Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

A new compressive sensing video coding framework based on Gaussian mixture model

Xiangwei Li[a,b], Xuguang Lan[a,*], Meng Yang[a], Jianru Xue[a], Nanning Zheng[a]

[a] *Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, PR China*
[b] *Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi 710049, PR China*

ARTICLE INFO

Keywords:
Compressive sensing video
Gaussian mixture model
Lossy compression
Video coding
Product vector quantizer

ABSTRACT

In this paper, we specifically design an efficient compressive sensing video (CSV) coding framework for the CSV system, by considering the distribution characteristics of the CSV frame. To explore the spatial redundancy of the CSV, the CSV frame is first divided into blocks and each block is modeled by a Gaussian mixture model (GMM), and then it is compressed by a product vector quantization. We further explore the temporal redundancy of the CSV by encoding the adjacent CSV frames by the differential pulse code modulation technique and the arithmetic encoding technique. Experiment results show that the proposed CSV coding solution maintains low coding complexity, which is required by the CSV system. Meanwhile, it achieves significant BD-PSNR improvement by about 7.13–11.41 dB (or equivalently 51.23–66.96% bitrate savings) compared with four existing video coding solutions, which also have low computational complexity and suit for the CSV system.

1. Introduction

Digital video acquisition and compression is an important research topic and has been well studied in the past decades. A classical video system often comprises two steps: capturing each frame of the video scene at certain temporal/spatial resolution by a video camera in acquisition process, and then massively dumpling the temporal/spatial redundancy information of the captured frames in a compression process. According to the Shannon-Nyquist sampling theorem [1], the temporal/spatial sampling rate of the video acquisition needs to be at least twice as high as the highest temporal/spatial frequency of the video scene so that it can be reconstructed accurately. The cost and computational complexity of the video system rises dramatically with the increase of the temporal/spatial resolution of the video camera. Thus it may not be suitable to the requirements of many modern applications with computational resource and energy limitations, e.g., wireless video sensor networks [2,3], wireless image/video broadcasting [4,5], aerial photography [6,7] and high-speed photography [8–10]. Recently, compressive sensing (CS) has emerged as an effective sampling theory for the acquisition of signals which can be sparsely represented [11,12]. Based on CS theory, a new architecture called compressive sensing video (CSV) system has been proposed for low complexity video acquisition, which enables acquiring and compressing video scene simultaneously [13–19]. Different from classical video system, the CSV system [16–19] is able to acquire the high-frame-rate video with a low-frame-rate camera, and its video acquisition has a lower computational complexity.

References [16–18] are typical CSV cameras for the CSV acquisition, which enables temporal compression while video acquisition. A CSV system mainly contains three steps: CSV acquisition, CSV coding (encoding/decoding) and CSV reconstruction, as shown in Fig. 1.

1.1. CSV acquisition process

In the CSV system with temporal compression ratio T, each of the T input video frames is first modulated through a coded aperture with random mask, and then the CSV camera sums the T modulated frames and produces one CSV frame as shown in Fig. 1. Let n_x and n_y represent the spatial resolution of the input video frame in the horizontal and vertical direction, respectively. Let the video cube $\mathbf{X} \in \Re^{n_x \times n_y \times T}$ denote the original T input video frames and $x_{i,j,k}$ denote (i,j)-th pixel in the k-th input video frame. Let $\mathbf{H} \in \Re^{n_x \times n_y \times T}$ denote the T random mask cube for modulating the input video frames, while $h_{i,j,k} \in \{0, 1\}$ is (i,j)-th component in the k-th random mask. Then the CSV frame $\mathbf{Y} \in \Re^{n_x \times n_y}$ is acquired as follows.

$$y_{i,j} = \sum_{k=1}^{T} x_{i,j,k} h_{i,j,k} + e_{i,j}, \ \forall \ i = 1, \ldots, n_x; \ j = 1, \ldots, n_y;$$

where $y_{i,j}$ is (i,j)-th pixel in the acquired CSV frame $\mathbf{Y}$. $h_{i,j,k}$ is assumed to be fixed in space during each time period of the input video frame. $e_{i,j}$ is

[*] Corresponding author.
E-mail addresses: xglan@mail.xjtu.edu.cn, xuguang.lan@gmail.com (X. Lan).

http://dx.doi.org/10.1016/j.image.2017.03.009
Received 29 April 2016; Received in revised form 16 February 2017; Accepted 13 March 2017
Available online 18 March 2017

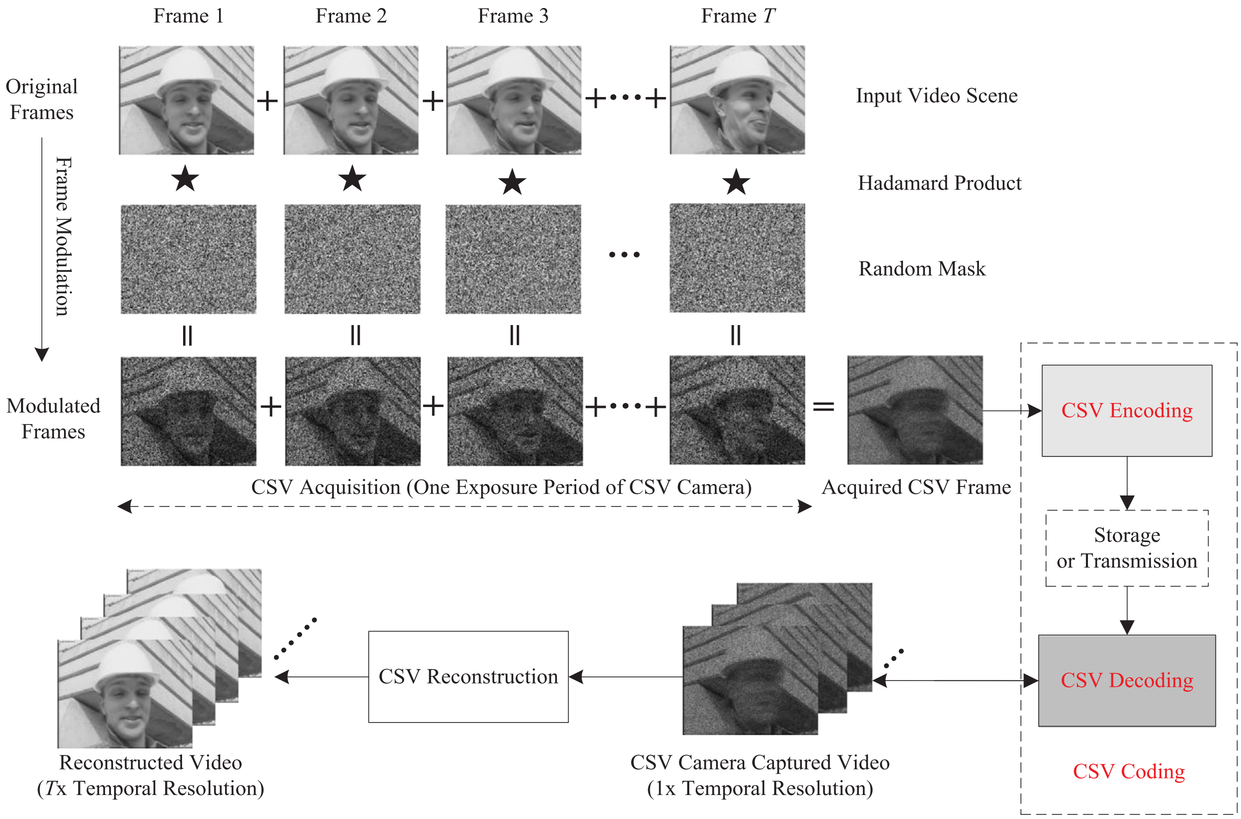

Fig. 1. The architecture of a typical CSV system which enables temporal compression while acquisition.

the noise while measuring the CSV frame. Thus each pixel value $y_{i,j}$ is a weighted sum of the pixels at the same spatial position in $\mathbf{X}$.

In order to reduce the computational complexity of the CSV system, we apply the block based CSV acquisition and reconstruction [24]. We divide the video cube $\mathbf{X}$ into L non-overlap patches of the size $s \times s \times T$, and the random mask cube $\mathbf{H}$ are also divided into the same size. Therefore the acquired CSV frame $\mathbf{Y}$ is divided into a set of L non-overlap blocks of the size $s \times s$. Let $\mathbf{x}_i \in \Re^{s^2T}$ and $\mathbf{y}_i \in \Re^{s^2}$ represent the vectorized i-th patch of the video cube $\mathbf{X}$ and i-th block of CSV frame $\mathbf{Y}$, respectively. The block based CSV acquisition process for each block $\mathbf{y}_i$ can be expressed as

$$\mathbf{y}_i = \mathbf{h}_i\mathbf{x}_i + \varepsilon_i, \forall\ i = 1, \ldots, L \tag{1}$$

where $\mathbf{h}_i \in \{0, 1\}^{s^2 \times s^2T}$ denotes the matrix formed by i-th patch of the random mask cube $\mathbf{H}$. The vectorized noise for measuring the i-th patch $\mathbf{x}_i$ is denoted by $\varepsilon_i \in \Re^{s^2}$, and in this work we assume that $\varepsilon_i \sim N(\varepsilon_i|0, \mathbf{R})$, where $\mathbf{R} \in \Re^{s^2 \times s^2}$ is the covariance matrix of the noise.

We assume that $\mathbf{x}_i$ is drawn from a GMM with K Gaussian components as in [24].

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} w_k N(\mathbf{x}_i|\mathbf{m}_k, \mathbf{S}_k)$$

where $\mathbf{m}_k \in \Re^{s^2T}$, $\mathbf{S}_k \in \Re^{s^2T \times s^2T}$ and w_k, ($w_k > 0$ and $\sum_{k=1}^{K} w_k = 1$) denote the mean, the covariance matrix, and the weight of the k-th Gaussian component in the GMM, respectively.

We then model the probability density function (PDF) of $\mathbf{y}_i$ in (1) by a GMM with K Gaussian components (p-dimensional Gaussian distribution $p = s^2$) [24,28].

$$p(\mathbf{y}_i) = \sum_{k=1}^{K} w_k N(\mathbf{y}_i|\mu_k, \Sigma_k) \tag{2}$$

where $\mu_k = \mathbf{h}_i\mathbf{m}_k \in \Re^p$, $\Sigma_k = \mathbf{h}_i\mathbf{S}_k\mathbf{h}_i^T + \mathbf{R} \in \Re^{p \times p}$ and w_k, ($w_k > 0$ and $\sum_{k=1}^{K} w_k = 1$) denote the mean, the covariance matrix, and the weight of the k-th Gaussian component in the GMM, respectively. $\mathbf{R}$ is the precision matrix which is the noise covariance of ε_i. One can estimate the GMM parameters in (2) by EM algorithm [25] based on a training set of blocks from the CSV frames. The number of Gaussian components K is a design parameter in the GMM parameters estimation. The main advantage of using GMM for modeling the CSV patch $\mathbf{y}_i$ is that it benefits a low computational lossy compression algorithm design based on GMM.

1.2. CSV coding process

Since the CSV camera is always implemented on analog device and the acquired CSV frames are real value, a CSV coding process is required for compressing the CSV frames for further storage or transmission. In CSV coding, the acquired CSV frames are compressed and encoded into a set of binary codes, and these binary codes can be conveniently stored or transmitted to the receiver for CSV reconstruction. At present, due to the low computational complexity requirement of the video acquisition and compression processes in the CSV system, existing coding solutions which suit for the CSV system are: 1) JPEG [20] based solution MJPEG-differential pulse code modulation (MJPEG-DPCM) [21], 2) uniform scalar quantization (USQ) based solution USQ-DPCM [22], 3) H.264/AVC Intra codec [23], and 4) HEVC Intra codec [41]. There are also some other efficient coding solutions like the H.264/AVC inter codec [23] and HEVC Lowdelay codec [41]. They may not be suit for the CSV coding application in CSV system due to the high computational complexity. However, many fast algorithms have been proposed for video coding [44–46]. However, these methods still rely on motion estimation (ME), which is not suitable for resource limited CSV system. Thus, these ME-based video coding methods are not taken into consideration in this paper. Although the above mentioned four coding solutions have a low computational complexity, they do not explore the distribution characteristics of the CSV frames, thus the compression efficiency will be degraded. The distribution of the discrete cosine transform (DCT) coefficients of the CSV frame is very different from that of natural video frame, as shown