-JPEG and the proposed CVS scheme.

ever, as the experimental results revealed, the reconstruction quality of the proposed lightweight remote imaging system is high enough so as to be effective in performing tasks such as detection and classification, which are also of importance in several surveillance applications, and require a more thorough study.

ACKNOWLEDGEMENTS

This work is supported by CS-ORION Marie Curie Industry-Academia Partnerships and Pathways (IAPP) project funded by the European Commission in FP7 (PIAP-GA-2009-251605).

REFERENCES

Blumensath, T. and Davies, M. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274.

Candés, E., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, 52:489–509.

Do, T., Chen, Y., Nguyen, D., Nguyen, N., Gan, L., and Tran, T. (2009). Distributed compressed video sensing. In *43rd Annual Conf. Inf. Sci. and Sys. (CISS'09)*, Baltimore, MD.

Do, T., Tran, T., and Gan, L. (2008). Fast compressive sampling with structurally random matrices. In *IEEE Int. Conf. on Acoust., Speech and Sig. Proc. (ICASSP'08)*, Las Vegas, NV.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Freeman, T., Jones, T., and Pasztor, E. (2002). Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65.

Jung, H. and Ye, J.-C. (2010). Motion estimated and compensated compressed sensing dynamic magnetic resonance imaging: What we can learn from video compression techniques. *Intl. J. of Imaging Systems and Technology*, 20(2):81–98.

Kang, L.-W. and Lu, C.-S. (2009). Distributed compressive video sensing. In *IEEE Int. Conf. on Acoust., Speech and Sig. Proc. (ICASSP'09)*, Taipei.

Prades-Nebot, J., Ma, Y., and Huang, T. (2009). Distributed video coding using compressive sampling. In *Picture Coding Symp. (PCS'09)*, Chicago, IL.

Skretting, K., Husøy, J. H., and Aase, S. O. (1999). Improved Huffman coding using recursive splitting. In *Norwegian Signal Proc. Symp.*, Norway.

Stanković, V., Stanković, L., and Cheng, S. (2008). Compressive video sampling. In *European Sig. Proc. Conf. (EUSIPCO'08)*, Lausanne.

Wang, P., Hu, X., Xuan, B., Mu, J., and Peng, S. (2011). Super resolution reconstruction via multiple frames joint learning. In *Intl. Conf. Multimedia and Signal Proc. (CMSP'11)*, Guilin, Guangxi.

Yang, J., Wright, J., Huang, T., and Ma, Y. (2008). Image super-resolution as sparse representation of raw image patches. In *Computer Vision and Pattern Rec. (CVPR'08)*, Anchorage, AK.

Zhang, H., Zhang, Y., and Huang, T. (2011). Efficient sparse representation based image super resolution via dual dictionary learning. In *Intl. Conf. Multimedia and Expo (ICME'11)*, Barcelona.