J. Vis. Commun. Image R. 47 (2017) 23–35

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Per-pixel mirror-based method for high-speed video acquisition ☆

J.A.S. Lima [a,*], C.J. Miosso [b], M.C.Q. Farias [c]

[a] Department of Computer Science, University of Brasilia, Brasilia, Brazil
[b] University of Brasilia at Gama, Gama, Brazil
[c] Department of Electrical Engineering, University of Brasilia, Brasilia, DF, Brazil

ARTICLE INFO

Article history:
Received 5 August 2016
Revised 12 April 2017
Accepted 8 May 2017
Available online 10 May 2017

Keywords:
Compressive sensing
Computational photography
High-speed imaging

ABSTRACT

High-speed imaging requires high-bandwidth, fast image sensors that are generally only available in high-end specialized cameras. Nevertheless, with the use of compressive sensing theory and computational photography techniques, new methods emerged that use spatial light modulators to reconstruct high-speed videos with low speed sensors. Although these methods represent a big step in the field, they still present some limitations, such as low light efficiency and the generation of measurements with time dependency. To tackle these problems, we propose a per-pixel mirror-based acquisition method that is based on a new kind of light modulator. The proposed method uses moving mirrors to scramble the light coming from different positions, thus ensuring better light efficiency and generating time independent measurements. Our results show that the proposed method and its variations perform better than methods available in the literature, generating videos that are less noisy and that display better content separation.

1. Introduction

Several applications in science and industry require the acquisition of videos with time resolutions that are from tens to a few hundred times higher than those of typical consumer cameras. Examples include the study of blood flow in cellular structures, imaging of combustion processes, evaluation of precise movements in biomechanical structures, analysis of the mechanics of novel fluids, detection of movements causing structural fatigue, visual microphones, etc. [1–3]. In these applications, the most common solution for acquiring high-speed videos involves using special high-speed video cameras, which have highly sensitive sensors capable of acquiring thousands of frames per second (FPS). Unfortunately, the cost of such sensors still prevents their use in most applications. An alternative approach uses a synchronized array of cameras [4,5], which is also an expensive solution because it requires, typically, 64 to 128 cameras.

A more recent solution consists of using compressive sensing to reconstruct high-speed videos from measurements obtained using sub-60 FPS cameras [6–12]. Most of these solutions use shutters to scramble the light rays that reach the sensors. One of the devices that uses this approach is the *flutter shutter*, which divides the camera frame time into short-term periods, during which sensors can either receive light or not. Some commercial cameras have implementations of the flutter shutter device [12] that can be used as a compressive sensing acquisition method for reconstructing high-speed periodic scenes [10] and videos with no motion restrictions [11].

Another device that implements this approach is the *per-pixel shutter*, which selects short periods for light exposure using an independent control for each camera sensor. Some compressive sensing high-speed video reconstruction methods [13,14,6] are based on per-pixel shutter devices. Although methods that use per-pixel shutters provide better results than methods that use flutter shutters, per-pixel shutters are not currently implemented in commercial cameras. Experimental implementations of per-pixel shutters are performed by attaching an additional optical system to the camera.

Both flutter shutter and per-pixel shutter methods have common drawbacks. First, these methods discard around 50% of the light, reducing light efficiency and consequently image quality after reconstruction. Also, for both methods, the light captured at different time instants is integrated into a single pixel, what means that measurements are time dependent. Therefore, at the reconstruction stage it is difficult to separate the information coming from different time instants. If the goal of an application is to reconstruct high-speed videos, these drawbacks have to be addressed.

☆ This paper has been recommended for acceptance by M.T. Sun.
* Corresponding author.
E-mail address: jonathanalis@gmail.com (J.A.S. Lima).

http://dx.doi.org/10.1016/j.jvcir.2017.05.004
1047-3203/

Considering the high cost of high-speed cameras and the limitations and drawbacks of acquisition methods that use low-cost cameras, in this paper we propose a new acquisition method for compressive sensing reconstruction of high-speed videos. The proposed method, called *per-pixel mirror-based* (PPM) acquisition method, was described in its initial form in an earlier publication [15], which contains preliminary results. PPM is based on an acquisition strategy that uses a set of moving mirrors to redirect the light to certain pixels. The method does not discard any light and it separates the temporal information, generating time-independent measurements. In this paper, we detail several variations of the proposed method and compare them with currently available acquisition methods. We tested the proposed method for still images, synthetic videos and natural videos.

The remaining parts of the paper are organized as follows. Sections 2 and 3 briefly describe the compressive sensing theory and the currently available video compressive sensing acquisition methods. Section 4 describes the proposed method, whereas Section 5 presents the simulation results. Finally, Section 6 presents our conclusions and future works.

2. Compressive sensing

Let N be the dimension of the signal $\mathbf{x}$ to be acquired. This signal $\mathbf{x}$ is said to be sparse in the $\boldsymbol{\Psi}$ domain if only a few projections of $\mathbf{x}$ into the bases of $\boldsymbol{\Psi}$ are non-zero. If the total number of non-zero projections is $K < N$, then the signal is said to be K-sparse. The theory of compressive sensing [16–20] allows one to acquire and to reconstruct a sparse signal with a smaller number of measurements than the number of samples required by the Nyquist rate. More specifically, compressive sensing allows one to take only M linear measurements from $\mathbf{x}$, where $M \ll N$, and nonetheless all N components of $\mathbf{x}$ theoretically without any error.

Suppose that $\mathbf{x}$ is K-sparse in the basis $\boldsymbol{\Psi}$, with $K \ll N$. Let $\mathbf{y}$ be the vector of linear projections of $\mathbf{x}$ into M vectors $\boldsymbol{\Phi}_i$ $(i = [1, \cdots, M])$. If $\boldsymbol{\Phi}$ is the $M \times N$ matrix in which each row is one of those distinct M vectors, then

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{x} = \boldsymbol{\Phi}\boldsymbol{\Psi}\mathbf{s} = \boldsymbol{\Theta}\mathbf{s}, \tag{1}$$

where $\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ is an $M \times N$ matrix and $\mathbf{s} = \boldsymbol{\Psi}^{-1}\mathbf{x}$ is the sparse representation of $\mathbf{x}$ in the domain defined by $\boldsymbol{\Psi}$.

Note that (1) represents the relationship between the available measurements, $\mathbf{y}$, and the desired signal, $\mathbf{x}$, which is at first unknown. The acquisition process must provide the components of $\mathbf{y}$, whereas a reconstruction procedure must provide $\mathbf{x}$ based on $\mathbf{y}$ and on some signal properties. In Section 4, we detail how we obtain $\mathbf{y}$ in our formulation, and we provide the mathematical modeling that relates $\mathbf{y}$ to $\mathbf{x}$ in our proposed methods. In other words, we describe our particular measurement matrix $\boldsymbol{\Phi}$. For now, we explain the principles based on which we compute $\mathbf{x}$ from $\mathbf{y}$.

Regarding the reconstruction stage, since (1) is an underdetermined system, there are generally infinite signals $\mathbf{s}'$ that satisfy $\boldsymbol{\Theta}\mathbf{s}' = \mathbf{y}$. Amongst all solutions, we search for the sparsest one. Two properties must be satisfied so that this procedure is stable: the restricted isometry property (RIP) [21] and the incoherence property [22]. According to the RIP, $\boldsymbol{\Theta}$ should roughly preserve the lengths of the K-sparse vectors, within a predefined tolerance [22]. Incoherence, on the other hand, requires that the rows of $\boldsymbol{\Phi}$ do not have a sparse representation in the $\boldsymbol{\Psi}$ domain [20].

Once the RIP and the incoherence are satisfied, reconstructing a K-sparse signal using these M measurements corresponds to finding the sparsest signal that satisfies these measurements [20]. Directly searching for the sparsest solution, however, is generally unfeasible, as it corresponds to a combinatorial optimization problem [21]. Practical solutions use optimization algorithms based on ℓ_1- or ℓ_p-minimization problems. Such alternatives reduce the computational complexity, at the cost of increasing the number of linear measurements [16].

In image reconstruction, the ideal problem of finding the sparsest solution by ℓ_0-minimization can be replaced by the Total Variation (TV) minimization [23,24], which is the chosen method in this paper. The TV of an $N_1 \times N_2$ image $\mathbf{s}'$, denoted as $\|\mathbf{s}'\|_{TV}$, is related to the horizontal ($\mathbf{G_v}\mathbf{s}'$) and vertical ($\mathbf{G_h}\mathbf{s}'$) gradients of $\mathbf{s}'$. If we use ℓ_1 (approach that was already applied in compressive video sensing [25]) to combine $\mathbf{G_h}\mathbf{s}'$ and $\mathbf{G_v}\mathbf{s}'$, TV can be defined as:

$$||\mathbf{s}'||_{TV} = \sum_{i=2}^{N_1}\sum_{j=2}^{N_2}|\mathbf{G_v}\mathbf{s}'(i,j)| + |\mathbf{G_h}\mathbf{s}'(i,j)|. \tag{2}$$

The use of TV minimization for image reconstruction is based on the idea that the discrete gradient of natural images tends to generate sparser images. TV can be then viewed as the ℓ_1 of the image in a sparse domain [26]. For image reconstruction, the optimization problem can be described as

$$\hat{\mathbf{s}} = \operatorname{argmin}_{\mathbf{s}'}(||\mathbf{s}'||_{TV}), \text{ such that } \mathbf{y} = \boldsymbol{\Theta}\mathbf{s}' = \boldsymbol{\Phi}\boldsymbol{\Psi}\mathbf{s}'. \tag{3}$$

In this equation, $\boldsymbol{\Phi}$ is the acquisition matrix and $\boldsymbol{\Psi}$ is the transform basis.

Note that, in our proposed methods, once we obtain the measurements $\mathbf{y}$ described in (1), we apply a numerical optimization procedure such as (2) to compute the desired image $\mathbf{x}$. Our main contribution is a novel formulation for obtaining the measurements $\mathbf{y}$, as described in Section 4, which improves the objective quality in high-speed video reconstruction.

In our approach, we take the measurements in the spatio-temporal domain, i.e., the pixel domain. In other words, $\boldsymbol{\Psi}$ is the identity matrix. Taking $||\mathbf{s}||_{TV} = ||\mathbf{D}_i\mathbf{s}||_1$ and making $\mathbf{s}'' = \mathbf{D}_i\mathbf{s}'$, the optimization problem is then given by

$$\hat{\mathbf{s}} = \operatorname{argmin}_{\mathbf{s}''}(||\mathbf{s}''||_1) \text{ such that } \mathbf{y} = \boldsymbol{\Phi}\mathbf{D}_i^{-1}\mathbf{s}''. \tag{4}$$

This method is equivalent to the ℓ_1 minimization problem, where the sparsifying transform is the finite differences operator.

We use TV reconstruction techniques for all tested acquisition methods tested (see Section 5). We refer to the TV reconstruction in 2 spatial dimensions (finite differences of lines and rows in an image) as TV2D and to the TV reconstruction in 2 spatial and 1 temporal dimensions (finite differences among lines, rows, and subsequent subframes) as TV3D. TV2D takes advantage of spatial redundancies, while TV3D takes advantage of spatial and temporal redundancies.

3. Current video acquisition methods

Traditional video acquisition produces frame pictures containing the light captured by sensors during the exposition time. If a long exposure is used, frame pictures may appear blurry for video scenes with a lot of movement (motion blur). However, if we choose a very short exposure time, the picture may appear too dark or noisy, since the amount of light that reaches each sensor is reduced because of the shorter time interval. In order to use compressive sensing reconstruction for obtaining videos with a higher spatial or temporal resolution, we have to linearly combine the scene samples. Without linearly combining samples in a proper way, we cannot guarantee that the incoherence property would be satisfied [27].

In many applications, a higher control of the light flow is desirable. For example, in applications like deblurring [28] or compressive sensing video acquisition [11,29,10], it is necessary to start and finish exposure several times during one frame interval. A