ADAPTIVE LOW RANK AND SPARSE DECOMPOSITION OF VIDEO USING COMPRESSIVE SENSING

Fei Yang[1] *Hong Jiang*[2] *Zuowei Shen*[3] *Wei Deng*[4] *Dimitris Metaxas*[1]

[1]Rutgers University [2]Bell Labs [3]National University of Singapore [4]Rice University

ABSTRACT

We address the problem of reconstructing and analyzing surveillance videos using compressive sensing. We develop a new method that performs video reconstruction by low rank and sparse decomposition adaptively. Background subtraction becomes part of the reconstruction. In our method, a background model is used in which the background is learned adaptively as the compressive measurements are processed. The adaptive method has low latency, and is more robust than previous methods. We will present experimental results to demonstrate the advantages of the proposed method.

Index Terms— Compressive sensing, low rank and sparse decomposition, background subtraction

1. INTRODUCTION

In video surveillance, video signals are captured by cameras and transmitted to a processing center, where the videos are monitored and analyzed. Given a large number of cameras installed in public places, an enormous amount of data are generated and need to be transmitted in the network, raising a high risk of network congestion. Therefore, it is highly desirable to compress the video signals transmitted in the network.

The recently introduced compressive sensing theory proves that if a signal has a sparse representation in some basis, then it can be reconstructed from a small set of linear measurements [1][2]. The number of measurements can be much smaller than that required by Nyquist sampling rate. Since videos are known to have a sparse representation in some transform basis (e.g. total variation, wavelet or framelet, etc.), the compressive sensing theory can be applied to compress video at the cameras, for example to acquire video by compressive measurements which can then be used to reconstruct the video [3][4].

In this paper, we developed a framework for processing surveillance video using compressive measurements. Our system is shown in Fig. 1. At the camera, the video captured by a surveillance camera is either acquired as, or transformed to, the low dimensional measurements by using random projections. At the processing center, the frames of the video are reconstructed, and the moving objects are detected at the same time.

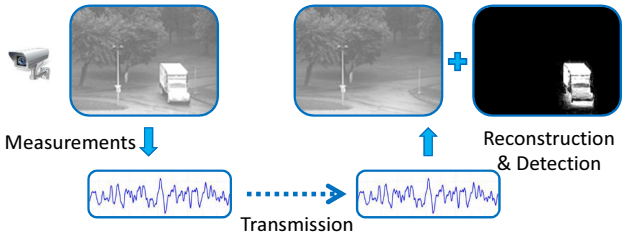

Fig. 1. The framework of the compressive sensing surveillance system. The video is compressed by using random projections, and then transmitted to the processing center. The frames are reconstructed and the moving objects are detected simultaneously.

Our method is based on three observations: 1). The background is nearly static over a short period. Thus the background images lie in a low dimensional subspace. 2). Natural images are sparse in a transform, such as tight wavelet frame, domain. 3). Generally the moving objects only occupies a small portion of the field of view of a surveillance camera. Based on these observations, we use a low rank model for background and a sparse model for moving objects. The reconstruction of background and moving objects is performed by a low rank and sparse decomposition similar to [**?**][5].

In the low rank model of [5], a large number of frames of video must be used in order to properly reconstruct the background because the low rank and sparse decomposition computes background frames as a low rank basis of the space spanned by the incoming video frames. This results in a long latency in the reconstruction.

In this paper, we introduce an adaptive background model in which the low rank and sparse decomposition is performed with a small number of video frames. This significantly reduces latency. In this adaptive method, the video frames are reconstructed by a few frames at a time. In each reconstruction, the compressive measurements from a small number of video frames are used to perform the low rank and sparse decomposition which produces a set of background frames. The background frames are further processed and the results are used in the low rank and sparse decomposition for the next set of frames. Therefore, effectively, a large number of background frames are participated (although not explicitly used)

in the computation of the low rank and sparse decomposition at each reconstruction, since the background frames from previous reconstructions are used. This makes it possible to accurately reconstruct background frames even with a small number of frames processed each time. The proposed method handles background changes very well because it is adaptive. Furthermore, the method reduces latency and computational complexity significantly.

In the remaining parts of the paper, we first introduce previous work related to our study. Then we introduce the framework of our video reconstruction method, followed by the background model and its adaption algorithm. The experimental results are given at the end.

2. RELATED WORK

Background subtraction. There has been extensive study on background subtraction from original videos [6]. The earliest background subtraction methods use frame difference to detect foreground [7]. Subsequent approaches aimed to model the variations and uncertainty in background appearance, such as mixture of Gaussian [8] and non-parametric kernel density estimation [9]. Currently state-of-art background subtraction methods are able to get satisfactory results for stationary cameras. However, these methods cannot be applied to compressive measurements.

Sparse reconstruction. Cevher et al. [10] casted the background subtraction as a sparse approximation problem and solved it based on convex optimization. Their method relies on a background model trained from pure background frames, which requires the prior knowledge of the background. Jiang et al. [5] developed a low rank and sparse decomposition based approach to detect moving objects from a video. Their method solves all the frames at the same time, which results in a long latency and expensive computational cost. In contrast, the approach in this paper does not require a clean background for training, and it reconstructs background adaptively, with a small number of frames of video processed at a time. This reduces latency and complexity.

3. LOW RANK AND SPARSE DECOMPOSITION

3.1. Compressive measurements

We consider a video consisting of m frames. Each frame has a total of n pixels. Let $x_j \in \Re^n$ be a vector formed by concatenating all pixels in frame j. Let $X = [x_1, ..., x_m] \in \Re^{n \times m}$ be a matrix containing m columns representing the m frames of the video. Let $\Phi \in \Re^{r \times n}$ be a sensing matrix. The compressive measurements of X are defined as

$$y = \Phi \circ X \triangleq [\Phi x_1, ..., \Phi x_m], \tag{1}$$

where $y \in \Re^{r \times n}$ is a matrix of measurements, with a much smaller row dimension than X, i.e., $r \ll n$. Each column of y contains r measurements of a frame of video. In our work, Φ is composed of a set of r randomly permutated rows of Walsh-Hadamard matrix.

3.2. Reconstruction

Given the measurements y , we want to reconstruct the original video X. X can be decomposed into background matrix X_1 and foreground matrix X_2:

$$X = X_1 + X_2. \tag{2}$$

In above, X_1 is a matrix each column of which is formed from the pixels of a background frame of the video. Similarly, X_2 is a matrix each column of which is formed from the pixels of a foreground frame of the video. Thus the objective is to solve X_1 and X_2, satisfying Eqs. (1) and (2). Apparently, this is an ill-posed problem which has infinite number of solutions. Therefore, we need some prior knowledge to find a proper solution.

Low rank background. We assume the background images have relative small changes over a short period, then the background matrix X_1 should have a low rank [**?**]. We use the nuclear norm to measure the rank of this matrix, which is defined as the sum of single values σ_i:

$$||X_1||_* = trace(\sqrt{X_1 X_1^T}) = \sum_i \sigma_i. \tag{3}$$

Sparsity in transformed domain. Previous work shows that natural images can be sparsely represented in a transformed space. We assume each background frame is sparse under transform W_1, and each foreground frame is sparse under a transform W_2 [5]. We use the the l_1-norm to measure the sparsity of the transformed background and foreground: $||W_1 \circ X_1||_1$, $||W_2 \circ X_2||_1$, where the l_1-norm is defined as

$$||Z||_1 \triangleq \sum_i \sum_j |z_{ij}|, \quad Z = [z_{ij}]. \tag{4}$$

Sparse foreground. We also assume the foreground only occupies a small portion of a frame, and therefore, we can also use l_1-norm as defined in Eq. (4) to measure the its sparsity: $||X_2||_1$.

Given these prior assumptions, X_1 and X_2 can be reconstructed by solving the following optimization problem:

$$\begin{aligned}(X_1, X_2) = \arg\min_{X_1, X_2} \ & \mu_1 ||X_1||_* + \mu_2 ||W_1 \circ X_1||_1 \\ & + \mu_3 ||W_2 \circ X_2||_1 + \mu_4 ||X_2||_1 \\ \text{such that} \quad & y = \Phi \circ (X_1 + X_2).\end{aligned} \tag{5}$$

In above, μ_1, μ_2, μ_3 and μ_4 are nonnegative weights. W_1 and W_2 are sparsifying operators. In our system, we set $W_1 = W_2 = W$ as the framelet transform [11][12][5].

Eq. (5) is a convex problem, so standard convex optimization algorithms such as the interior point method [13] can be applied to find a solution. However, these standard methods are computationally expensive. Instead, as shown in [14], singular value thresholding is more efficient for low rank decomposition. We apply the Augmented Lagrangian Alternating Direction (ALAD) algorithm introduced in Jiang et al. [5].

4. ADAPTIVE RECONSTRUCTION

For the optimization problem described in [5], a large number of frames (i.e., $m > 100$) are needed to find a proper solution, which leads to a high latency in the reconstruction. In addition, the computational complexity of singular value thresholding is $O(m^3)$, which makes the algorithm highly computationally expensive as the m becomes large.

To reconstruct the background and foreground by solving Eq. (5), a large number of frames (i.e., the number of columns of X_1) are required. This is because the solution to Eq. (5) captures the low rank basis in the space spanned by X_1. If the number of frames is small, a moving object may not change significantly, thus would be captured as part of background. Only when a large number of frames, the solution to (5) would reconstruct a background as expected. This is the reason that a large number of frames must be used in [5].

In this section, we introduce an adaptive method to reduce both latency and complexity. In order to reduce latency, we want to process a small number of frames each time. However, in order to improve accuracy of reconstructed background, we still need a large number of columns to be present in the calculation of the nuclear norm $||\cdot||_*$. For this purpose, we augment X_1 by the previously calculated background frames. In other words, we replace $||X_1||_*$ in Eq. (5) by $|||[M_b, X_1]|||_*$ where M_b is a matrix which is a model of previously calculated background frames.

The key idea of the paper is that M_b, a representation of previously calculated background frames, is low dimensional and is computed adaptively as more frames are processed. M_b may initially be an inaccurate approximation of the background frames, but as the adaptation proceeds, M_b becomes progressively better representation of background frames. Furthermore, as background changes, M_b changes accordingly with the background. Therefore, this method not only reduces latency and complexity, but also allows the reconstructed background frames to adapt quickly to the changes in the background of the video.

4.1. Augmented low rank decomposition

We assume that a set of k background frames, b_j, are already computed in processing the previous frames. We put them in a background matrix defined as:

$$X_b = [b_1, ..., b_k] \in \Re^{n \times k}.$$

The augmented background matrix $\hat{X_1}$ is formed by combining the previously computed background matrix X_b with the to-be-computed background X_1 of m new frames:

$$\hat{X_1} = [X_b, X_1] \in \Re^{n \times (k+m)}.$$

The use of the augmented matrix makes it possible to reconstruct X_1, X_2 even if X_1 has a very small number of columns. We now require $\hat{X_1}$, instead of X_1, to have a small rank. Therefore, the problem to solve is same as Eq. (5) but with $||X_1||_*$ replaced by $||\hat{X_1}||_*$. By using $\hat{X_1}$, there is no need for X_1 to have a large number of columns.

4.2. Low dimensional background model

The computational complexity to optimize the low rank of $\hat{X_1}$ is $O(k+m)^3$, which grows quickly as frames are continuously being processed. Therefore, we need to find a lower dimensional background model $M_b \in \Re^{n \times p}$ from the computed background frames X_b, for a new augmented matrix: $[M_b, X_1] \in \Re^{n \times (p+m)}$, where $p \ll k$. We need to find M_b such that the nuclear norm of $[M_b, X_1]$ could approximate the nuclear norm of $\hat{X_1}$, which leads to the following optimization problem:

$$M_b = \arg\min_{M_b} \left| ||\hat{X_1}||_* - |||[M_b, X_1]|||_* \right|. \quad (6)$$

We perform SVD decomposition of the background matrix X_b, and form M_b as

$$X_b = UDV^T, \quad (7)$$

$$M_b = U_p D_p. \quad (8)$$

In Eqs. (7) and (8), D is a diagonal matrix containing singular values of X_b, and U, V are orthogonal matrices. D_p is a diagonal matrix formed by the p largest single values, and U_p is consist of the first p columns of U.

Now, replacing $||X_1||_*$ by $|||[M_b, X_1]|||_*$ in Eq. (5), we have the low latency reconstruction given as:

$$\begin{aligned}(X_1, X_2) = \arg\min_{X_1, X_2} \ & \mu_1 |||[M_b, X_1]|||_* + \mu_2 ||W_1 \circ X_1||_1 \quad (9)\\ & + \mu_3 ||W_2 \circ X_2||_1 + \mu_4 ||X_2||_1,\\ \text{such that} \quad & y = \Phi \circ (X_1 + X_2).\end{aligned}$$

4.3. Optimization

We now use the Augmented Lagrangian Alternating Direction (ALAD) algorithm to solve the problem in Eq. (9). 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 splitting variables to make the objective function separable. We perform variable substitution as below:

$$Z_1 = [M_b, X_1],\ Z_2 = W_1 \circ X_1,\ Z_3 = W_2 \circ X_2. \quad (10)$$

The ALAD optimization is shown in Alg. 1. More details about the optimization framework can be found in [5].

4.4. Updating the background model

With the previously computed M_b, Eq. (9) can be used to compute current background frames X_1 by Alg. 1. Then the question is, how do we update M_b with current X_1 to obtain a new background model $M_b^{(new)}$ in order for us to solve Eq. (9) to reconstruct the next set of frames? We use an approach

Fig. 2. Results of video reconstruction and background subtraction. **Left**: original frames; **Middle**: background and foreground reconstructed using the method of this paper; **Right**: Foreground masks generated from original video with GMM.

Algorithm 1 Reconstructing X_1 and X_2 using ALAD.

Initialize $Z_i^{(0)}, \Lambda_i^{(0)}$,
repeat
  Update X_1, X_2, while fixing Z_i and Λ_i,
  Update Z_i, while fixing X_1, X_2 and Λ_i,
  Update Λ_i, while fixing X_1, X_2 and Z_i,
until converge

to update M_b similar to the incremental SVD [15]. Given the SVD decomposition $X_b \approx U_p D_p V_p^T$, the decomposition of the augmented matrix with current background frames X_1 can be used to update M_b as follows:

$$\begin{aligned} \begin{bmatrix} U^{(new)} & D^{(new)} \end{bmatrix} &= svd(\begin{bmatrix} w_b X_b & w_a X_1 \end{bmatrix}), \\ &\approx svd(\begin{bmatrix} w_b U_p D_p V_p^T & w_a X_1 \end{bmatrix}), \\ &= svd(\begin{bmatrix} w_b U_p D_p & w_a X_1 \end{bmatrix}), \\ &= svd(\begin{bmatrix} w_b M_b & w_a X_1 \end{bmatrix}). \\ M_b^{(new)} &= U_p^{(new)} D_p^{(new)}. \end{aligned} \quad (11)$$

In (11), $D_p^{(new)}$ is a diagonal matrix formed by the p largest single values, and $U_p^{(new)}$ is consist of the first p columns of $U^{(new)}$, similar to those in (8). w_a and w_b are weights controlling the updating rate.

It is important to point out that in the update (11), the large matrix V in SVD will never need to be computed, representing a significant reduction in complexity.

5. EXPERIMENTS

We perform experiments on three video clips from PETS2001 database. The results are shown in Fig. 2. The first column shows the original frames. The second and third columns show backgrounds and foregrounds reconstructed by the method of this paper. We use 5% measurements for the first two examples, and 10% measurements in the last example. Median filters are used to post-process the results of our method to reduce the noises. The last column shows the foregrounds generated by applying Gaussian Mixture model (GMM). [8].

Fig. 2 demonstrates that the results of our method are comparable to GMM. But our method are performed by only using 5%-10% of the original data, while GMM uses 100%.

6. CONCLUSION

In this paper, we address the problem of reconstructing and analyzing surveillance videos from compressive measurements. We propose a method that simultaneously performs reconstruction and background subtraction with low latency. Our method is built on a background model, which is continuously updated as new frames are reconstructed. The experiments have proved the effectiveness and efficiency of the proposed method.

7. REFERENCES

[1] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from

highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[2] David L. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] H. Jiang, C. Li, R. Haimi-Cohen, P. Wilford, and Y. Zhang, "Scalable video coding using compressive sensing," *Bell Labs Technical Journal*, vol. 16, no. 4, pp. 149–169, 2012.

[4] C. Li, H. Jiang, P. Wilford, Y. Zhang, and M. Scheutzow, "A new compressive video sensing framework for mobile broadcast," *IEEE Transactions on Broadcasting*, to appear 2013.

[5] H. Jiang, W. Deng, and Z. Shen, "Surveillance video processing using compressive sensing," *Inverse Problems and Imaging*, vol. 6, no. 2, pp. 201–214, 2012.

[6] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. CVPR*, 2011.

[7] R. Jain and H.H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 206–214, 1979.

[8] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[9] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of IEEE*, vol. 90, no. 7, pp. 1151–1163, 2002.

[10] V. Cevher, A. Sankaranarayanan, M. Duarte, D. Reddy, R. Baraniuk, and R. Chellappa, "Compressive sensing for background subtraction," in *Proc. ECCV*, 2008.

[11] A. Ron and Z. Shen, "Affine systems in $l_2(r^d)$: the analysis of the analysis operator," *Journal of Functional Analysis*, vol. 148, pp. 408–447, 1997.

[12] I. Daubechies, B. Han, A. Ron, and Z. Shen, "Framelets: Mra-based constructions of wavelet frames," *Applied and Computational Harmonic Analysis*, , no. 14, pp. 1–46, 2003.

[13] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C.A. Sagastizábal, *Numerical optimization: theoretical and practical aspects*, Springer, 2006.

[14] J.F. Cai, E.J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[15] M. Brand, "Fast low-rank modifications of the thin singular value decomposition," *Linear algebra and its applications*, vol. 415, no. 1, pp. 20–30, 2006.