◆ Surveillance Video Analysis Using Compressive Sensing With Low Latency

Hong Jiang, Songqing Zhao, Zuowei Shen, Wei Deng, Paul A. Wilford, and Raziel Haimi-Cohen

We propose a method for analysis of surveillance video by using low rank and sparse decomposition (LRSD) with low latency combined with compressive sensing to segment the background and extract moving objects in a surveillance video. Video is acquired by compressive measurements, and the measurements are used to analyze the video by a low rank and sparse decomposition of a matrix. The low rank component represents the background, and the sparse component, which is obtained in a tight wavelet frame domain, is used to identify moving objects in the surveillance video. An important feature of the proposed low latency method is that the decomposition can be performed with a small number of video frames, which reduces latency in the reconstruction and makes it possible for real time processing of surveillance video. The low latency method is both justified theoretically and validated experimentally. © 2014 Alcatel-Lucent.

Introduction

In a surveillance network, cameras transmit surveillance videos to a processing center where the video streams are processed and analyzed. The ability to detect moving objects in a scene quickly and automatically is of particular interest in surveillance video processing. Detection of moving objects is traditionally achieved by background subtraction methods [1, 2] which segment the background from moving objects in a sequence of surveillance video frames. The technique described in [1] stores, for each pixel, a set of values taken in the past in the same location or neighborhood. It then compares this set to the current pixel value in order to determine whether that pixel belongs to the background and adapts the model by randomly choosing which values to substitute from the background model. The mixture of Gaussians technique [22] assumes that each pixel has a distribution that is a sum of Gaussians and the background and foreground are modeled by the different size of the Gaussians. In low rank and sparse decomposition (LRSD) [6], the background is modeled by a low rank matrix, and the moving objects are identified by a sparse component.

These traditional background subtraction techniques apply to video in the pixel domain, and require the pixels in a surveillance video to be captured, transmitted, and analyzed. The ever-growing number of surveillance cameras generates an enormous amount of data that needs to be transported over the network. There is a high risk that congestion in the network will prevent timely detection of

Bell Labs Technical Journal 18(4), 63–74 (2014) © 2014 Alcatel-Lucent. Published by Wiley Periodicals, Inc.
Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/bltj.21646

moving objects. Addressing the congestion problem by conventional video coding methods makes the video transmission highly sensitive to varying channel conditions and significantly increases the complexity of the cameras as well as the processing center. Therefore, it is highly desirable to have a network of cameras in which each camera transmits a small amount of data with enough information for reliable detection and tracking of moving objects. Compressive sensing [7] allows us to achieve this goal. Compressive sensing has previously been used for both video processing [6, 7] and background subtraction [8, 15].

In [15], an LRSD of a matrix is used in processing compressive measurements to segment the background and extract moving objects. The method described in [15], which was motivated by the work in [6], assumes that the surveillance video is comprised of a low rank component (background) and a sparse component (the moving objects) which is possibly in a tight wavelet frame domain. Therefore, the background subtraction becomes part of the reconstruction. Furthermore, the reconstruction in [15] takes advantage of the knowledge that the video has a background, which helps to reduce the number of measurements required.

Since the method used in [15] reconstructs the background without a training process, compressive measurements from a large number of frames are needed in order to recover the background properly. Typically, the number of frames used in the reconstruction process is on the order of a hundred frames, representing a few seconds of video in real time. This causes an inherent latency of a few seconds, which is independent of and added to the computational time. Such latency may not be appropriate for real time applications.

In this paper, we propose a low latency LRSD method. This method extends the framework of [15] to reduce the latency needed in the reconstruction process. As in [15], segmentation of background is performed by using an LRSD of the matrix. However, in this paper, the low rank matrix is augmented with known background frames. The background frames may be learned via a training process, for example, by using the methods in [15] and [23]. By using the augmented low rank matrix, the reconstruction by LRSD can be carried out with compressive measurements from a few video frames, as few as one frame. In other words, as soon as the measurements from one video frame are available, we can start processing the measurements to reconstruct the background and to compute the silhouette of the moving objects in that frame. Therefore, the method presented in this paper paves the way for real time processing of compressive sensed surveillance video by using compressive LRSD.

The method we propose removes a fundamental barrier to real time processing of surveillance video in the methods that use LRSD, such as those in [6] and [15], by relaxing the requirement of the number of frames needed in the reconstruction process.

Panel 1. Abbreviations, Acronyms, and Terms

2D—Two-dimensional
ADM—Alternating Direction Method
JPEG—Joint Photographic Experts Group
LRSD—Low rank and sparse decomposition
MPEG—Moving Picture Experts Group
PCA—Principle component analysis
RGB—Red, green, blue

Related Work

We make a few notes to compare this paper with other work in the literature. The novelty of this paper is the low latency for LRSD with compressive sensing. It has been demonstrated in [15] that LRSD with compressive sensing achieves what the traditional methods cannot. For example, in the Daniel video sequence [11], a sudden illumination change in the background was falsely detected as a moving object in the foreground when using the principal component analysis (PCA) method [11], while in [15], where LRSD was used, the event did not register. This

demonstrates the advantage of LRSD. Furthermore, traditional background subtraction methods such as [1, 2, 22] do not use compressive sensing, so video data must be processed in the pixel domain, not in the compressed domain. Although compressive sensing has been used in video processing, existing compressive video sensing methods reconstruct the video frames, but do not segment the background and foreground, thus requiring the use of an additional pixel domain background subtraction method for background segmentation. Our method performs the background segmentation as part of the reconstruction. This paper differs from [15] in that a large number of frames must be used to reliably detect moving objects in [15], which results in a long latency, making real-time processing difficult. This paper is an improvement over [15] because of low latency achieved by processing a small number of frames at a time, which makes it viable for real time processing. Reference [23] is a companion paper to this work for the reason that this paper establishes a theoretical basis for low latency LRSD by assuming that the background is known, while [23] relies on this theoretical basis and develops an algorithm to adaptively train the background model. Therefore, this work is a theoretical justification for the development in [23], while [23] in turn provides more experimental results to validate the theory in this work.

The paper is organized as follows. We will introduce notations and review the framework for reconstruction of compressively-sensed video using LRSD. We next describe a low latency method for LRSD, and follow with a theoretical justification. We also report results from numerical experiments.

Low Rank and Sparse Decomposition

In this section, we introduce notations and review the work of [15]. In analysis, we will treat video as black and white, having only the luminance component. Color video can be treated separately for each red, green, blue (RGB) component, or jointly as discussed in [15]. The framework for analyzing surveillance video using compressive sensing is shown in **Figure 1**. Network transmission of compressively sensed video can be found in, for example, [16] and [19].

Compressive Measurements

We consider a video volume of a video sequence consisting of a number of consecutive frames. Let $x_j \in \Re^n$ be a vector formed from the pixels of a sub-region in frame j of the video sequence, for $j = 1, ..., J$, where J is the total number of frames and n is the total number of pixels in the sub-region. Let $X = [x_1, ..., x_J] \in \Re^{n \times J}$ be the matrix of dimension $n \times J$, the columns of which are the pixels in the video volume. The total number of entries in X is $N = nJ$.

Let ϕ be an $M \times N$ measurement matrix with M rows and N columns, where $M < N$. The measurement matrix ϕ may be chosen as a random matrix such as a randomly permutated Walsh-Hadamard matrix [16, 17]. Let $\phi = [\phi_1, ..., \phi_J]$, where $\phi_j \in \Re^{M \times n}$ is a sub-matrix of dimension $M \times n$.

The compressive measurements of the video volume are defined as

$$y = \phi \circ X \triangleq \sum_{j=1}^{J} \phi_j x_j, \tag{1}$$

where y is a vector of length M which is the number of measurements and is much less than the total number of pixels of the video volume, N.

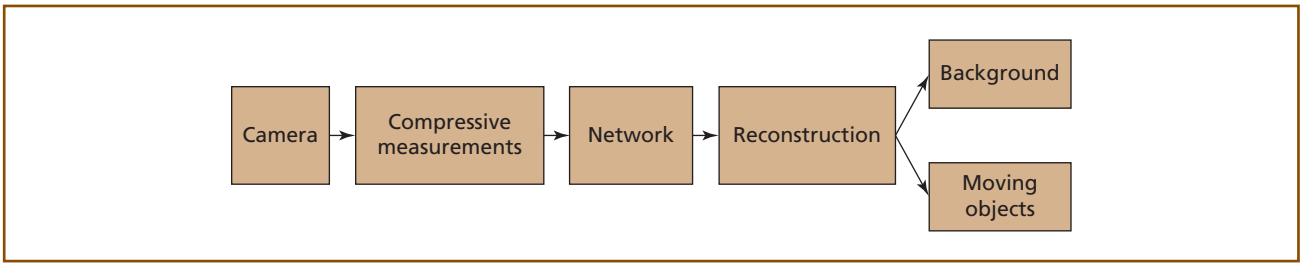


Figure 1.
Analysis of surveillance video using compressive sensing.

Reconstruction

Given the measurement vector y, the video volume X can be reconstructed by solving the following minimization problem:

$$\min_{X_1, X_2} \mu_1 \|X_1\|_* + \mu_2 \|W_1 X_1\|_1 + \mu_3 \|W_2 X_2\|_1,$$
$$\text{such that} \quad y = \phi \circ (X_1 + X_2), \tag{2}$$

$$X = X_1 + X_2. \tag{3}$$

In equation 2, $\|A\|_*$ is the nuclear norm of a matrix $A \in \Re^{n \times J}$ defined by

$$\|A\|_* \triangleq \text{trace}\,(\sqrt{A^* A}) = \sum_{i=1}^{J} \sigma_i, \tag{4}$$

where σ_i are the singular values of matrix A. The nuclear norm of A is the ℓ_1-norm of its singular values. Denote by $\|A\|_1$ the ℓ_1-norm when A is considered to be a vector, i.e.,

$$\|A\|_1 \triangleq \sum_{i=1}^{n} \sum_{j=1}^{J} |a_{ij}|. \tag{5}$$

The parameters μ_1, μ_2 and μ_3 are some nonnegative constants, and W_i, $i = 1, 2$ are transforms that give sparse representations of underlying frames of video. The transform used is the wavelet frame transform constructed by [10] which will be described later.

In equation 3, X_1 and X_2 represent two different components of the reconstructed video volume. The low rank component X_1 is a relatively stationary component, which represents the background of the video. Matrix X_2 of equation 3 is the sparse component in some tight wavelet domain, which represents moving objects in the video volume. The terms with ℓ_1-norm in equation 2 are used to sparsify each frame in the video in the tight wavelet frame domain.

The minimization problem defined in equation 2 is a convex problem, so standard convex optimization algorithms such as the interior point methods can be used. However, these standard methods are computationally expensive. Instead, as shown in [3], the singular value threshold method is very efficient in low rank matrix completion and low rank matrix and sparse matrix decomposition. We leverage the idea of using a singular value threshold-based first order method using the augmented Lagrangian Alternating Direction Method (ADM) for solving this minimization problem. More details on the ADM algorithm applied to equation 2 can be found in [15], and an extension of it will be described in the next section.

Sparsifying Operators

A background image may be sparse in some transformed space, for example, in a wavelet transform space. Similarly, the moving objects represented by X_2 may have spatial correlations which can also be sparsified by a transform.

The operators W_i, $i = 1, 2$ in equation 2 are spatial sparsifying operators. Because W_i, $i = 1, 2$ have the same form, for simplicity, we use W to denote each of W_i, $i = 1, 2$ when there is no risk of confusion. For a given matrix $A \in \Re^{n \times J}$, the operator W work on columns of the matrix A. Specifically, let

$$A = [a_1, \ldots, a_J], \quad a_j \in \Re^n. \tag{6}$$

Then the spatial operator W is defined as

$$WA \triangleq [W^{(1)} a_1, \ldots, W^{(J)} a_J], \; W^{(j)} \in \Re^{n' \times n}, \; j = 1, \ldots, J. \tag{7}$$

In other words, the spatial operator W is defined by J linear operators which can be represented by matrices of dimension $n' \times n$, where often $n' \geq n$. The matrices $W^{(j)}$, $j = 1, \ldots, J$ may be different for each of W_i, $i = 1,2$, but they also may be the same. In this paper, all the matrices $W^{(j)}$ are identical, i.e., $W^{(j)} = W^{(0)}$, for all $j = 1, \ldots, J$. The matrices $W^{(j)}$ are a tight frame transform such as that described in [13] and [20]. When the tight frame transform is used, the operation $W^{(j)} a_j$ is carried out by regarding the column vector a_j as a two-dimensional (2D) image. When wavelet tight frames are used, the matrix multiplications are computed by a wavelet fast decomposition algorithm on each image frame as described in [9]. The wavelet frames are used in image restorations, since they give sparse approximations for many images. More details on applications of a wavelet frame for image restorations can be found in [10] and [21].

Low Latency Reconstruction

If the framework introduced in the section on Low Rank and Sparse Decomposition is applied

directly, such as in [15], a large number of video frames are needed for the reconstruction of the low rank matrix X_1, as given in equation 2, to properly segment the background from moving objects. Experiments show that the number of frames required is on the order of one hundred, i.e., $J \geq 100$. This represents a high latency in the reconstruction, which is unsuitable for real time processing. In this section, we introduce an augmented low rank matrix which allows the reconstruction to start with as few as one video frame worth of measurements.

Linearly Independent Background Frames

We assume that a set of background frames is available. As described previously, the background frames may be obtained via a training process. The background of a surveillance video may consist of more than one frame. For example, if the video is a surveillance of a room, the background may consist of a frame of the room with lights on and a frame of the room with lights off.

Let K be the number of linearly independent background frames, and b_k, $k = 1, ..., K$, be the vectors formed from these background frames. We define the matrix of linearly independent background frames as

$$X_b = [b_1, ..., b_K] \in \Re^{n \times K}. \tag{8}$$

For example, if the background is computed by a training process, then $K = rank\ (X_1)$, and X_b is a submatrix of X_1 with the maximum number of linearly independent columns. The columns b_k may be computed from X_1 but they may not necessarily be any columns of X_1.

Augmented Low Rank Matrix

When the matrix of the background X_b is known, a new video volume can be reconstructed by using an augmented matrix. Let $X = [x_1, ..., x_J] \in \Re^{n \times J}$ be a matrix, and $B \geq 0$ a real number, the augmented matrix of X is defined as

$$\hat{X} = \hat{X}(B) = [\sqrt{B}X_b, X] \in \Re^{n \times (J+K)}. \tag{9}$$

If B is a positive integer, the nonzero singular values of $\hat{X}(B)$ in equation 9 are the same as those of the matrix

$$[X_b, ..., X_b, X] \in \Re^{n \times (J+KB)}, \tag{10}$$

in which X_b is repeated B times. The purpose of the parameter B is to give a weight to the known background matrix X_b.

Once again, let y be the compressive measurements of a video volume X given by equation 1. Then low latency reconstruction can be carried out by the following minimization problem:

$$\min_{X_1, X_2} \mu_1 \|\hat{X}_1(B)\|_* + \mu_2 \|W_1 X_1\|_1 + \mu_3 \|W_2 X_2\|_1,$$
$$\text{such that} \quad y = \phi \circ (X_1 + X_2), \tag{11}$$

$$X = X_1 + X_2. \tag{12}$$

Note that in equation 11, the augmented matrix of X_1, $\hat{X}_1(B) = [\sqrt{B}X_b, X_1]$, as opposed to X_1 itself, is used in the nuclear norm. The use of the augmented matrix $\hat{X}_1(B)$ makes it possible to reconstruct X_1, X_2 accurately using only a small number of frames in the video volume. In fact, when the background X_b is known, and by properly choosing parameters B, μ_1, μ_2, μ_3, equation 11 may be used for accurate reconstruction for any values of J, which is the number of frames in the video volume, even for $J = 1$. Due to this property, the reconstruction method used in equation 11 and equation 12 is an LRSD with low latency, and it may be performed in real time as soon as measurements of a video frame become available.

Minimization Algorithm

We now extend the Alternating Direction Method (ADM) described in [15] to the low latency reconstruction model, equation 11 and equation 12. The main difficulty is that the nuclear norm term involves an augmented matrix having both known columns and unknown columns. However, this can be handled by replacing the augmented matrix with a new variable. In addition, we introduce some splitting variables to make the objective function separable. Specifically, we transform the original problem (equation 11) into the following one:

$$\min \mu_1 \|Z_1\|_* + \mu_2 \|Z_2\|_1 + \mu_3 \|Z_3\|_1,$$
$$\text{s.t. } [\sqrt{B}X_b, X_1] = Z_1,\ W_1 X_1 = Z_2,$$
$$W_2 X_2 = Z_3,\ \phi \circ (X_1 + X_2) = y. \tag{13}$$

The augmented Lagrangian function is given by

$$L_A(X, Z, \Lambda) = \mu_1 \|Z_1\|_* + \mu_2 \|Z_2\|_1 + \mu_3 \|Z_3\|_1 - < \Lambda_1,$$
$$[\sqrt{B}X_b, X_1] - Z_1 > + \frac{\beta_1}{2} \|[\sqrt{B}X_b, X_1] - Z_1\|_2^2$$

$$
\begin{aligned}
&- < \Lambda_2,\ W_1 X_1 - Z_2 > + \frac{\beta_2}{2} \|W_1 X_1 - Z_2\|_2^2 \\
&- < \Lambda_3,\ W_2 X_2 - Z_3 > + \frac{\beta_3}{2} \|W_2 X_2 - Z_3\|_2^2 \\
&- < \Lambda_4,\ \phi \circ (X_1 + X_2) - y \qquad (14) \\
&> + \frac{\beta_4}{2} \|\phi \circ (X_1 + X_2) - y\|_2^2.
\end{aligned}
$$

where Λ_i, $(i = 1, ..., 4)$ are Lagrangian multipliers, and $\beta_i > 0$ $(i = 1, ..., 4)$ are penalty parameters. All operations in equation 14, except the nuclear norm $\| \|_*$, are defined as vector operations, in which each variable is considered to be a column vector formed by concatenating its elements. The ADM algorithm can be similarly derived as in [15], and is similar to the split Bregman method used in image restorations [4, 14], which we briefly describe as follows.

Initialize: $Z_1^{(0)}$, $Z_2^{(0)}$, $Z_3^{(0)}$, $\Lambda_i^{(0)}$, β_i $(i = 1, ..., 4)$, $k = 0$.

While stopping criterion is not met, do

$$
\begin{aligned}
(X_1^{(k+1)}, X_2^{(k+1)}) &= \arg\min_X L_A\,(X, Z^{(k)}, \Lambda^{(k)}); \\
(Z_1^{(k+1)}, Z_2^{(k+1)}, Z_3^{(k+1)}) &= \arg\min_Z L_A\,(X^{(k)}, Z, \Lambda^{(k)});
\end{aligned}
$$

update $\Lambda_i^{(k+1)}$ $(i = 1, ..., 4)$ by

$$
\begin{aligned}
\Lambda_1^{(k+1)} &= \Lambda_1^{(k)} - \gamma\beta_1\,(X_1^{(k+1)} - Z_1^{(k+1)}), \\
\Lambda_2^{(k+1)} &= \Lambda_2^{(k)} - \gamma\beta_2\,(W_1 X_1^{(k+1)} - Z_2^{(k+1)}), \\
\Lambda_3^{(k+1)} &= \Lambda_3^{(k)} - \gamma\beta_3\,(W_2 X_2^{(k+1)} - Z_3^{(k+1)}), \\
\Lambda_4^{(k+1)} &= \Lambda_4^{(k)} - \gamma\beta_4\,[\phi \circ (X_1^{(k+1)} + X_2^{(k+1)}) - Z_1^{(k+1)}];
\end{aligned}
$$

$k \leftarrow k + 1$.

End

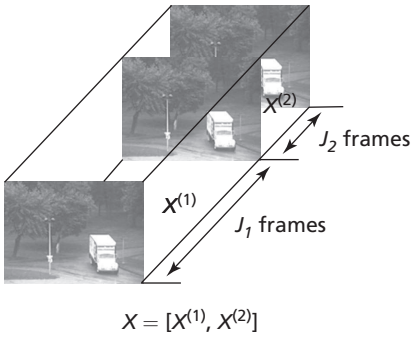
Theoretical Justification

In this section, we provide a theoretical justification for the low latency method that is presented in the section on Low Latency Reconstruction. We will show that if a large number of frames are used to obtain the background frames as in [15], then subsequent LRSD may be performed with any number of frames.

We start by making some definitions. Let J, J_1, $J_2 > 0$ be positive integers, such that

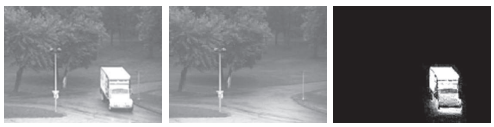
$$J = J_1 + J_2. \qquad (15)$$

Next, let $\bar{X} = \bar{X}_1 + \bar{X}_2 \in \Re^{n \times J}$ be a solution of the minimization problem in equation 2 and equation 3,



$X = [X^{(1)}, X^{(2)}]$

(a) The matrix X whose columns are the frames of a video sequence is divided into submatrix $X^{(1)}$ of J_1 frames and submatrix $X^{(2)}$ of J_2 frames.



$X = X_1 + X_2$

(b) Video frames X are segmented into background frames X_1 and foreground frames X_2.

Figure 2.
Illustration of notations.

and define $\bar{X}^{(i)}$, $\bar{X}_j^{(i)}$, $\phi^{(i)} \in \Re^{n \times J_i}$, $i, j = 1, 2$, and $y^{(2)} \in \Re^M$ such that

$$
\begin{aligned}
\bar{X} &= [\bar{X}^{(1)}, \bar{X}^{(2)}] = [\bar{X}_1^{(1)} + \bar{X}_2^{(1)}, \bar{X}_1^{(2)} + \bar{X}_2^{(2)}], \\
\phi &= [\phi^{(1)}, \phi^{(2)}], \qquad (16) \\
y^{(2)} &= \phi^{(2)} \circ \bar{X}^{(2)}.
\end{aligned}
$$

Note that in the definitions above, the superscripts 1 and 2 denote the first J_1 frames and the last J_2 frames of a matrix, respectively. The subscripts 1 and 2 denote the background and foreground, respectively. This is illustrated in **Figure 2.**

With these definitions, we have the following result.

Theorem 1:
Let $\bar{X}_1^{(1)}$ be defined in equation 16 from a solution, $\bar{X}$, to the minimization problem of equation 2 and equation 3, i.e., $\bar{X} = \bar{X}_1 + \bar{X}_2$, where $(\bar{X}_1, \bar{X}_2)$ is a solution to the following minimization problem:

$$
\begin{aligned}
&\min_{X_1, X_2} \mu_1 \|X_1\|_* + \mu_2 \|W_1 X_1\|_1 + \mu_3 \|W_2 X_2\|_1, \\
&\text{such that} \quad y = \phi \circ (X_1 + X_2). \qquad (17)
\end{aligned}
$$

If $X^* = X_1^* + X_2^*$ is a solution of the minimization problem

$$\min_{X_1, X_2} \mu_1 \|[\bar{X}_1^{(1)}, X_1]\|_* + \mu_2 \|W_1 X_1\|_1 + \mu_3 \|W_2 X_2\|_1,$$
$$\text{s.t.} \quad y^{(2)} = \phi^{(2)} \circ (X_1 + X_2), \tag{18}$$

$$X = X_1 + X_2 \in \Re^{n \times J_2}, \tag{19}$$

then the matrix

$$\hat{X} \triangleq [\bar{X}^{(1)}, X^*] \in \Re^{n \times J} \tag{20}$$

is a solution to the minimization problem of equation 2 and equation 3, i.e., $\hat{X} = \hat{X}_1, + \hat{X}_2$, where $(\hat{X}_1, \hat{X}_2)$ is a solution to the minimization problem defined in equation 17.

If we assume that there is a unique solution to LRSD minimization given in equation 2 and equation 3, which is justified by the work of [5] and [6] with the condition that ϕ and W_i, $i = 1,2$, are incoherent [5], then we have a stronger result as follows.

Corollary 1:
With notations of theorem 1, if the minimization problem in equation 2 and equation 3 has a unique solution, then the minimization problem in equation 18 and equation 19 also has a unique solution, and furthermore, the solution satisfies

$$X^* = \bar{X}^{(2)}, \text{ and } [\bar{X}^{(1)}, X^*] = \bar{X}. \tag{21}$$

Corollary 1 can be interpreted as follows. Let the background and foreground frames be properly segmented by using a large number of frames to compute the solution $\bar{X}$ of equation 2 and equation 3 with $J \gg 1$. The computed solution $\bar{X}$ has a large number of frames, from which we take the first J_1 background frames and use them as known background frames. We then form an augmented matrix $[\bar{X}_1^{(1)}, X]$ to process J_2 frames of video by solving equation 18 and equation 19, which is a problem for a smaller number of video frames. Corollary 1 shows that the J_2 video frames computed from equation 18 and equation 19 must be the same as the last J_2 frames of $\bar{X}$. Since J_2 can be any positive integer in theorem 1 and corollary 1, we can choose a small $J_2 \ll J$. Hence, the problem defined by equation 18 and equation 19 has low latency because J_2 can be small, as small as 1. Therefore, corollary 1 shows that if the background frames are known, the augmented matrix can be used to form a problem where the LRSD can be performed with low latency.

Proof of theorem 1:

We want to show that $\hat{X}$ of equation 20 is a solution to equation 2 and equation 3. Towards this purpose, we will first show that $\hat{X}$ satisfies the constraint of equation 2, and then we will show that $\hat{X}$ minimizes the cost function in equation 2.

First, from the definition of operator "$\circ$" given in equation 1, we have

$$\phi \circ [X^{(1)}, X^{(2)}] = \phi^{(1)} \circ X^{(1)} + \phi^{(2)} \circ X^{(2)},$$
$$\text{for any } X^{(1)} \in \Re^{n \times J_1}, X^{(2)} \in \Re^{n \times J_2}. \tag{22}$$

Then because X^* is the solution of the problem given by equation 18 and equation 19, we also have

$$\phi^{(2)} \circ X^* = y^{(2)} = \phi^{(2)} \circ \bar{X}^{(2)}. \tag{23}$$

In the equation above, the first equality is from the constraint in equation 18, and the second equality is from the definition of $y^{(2)}$ in the last equation of equation 16.

Now, because $\bar{X}$ is the solution to equation 2 and equation 3, by using equation 22 and equation 23, we can derive

$$\begin{aligned} y &= \phi \circ \bar{X} \\ &= \phi \circ [\bar{X}^{(1)}, \bar{X}^{(2)}] \\ &= \phi^{(1)} \circ \bar{X}^{(1)} + \phi^{(2)} \bar{X}^{(2)} \\ &= \phi^{(1)} \circ \bar{X}^{(1)} + \phi^{(2)} X^* \\ &= \phi \circ [\bar{X}^{(1)}, X^*] \\ &= \phi \circ \hat{X}, \end{aligned} \tag{24}$$

which shows that $\hat{X}$ satisfies the constraints of equation 2.

Next, we show that $\hat{X}$ has an expression in the form of equation 3 which minimizes the cost function in equation 2. Let

$$\hat{X}_1 = [\bar{X}_1^{(1)}, X_1^*], \hat{X}_2 = [\bar{X}_2^{(1)}, X_2^*]. \tag{25}$$

Then from the definition of $\hat{X}$, we have

$$\hat{X} = [\bar{X}^{(1)}, X^*] = [\bar{X}_1^{(1)} + \bar{X}_2^{(1)}, X_1^* + X_2^*] = \hat{X}_1 + \hat{X}_2, \tag{26}$$

which shows $\hat{X}$ can be expressed in the form of equation 3, but we still need to show that $\hat{X}_1$, $\hat{X}_2$ also

minimizes the cost function of equation 2. In order to do so, we need the following property which can be derived from the definition of $\| \ \|_1$ and W_i given in equation 5 and equation 7, respectively:

$$\|W_i\ [X^{(1)}, X^{(2)}]\|_1 = \|W_i\ X^{(1)}\|_1 + \|W_i\ X^{(2)}\|_1,\ i = 1,2, \quad \text{for any } X^{(1)} \in \Re^{n\times J}1,\ X^{(2)} \in \Re^{n\times J}2 \tag{27}$$

Now, since $X^* = X_1^* + X_2^*$ is a solution to equation 18, it minimizes the cost function of equation 18. Thus, the value of the cost function in equation 18, when evaluated at X^*, must be no greater than its value when evaluated at $\bar{X}^{(2)} = \bar{X}^{(2)}_1 + \bar{X}^{(2)}_2$, because $\bar{X}^{(2)}$ also meets the constraint of equation 18 by virtue of the last equation in equation 16. Therefore, we have

$$\begin{aligned}&\mu_1 \|[\bar{X}^{(1)}_1, X_1^*]\|_* + \mu_2 \|W_1\ X_1^*\|_1 + \mu_3 \|W_2\ X_2^*\|_1 \\ &\le \mu_1 \|[\bar{X}^{(1)}_1, \bar{X}^{(2)}_1]\|_* + \mu_2 \|W_1\ \bar{X}^{(2)}_1\|_1 + \mu_3 \|W_2\ \bar{X}^{(2)}_2\|_1.\end{aligned} \tag{28}$$

Adding $\mu_2 \|W_1\ \bar{X}^{(1)}_1\|_1 + \mu_3 \|W_2\ \bar{X}^{(1)}_2\|_1$ to both sides of the equation in equation 28, and making use of equation 27, we have

$$\begin{aligned}&\mu_1 \|[\bar{X}^{(1)}_1, X_1^*]\|_* + \mu_2 \|W_1\ [\bar{X}^{(1)}_1, X_1^*]\|_1 + \mu_3 \|W_2\ [\bar{X}^{(1)}_2, X_2^*]\|_1 \\ &\le \mu_1 \|[\bar{X}^{(1)}_1, \bar{X}^{(2)}_1]\|_* + \mu_2 \|W_1\ [\bar{X}^{(1)}_1, \bar{X}^{(2)}_1]\|_1 \\ &+ \mu_3 \|W_2\ [\bar{X}^{(1)}_2, \bar{X}^{(2)}_2]\|_1,\end{aligned} \tag{29}$$

which is equivalent to

$$\begin{aligned}&\mu_1 \|\hat{X}\|_* + \mu_2 \|W_1\ \hat{X}_1\|_1 + \mu_3 \|W_2\ \hat{X}_2\|_1 \\ &\le \mu_1 \|\bar{X}\|_* + \mu_2 \|W_1\ \bar{X}_1\|_1 + \mu_3 \|W_2\ \bar{X}_2\|_1.\end{aligned} \tag{30}$$

Since $\bar{X} = \bar{X}_1 + \bar{X}_2$ minimizes the cost function in equation 2, equation 30 implies that $\hat{X} = \hat{X}_1 + \hat{X}_2$ also minimizes it. This shows that $\hat{X} = \hat{X}_1 + \hat{X}_2$ is a solution to equation 2, and concludes the proof.

Proof of corollary 1:
From theorem 1, $\hat{X}$ is a solution to equation 2 and equation 3. Since equation 2 and equation 3 have a unique solution which is $\bar{X}$, we conclude $\hat{X} = \bar{X}$, which is equivalent to $[\bar{X}^{(1)}, X^*] = \bar{X}$, and in particular $X^* = \bar{X}^{(2)}$, and this hence proves equation 21. This holds true for any solution of equation 18 and equation 19, and therefore, there is a unique solution to equation 18 and equation 19, which concludes the proof.

Experimental Result

In this section, we present results from numerical experiments. The purpose of the experiments in this paper is to demonstrate the effectiveness of theorem 1 by numerical experiments. The work in [15] contains experiments for validating the LRSD itself.

Setup

The experiments are performed with three surveillance video sequences: Browse2, Shop, and Traffic, which were obtained from publicly available databases [12, 18]. For each video sequence, two tests are performed.

Test 1. In the first test, the total number of J frames are used in the minimization problem of equation 2 and equation 3 to compute the video frames $\bar{X}$ with the background $\bar{X}_1$ and foreground $\bar{X}_2$. This test has high latency because J is large.

Test 2. In the second test, we divide the video frames $\bar{X}$ computed from test 1 into two blocks: $\bar{X}^{(1)}$ of J_1 frames and $\bar{X}^{(2)}$ of J_2 frames. We use the J_1 background frames $\bar{X}^{(1)}_1$ that are computed from test 1 as the known background to form augmented matrix $[\bar{X}^{(1)}_1, X]$. Then we perform the minimization of equation 18 and equation 19 to compute J_2 frames of video X^* with background and foreground segmentation. For all video sequences in test 2, $J_2 = 10$. This represents low latency results.

We then compare the results computed in test 1 and test 2 for each video sequence. Corollary 1 stipulates that $X^* = \bar{X}^{(2)}$. That is, the background and foreground frames computed from test 2 should exactly match the non-used J_2 frames from test 1. In practice, however, there is some difference between test 1 and test 2 due to numerical errors from the minimization process, but the difference is small and negligible for the purpose of detecting moving objects.

To demonstrate the capability of detecting moving objects, we will show the images of the background X_1, and the silhouette of the moving objects obtained from the sparse component X_2.

The silhouette of frame n, S_n, is a binary image obtained from X_2 by the following equation.

$$S_n = T_\delta\,(Med(X_2\,(n))). \tag{31}$$

In equation 31, $X_2(n)$ is frame n (i.e., the n^{th} column) of the sparse component X_2. $Med(\cdot)$ is a median filter, and $T_\delta(\cdot)$ is a threshold operator defined as

$$T_\delta\,(X)(i,j) = \begin{cases} 1, \text{ if } |X(i,j)| \geq \delta \\ 0, \text{ if } |X(i,j)| < \delta \end{cases}. \tag{32}$$

For all experiments, a permutated Walsh-Hadamard matrix [16, 17] is used to take measurements of the video volume. The number of measurements used in the reconstruction of LRSD is expressed as a percentage of the total number of pixels in the video volume. For example, 100 percent means the number of measurements is equal to the total number of pixels in the video volume.

Table I. Configuration of experiments.

Name	Browse	Shop	Traffic
Resolution	384×288	384×258	378×282
Total frames (J)	100	120	190
J_1	90	110	180
J_2	10	10	10
Measurements (%)	4	4	4

In the reconstruction, the parameters μ_1, μ_2, μ_3 are fixed for all experiments, and they are given by

$$\mu_1 = 1,\ \mu_2 = 0,\ \mu_3 = 1e-3. \tag{33}$$

Table I shows the configuration used in the experiments.

Browse2

Browse2 [12] is a color sequence from a camera monitoring a building lobby. The original is a Moving Picture Experts Group (MPEG) file with a resolution of 384 × 288 and it is more than six minutes in length. We take $J = 100$ frames from the sequence, and process only the luminance component. The total number of pixels is $N = 384 \times 288 \times 100 = 11059200$. The total number of measurements used in the reconstruction is 1/25 (four percent) of the total number of pixels, i.e., the total number of measurements is $M = 442368$.

The results from test 1 and test 2 are shown in **Figure 3**. The results demonstrate that the low latency results (from test 2) are nearly identical to those from test 1 with long latency.

Shop

Shop [12] is a color sequence from a camera in a shopping mall. The original is an MPEG file with a resolution of 384 × 258 and it is about one minute in length. We take $J = 120$ frames from the sequence, and process only the luminance component.

Figure 3.
Result for Browse2 sequence. Top: frame 91 in test 1, which is the same as frame 1 in test 2. Bottom: frame 100 in test 1, which is the same as the frame 10 in test 2. (a) and (b): background and foreground frames from test 1. (c): the original video frames. (d) and (e): foreground and background frames from test 2.

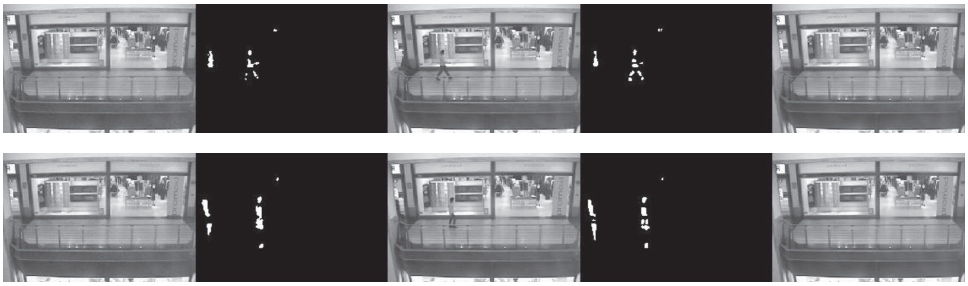


Figure 4.
Result for Shop sequence. Top: frame 111 in test 1, which is the same as frame 1 in test 2. Bottom: frame 120 in test 1, which is the same as frame 10 in test 2. (a) and (b): background and foreground frames from test 1. (c): the original video frames. (d) and (e): foreground and background frames from test 2.

Figure 5.
Results from Traffic sequence. Top: frame 181 in test 1, which is the same as frame 1 in test 2. Bottom: frame 190 in test 1, which is the same as frame 10 in test 2. (a) and (b): background and foreground frames from test 1. (c): the original video frames. (d) and (e): foreground and background frames from test 2.

The total number of measurements used in the reconstruction is four percent of the total number of pixels. The results from test 1 and test 2 are shown in **Figure 4**. The results demonstrate that the low latency results (from test 2) are nearly identical to those from test 1 with long latency.

Traffic

Traffic [18] is a black and while sequence from a traffic camera in a highway intersection. The original is a sequence of $J = 190$ Joint Photographic Experts Group (JPEG) frames with a resolution of 378×282. The total number of measurements used in the reconstruction is four percent (1/25) of the total number of pixels. The results from test 1 and test 2 are shown in **Figure 5**. The results demonstrate that the low latency results (from test 2) are nearly identical to those from test 1 with long latency.

Conclusions

We presented a low latency method for analyzing surveillance video by using compressive sensing in which background and foreground is segmented by LRSD. Once the background is available, it can be

used to form an augmented matrix in a minimization problem to compute a small number of video frames, and therefore result in a low latency method. Low latency makes it possible to analyze video in real time. Both theoretical justification and experimental validation of the low latency method are provided.

References

[1] Barnich and M. Van Droogenbroeck, "ViBe: A Universal Background Subtraction Algorithm for Video Sequences," IEEE Trans. Image Process., 20:6 (2011), 1709–1724.

[2] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Comparative Study of Background Subtraction Algorithms," J. Electron. Imaging, 19:3 (2010).

[3] J.-F. Cai, E. J. Candès, and Z. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," SIAM J. Optim., 20:4 (2010), 1956–1982.

[4] J.-F. Cai, S. Osher, and Z. Shen, "Split Bregman Methods and Frame Based Image Restoration," Multiscale Model. Simul., 8:2 (2009), 337–369.

[5] E. J. Candès, "Compressive Sampling," Proc. Internat. Congress of Mathematicians (ICM '06) (Madrid, Spn., 2006), vol. 3, pp. 1433–1452.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?," J. ACM, 58:3 (2011), article 11.

[7] E. Candès, J. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," Commun. Pure Appl. Math., 59:8 (2005), 1207–1223.

[8] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa, "Compressive Sensing for Background Subtraction," Proc. 10th Eur. Conf. on Comput. Vision (ECCV '08) (Marseille, Fra., 2008), pp. 155–168.

[9] I. Daubechies, B. Han, A. Ron, and Z. Shen, "Framelets: MRA-Based Constructions of Wavelet Frames," Appl. Comput. Harmon. Anal., 14:1 (2003), 1–46.

[10] B. Dong and Z. Shen, "MRA-Based Wavelet Frames and Applications," Summer Program—The Mathematics of Image Processing, Institute for Advanced Study (IAS)/Park City Mathematics Series vol. 19, 2010, <http://math.arizona.edu/∼dongbin/Publications/IASLectureNotes.pdf>.

[11] Y. Dong, T. X. Han, and G. N. DeSouza, "Illumination Invariant Foreground Detection Using Multi-Subspace Learning," Internat. J. Knowledge-Based and Intell. Eng. Syst., 14:1 (2010), 31–41.

[12] European Commission, Sixth Framework Programme, "CAVIAR: Context Aware Vision Using Image-Based Active Recognition," Information Society Technology Programme Project FP6-IST- 2001-37540, Oct. 2002–Sept. 2005, <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>.

[13] H. Gao, J.-F. Cai, Z. Shen, and H. Zhao, "Robust Principal Component Analysis-Based Four-Dimensional Computed Tomography," Phys. Med. Biol., 56:11 (2011), 3181–3198.

[14] T. Goldstein and S. Osher, "The Split Bregman Algorithm for L1-Regularized Problems," SIAM J. Imaging Sci., 2:2 (2009), 323–343.

[15] H. Jiang, W. Deng, and Z. Shen, "Surveillance Video Processing Using Compressive Sensing," Inverse Probl. Imaging, 6:2 (2012), 201–214.

[16] H. Jiang, C. Li, R. Haimi-Cohen, P. A. Wilford, and Y. Zhang, "Scalable Video Coding Using Compressive Sensing," Bell Labs Tech. J., 16:4 (2012), 149–169.

[17] C. Li, H. Jiang, P. Wilford, Y. Zhang, and M. Scheutzow, "A New Compressive Video Sensing Framework for Mobile Broadcast," IEEE Trans. Broadcasting, 59:1 (2013), 197–205.

[18] V. Mahadevan and N. Vasconcelos, "Spatiotemporal Saliency in Dynamic Scenes," IEEE Trans. Pattern Anal. Mach. Intell., 32:1 (2010), 171–177.

[19] S. Pudlewski, T. Melodia, and A. Prasanna, "Compressed-Sensing-Enabled Video Streaming for Wireless Multimedia Sensor Networks," Wireless Networks and Embedded Systems Laboratory, Department of Electrical Engineering, State University of New York (SUNY) Buffalo, 2011, <http://www.eng.buffalo.edu/wnesl/papers/TMC-2011.pdf>.

[20] A. Ron and Z. Shen, "Affine Systems in $L_2(\mathrm{R}^d)$: The Analysis of the Analysis Operator," J. Funct. Anal., 148:2 (1997), 408–447.

[21] Z. Shen, "Wavelet Frames and Image Restorations," Proc. Internat. Congress of Mathematicians (ICM '10) (Hyderabad, Ind., 2010), vol. 4, pp. 2834–2863.

[22] C. Stauffer and W. E. L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," Proc. IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recognition (CVPR '99) (Fort Collins, CO, 1999), vol. 2, pp. 246–252.
[23] F. Yang, H. Jiang, Z. Shen, W. Deng, and D. Metaxas, "Adaptive Low Rank and Sparse Decomposition of Video Using Compressive Sensing," Proc. IEEE Internat. Conf. on Image Process. (ICIP '13) (Melbourne, Aus., 2013), paper TA.PC.4.

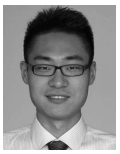
(Manuscript approved September 2013)

HONG JIANG is a researcher with Alcatel-Lucent Bell Labs in Murray Hill, New Jersey. He received his B.S. from Southwestern Jiaotong University, Chengdu, China, M. Math from the University of Waterloo, and Ph.D. from the University of Alberta in Canada. Dr. Jiang is currently conducting research on digital communications, image and video processing. He has authored more than 50 technical papers in scientific and engineering journals, and has more than 40 U.S. patents in digital communications.

SONGQING ZHAO received B.E. and M.E. degrees in control science and engineering from Huazhong University of Science and Technology, Wuhan, China, and received his Ph.D. degree in electrical and computer engineering from University of Illinois at Chicago. He was formerly a member of technical staff at Alcatel-Lucent in Murray Hill, New Jersey where his research focus was on fourth generation (4G) Long Term Evolution (LTE) and video quality. Prior to joining Bell Labs, he worked as a research intern at Mitsubishi Electric Research Labs in Boston, Massachusetts, and for the Institute for Telecommunications Research, Adelaide, Australia. His research interests include wireless multimedia communications, information theory, video quality, quality of experience, video transmission, video processing, and pattern recognition.

ZUOWEI SHEN is the Tan Chin Tuan Centennial Professor at the National University of Singapore where he has been on the faculty at the Department of Mathematics since 1993. His primary research interests include wavelet frames, Gabor frames, and applications. More recently his research has focused on imaging science using wavelet and Gabor frames.

WEI DENG is a Ph.D. student in the Department of Computational and Applied Mathematics at Rice University, Houston, Texas. He received a B.S. degree in mathematics from Nanjing University, Nanjing, China, and an M.A. degree in computational and applied mathematics from Rice University, Houston, Texas. He is currently conducting research on developing and analyzing numerical optimization algorithms for various applications including compressive sensing, image and video processing, and machine learning.



PAUL A. WILFORD is the director of Multimedia Research at Alcatel-Lucent Bell Labs in Murray Hill, New Jersey. He received his B.S. and M.S. in electrical engineering from Cornell University, Ithaca, New York. His research focus was communication theory and predictive coding. Mr. Wilford is a Bell Labs fellow. He has made extensive contributions in the development of digital video processing and multimedia transport technology. He was a key leader in the development of Lucent Technologies' first high-definition television (HDTV) broadcast encoder and decoder. Under his leadership, Bell Labs then developed the world's first Moving Picture Experts Group 2 (MPEG2) encoder. He has made fundamental contributions in the high speed optical transmission area. Currently he is leading a department working on next-generation video transport systems, hybrid satellite-terrestrial networks, and high-speed mobility networks.

RAZIEL HAIMI-COHEN is a researcher with Alcatel-Lucent Bell Labs in Murray Hill, New Jersey and a member of the Alcatel-Lucent Technical Academy. His current research is in compressive sensing of video. Previously, he worked in the areas of video delivery and processing, audio and speech compression, speech recognition, signal processing, and cellular communication. Dr. Haimi-Cohen holds a B.Sc. in mathematics from Tel-Aviv University in Israel, an M.Sc. in applied mathematics from Cornell University, Ithaca, New York, and a Ph.D. in electrical engineering from Ben-Gurion University in Israel. ◆