ON THE DETECTION OF ABANDONED OBJECTS WITH A MOVING CAMERA USING ROBUST SUBSPACE RECOVERY AND SPARSE REPRESENTATION

Eric Jardim[1], *Xiao Bian*[2], *Eduardo A. B. da Silva*[1], *Sergio L. Netto*[1], *and Hamid Krim*[2]

[1]Federal University of Rio de Janeiro
{eric.jardim, eduardo, sergioln}@smt.ufrj.br

[2]North Carolina State University
{xbian, ahk}@ncsu.edu

ABSTRACT

We consider the application of sparse-representation and robust-subspace-recovery techniques to detect abandoned objects in a target video acquired with a moving camera. In the proposed framework, the target video is compared to a previously acquired reference video, which is assumed to have no abandoned objects. The detection method explores the low-rank similarities among the reference and target videos, as well as the sparsity of the differences between the two video sequences caused by the unexpected object in the target video. A three-step procedure is then presented adapting a previous low-rank and sparse image representation to the problem at hand. Performance of the proposed technique is verified using a large video database for abandoned-object detection in a cluttered environment. Results demonstrate the technique effectiveness even in the presence of some significant camera shake along its trajectory.

1. INTRODUCTION

Fixed surveillance cameras for monitoring purposes are ubiquitous in today's world. Inspection of the acquired content, however, is highly inefficient when it depends on direct human supervision, particularly when a large number of cameras must be continuously monitored. This is so because, among other things, such an activity can be tedious and error-prone. To alleviate this problem, automatic anomaly-detection techniques have been proposed to substitute or complement the human-based detection task [1, 2].

Visual-surveillance systems based on computer-vision techniques constitute a trendy research topic due to is inherent difficulty, as it depends on both the searched anomaly and the environment complexity. In addition, some monitoring activities may require several viewpoints for an efficient inspection, particularly within a cluttered environment where there may be uncontrolled occlusions. In these situations, pose-changing or wide-angle cameras may not provide an acceptable solution. An interesting alternative to multiple cameras is the use of a single camera mounted on a moving (robotic) platform.

Sparse representations have been successfully used in problems which involve high-dimensional signals with low-dimensional representation models, such as face recognition [3, 4, 5] and background subtraction when static surveillance cameras are used [6]. In this paper we investigate how these techniques can be adapted to detect abandoned objects within a video sequence acquired by a moving surveillance camera. In the proposed framework, an initial representation is determined for the reference video, that we consider to have no abandoned objects. On a second step, this representation is employed to model the target video. Assuming pure translational movement, every object present in both videos is properly modeled by the initial representation, irrespective of any temporal misalignment between the two sequences. On the other hand, the abandoned objects cannot be modeled properly, since they are only found in the target video, appearing as a residue in this representation process.

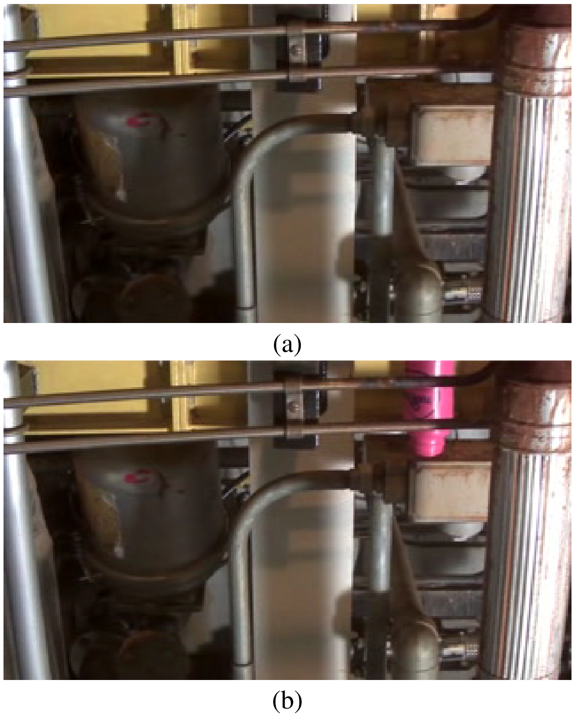
In order to introduce the proposed abandoned object detection scheme, this paper is organized as follows: Section 2 describes the video database employed to assess the performance of the proposed abandoned-object detection system. Section 3 revisits the so-called robust subspace recovery (RoSuRe) algorithm [7], whereas Section 4 considers its proposed adaptation to the object-detection problem of interest. Finally, Section 5 summarizes the results achieved with the proposed formulation, and Section 6 concludes the paper emphasizing its main contributions.

2. ABANDONED OBJECT SCENARIO

In this work, the abandoned-object detection problem is assessed using an existing video database with several recordings of a complex industrial-like environment [8, 9]. The database was acquired with a rigid camera mounted on a robotic iRobot Roomba platform with a back-and-forth linear movement on a 6m-long hanging rail. Two different IP cameras were employed, having the same 1280×720-pixel resolution and a frame rate of 24 frames per second. An industrial environment was considered, comprised of several pipes and valves. Twenty four distinct abandoned objects were employed in the recordings, which total approximately 8.2 hours of video. This database, along with the annotations of the abandoned objects, can be downloaded from [9].

In this database, the recordings were divided into two groups - reference and target sequences, as exemplified in Fig. 1. The reference sequences have no abandoned objects, as validated by human supervision, while the target sequences contain one or more objects to be detected automatically by the proposed algorithm. Due to track imperfections and mechanical friction with the robot wheels, the captured sequences present considerable camera shake, which poses an additional challenge to the detection scheme.

978-1-4673-6997-8/15/$31.00 ©2015 IEEE


(a)

(b)

Fig. 1. Example of video frames from the database in [9]: (a) Reference frame; (b) Target frame with the pink bottle on the top right as an abandoned object.

3. ROBUST SUBSPACE RECOVERY

Recent works (e.g. [10]) have shown that traditional tools for dimensionality reduction, such as principal component analysis (PCA), can yield incorrect results in the presence of measurement errors and noise, which are quite common in real data. To address this problem, solutions like robust PCA [10] (RPCA) have been proposed. If $X = [x_1 x_2 \dots x_n]$ is a data matrix where each column x_i is a sample vector in $\mathbb{R}^m$, the RPCA decomposes X as $X = L + E$, where L is a low-rank data matrix and E is an sparse error matrix. RPCA can be effective in cases where the data can be modeled as belonging to a single low-rank subspace. However, in the more general case where the data can only be modeled as lying in a union of multiple subspaces (UoS) [11, 12] the RPCA can lead to incorrect model recovery.

Such case is addressed by the robust subspace recovery (RoSuRe) algorithm [7], which represents the data matrix as

$$X = LW + E, \qquad (1)$$

where $L = [L_1| \dots |L_k]$ denotes the union of subspaces, each L_j being a representative sampling matrix of a subspace S_j, and W is a sparse block-diagonal matrix such that $LW = L$, with $W_{ii} = 0$. The RoSuRe algorithm assumes the sparsity of both W and E matrices [7] to obtain the representation in Eq. (1). It does so by solving the non-convex optimization problem

$$\min_{W,E} ||W||_1 + \lambda ||E||_1, \text{ s.t. } X = L + E, LW = L, W_{ii} = 0, \qquad (2)$$

using the augmented Lagrangian multiplier method with the augmented Lagrangian function [7]

$$\mathcal{L}(W, E, Y, \mu) = ||W||_1 + \lambda ||E||_1 + \underbrace{\langle LW - L, Y\rangle + \frac{\mu}{2}||LW - L||^2}_{f(W,E)}, \qquad (3)$$

where f is the differential part of $\mathcal{L}$, which is bilinear in W and E. Note that L is redundant and could be replaced by $(X - E)$ in f. Using the soft-threshold operator τ_α and letting $\hat{W}_k = I - W_k$, one can define the update steps of W and E such as [7]:

$$\begin{aligned} W_{k+1} &= \arg\min_{W} ||W||_1 + f(W, E) \\ &= \tau_{\frac{\lambda}{\mu\eta_1}} \left[W_k - \frac{1}{\eta_1} \nabla_W f(W_k, E_k) \right] \\ &= \tau_{\frac{\lambda}{\mu\eta_1}} \left[W_k + \frac{1}{\eta_1} L_{k+1}^T \left(L_{k+1} \hat{W}_k - \frac{Y_k}{\mu_k} \right) \right], \end{aligned} \qquad (4)$$

$$\begin{aligned} E_{k+1} &= \arg\min_{E} \lambda ||E||_1 + f(W, E) \\ &= \tau_{\frac{1}{\mu\eta_2}} \left[E_k - \frac{1}{\eta_2} \nabla_E f(W_{k+1}, E_k) \right] \\ &= \tau_{\frac{1}{\mu\eta_2}} \left[E_k + \frac{1}{\eta_2} \left(L_{k+1} \hat{W}_{k+1} - \frac{Y_k}{\mu} \right) \hat{W}_{k+1}^T \right], \end{aligned} \qquad (5)$$

where $\eta_1 \geq ||L||_2^2$ and $\eta_2 \geq ||\hat{W}||_2^2$, as summarized in Algorithm 1.

Algorithm 1 - Robust Subspace Pursuit (RoSuRe) [7].

```
Input: $X, \lambda, \rho > 1, \eta_1, \eta_2$
while not converged do
    $L_{k+1} = X - E_k$
    $W_{k+1} = \tau_{\frac{\lambda}{\mu\eta_1}} \left[ W_k + \frac{1}{\eta_1} L_{k+1}^T \left( L_{k+1} \hat{W}_k - \frac{Y_k}{\mu_k} \right) \right]$
    $(W_{k+1})_{ii} = 0$
    $\hat{W}_{k+1} = I - W_{k+1}$
    $E_{k+1} = \tau_{\frac{1}{\mu\eta_2}} \left[ E_k + \frac{1}{\eta_2} \left( L_{k+1} \hat{W}_{k+1} - \frac{Y_k}{\mu_k} \right) \hat{W}_{k+1}^T \right]$
    $Y_{k+1} = Y_k + \mu_k \left( L_{k+1} W_{k+1} - L_{k+1} \right)$
    $\mu_{k+1} = \rho \mu_k$
end
```

4. ROSURE MODIFICATION FOR THE MOVING CAMERA CASE

The basic assumption underlying the proposed modification is that the moving surveillance camera is sufficiently slow. This way, adjacent video frames can be considered sufficiently similar to share the same low-rank representation. In this scenario, let X_r be a data matrix such that each column is a reference video frame. We start by decomposing X_r according to Eq. (1) using the RoSuRe algorithm, that is, $X_r = L_r W_r + E_r$. L_r is the low-rank linear part of the reference video and $E_r = (X_r - L_r)$ is its non-linear complement, that provides a sparse error signal.

Algorithm 2 - Sparse representation of X given the low-rank representation L.

```
*Input*: $L, X, \lambda, \rho > 1, \eta_1, \eta_2$
**while** not converged **do**
    $L'_{k+1} = X - E_k$
    $W_{k+1} = \tau_{\frac{\lambda}{\mu\eta_1}}\left[W_k - \frac{1}{\eta_1}L^T\left(LW_k - L'_{k+1} + \frac{Y_k}{\mu_k}\right)\right]$
    $E_{k+1} = \tau_{\frac{1}{\mu\eta_2}}\left[E_k - \frac{1}{\eta_2}\left(LW_{k+1} - L'_{k+1} + \frac{Y_k}{\mu_k}\right)\right]$
    $Y_{k+1} = Y_k + \mu_k\left(LW_{k+1} - L'_{k+1}\right)$
    $\mu_{k+1} = \rho\mu_k$
**end**
```

Assuming that the target frame sequence X_t shares the same low-rank structure with its reference counterpart, one can rewrite the low-rank part of X_t as a combination of the linear low-rank information of the reference sequence plus an error signal. In other words, one can find a W_t matrix, such that target-video matrix can be written as $X_t = L_rW_t + E_t$, with both W_t and E_t as sparse matrices. Using this description, all anomalies in X_t, such as an abandoned object, are encapsulated into E_t. In order to perform this alternate representation of X_t, taking advantage of L_r determined in the sparse representation of X_r, the RoSuRe algorithm must be modified to work with a given fixed low-rank term L_r. To do so, the cost function in Eq. (2) should be modified to

$$\min_{E,W} = ||W||_1 + \lambda||E||_1 \quad \text{s.t. } L_rW = X - E, \tag{6}$$

and, following equation (3), the augmented Lagrangian function becomes

$$\mathcal{L}'(E, W, Y, \mu) = ||W||_1 + \lambda||E||_1 \\ \underbrace{+\langle L_rW - X + E, Y\rangle + \frac{\mu}{2}||L_rW - X + E||^2}_{g(W,E)}. \tag{7}$$

In a similar way to f in Eq. (3), g is the smooth part of the Lagrangian and will be used to compute the update steps of W_k and E_k as follows:

$$\begin{aligned} W_{k+1} &= \arg\min_W ||W||_1 + g(W, E) \\ &= \tau_{\frac{\lambda}{\mu\eta_1}}\left[W_k - \frac{1}{\eta_1}\nabla_W g(W_k, E_k)\right] \\ &= \tau_{\frac{\lambda}{\mu\eta_1}}\left[W_k - \frac{1}{\eta_1}{L_r}^T\left(L_rW_k - X + E_k + \frac{Y_k}{\mu_k}\right)\right], \quad (8) \\ E_{k+1} &= \arg\min_E \lambda||E||_1 + g(W, E) \\ &= \tau_{\frac{1}{\mu\eta_2}}\left[E_k - \frac{1}{\eta_2}\nabla_E g(W_{k+1}, E_k)\right] \\ &= \tau_{\frac{1}{\mu\eta_2}}\left[E_k - \frac{1}{\eta_2}\left(L_rW_{k+1} - X + E_k + \frac{Y_k}{\mu}\right)\right], \quad (9) \end{aligned}$$

as summarized in Algorithm 2.

Summing it up, in the proposed abandoned-object detection algorithm, we first perform Algorithm 1 to decompose X_r into a low-rank data matrix L_r and its sparse, non-linear complement E_r, such that

$$X_r = L_rW_r + E_r. \tag{10}$$

Second, we use the low-rank L_r matrix obtained in Eq. (10) as the input L for Algorithm 2 to decompose X_t into

$$X_t = L_rW_t + E_t. \tag{11}$$

As mentioned before, this procedure tends to isolate in E_t all the target-sequence information that is not present in L_r (or X_r). However, besides the residual generated by the abandoned-object, E_t will also have the sequence's high frequency information that could not be captured by the low-rank representation L_rW_t. The abandoned-object information contained in E_t can be separated from its inherent high frequency information by noting that, as X_t and X_r are similar by assumption, E_t will in general look quite similar to E_r, except around the abandoned object. Therefore, we also perform a third and last step, decomposing E_t using E_r as the input dictionary L of Algorithm 2, yielding

$$E_t = E_rW + E. \tag{12}$$

The remaining sparse error E tends to contain, as desired, just the abandoned objects in X_t not present in X_r, as illustrated in Section 5.

5. EXPERIMENTAL RESULTS

The proposed strategy was tested for several video sequences from the database presented in [8] and available from [9]. In our experiments, the video sequences were spatially subsampled to 320×180 pixels. In order to guarantee a proper representation of the target video sequence using the low-rank approximation of the reference video, we have to guarantee that all frames of the target video sequence have a corresponding frame in the reference video sequence. One way of guaranteeing this is to make sure that the reference video contains at least one complete turn of the camera through the surveillance area. In case there are computational complexity limitations and a complete surveillance turn of the reference video cannot be processed, one can follow a divide-and-conquer approach. We do so by segmenting both the reference and target videos with the care that each segment of the reference sequence contains all the corresponding frames of the target sequence.

Results for four different abandoned-object situations are illustrated in Fig. 2. In each case, a 70-frame reference sequence is used to model a 50-frame target sequence contained within the reference sequence. In all steps, the representation Algorithms 1 and 2 employed $\lambda = 1$ and $\rho = 1.5$.

In this figure, each column represents a given experiment and each row shows a sample frame of the matrices described in equations (10), (11), and (12). Error matrices are visualized in terms of their absolute values, and the last row contains the objects and other differences detected in each of the target frames shown in the second row. One can see that the proposed method can detect the abandoned objects while having very few false positives. Illumination changes, as well as major camera misalignments, may also lead to false positives.

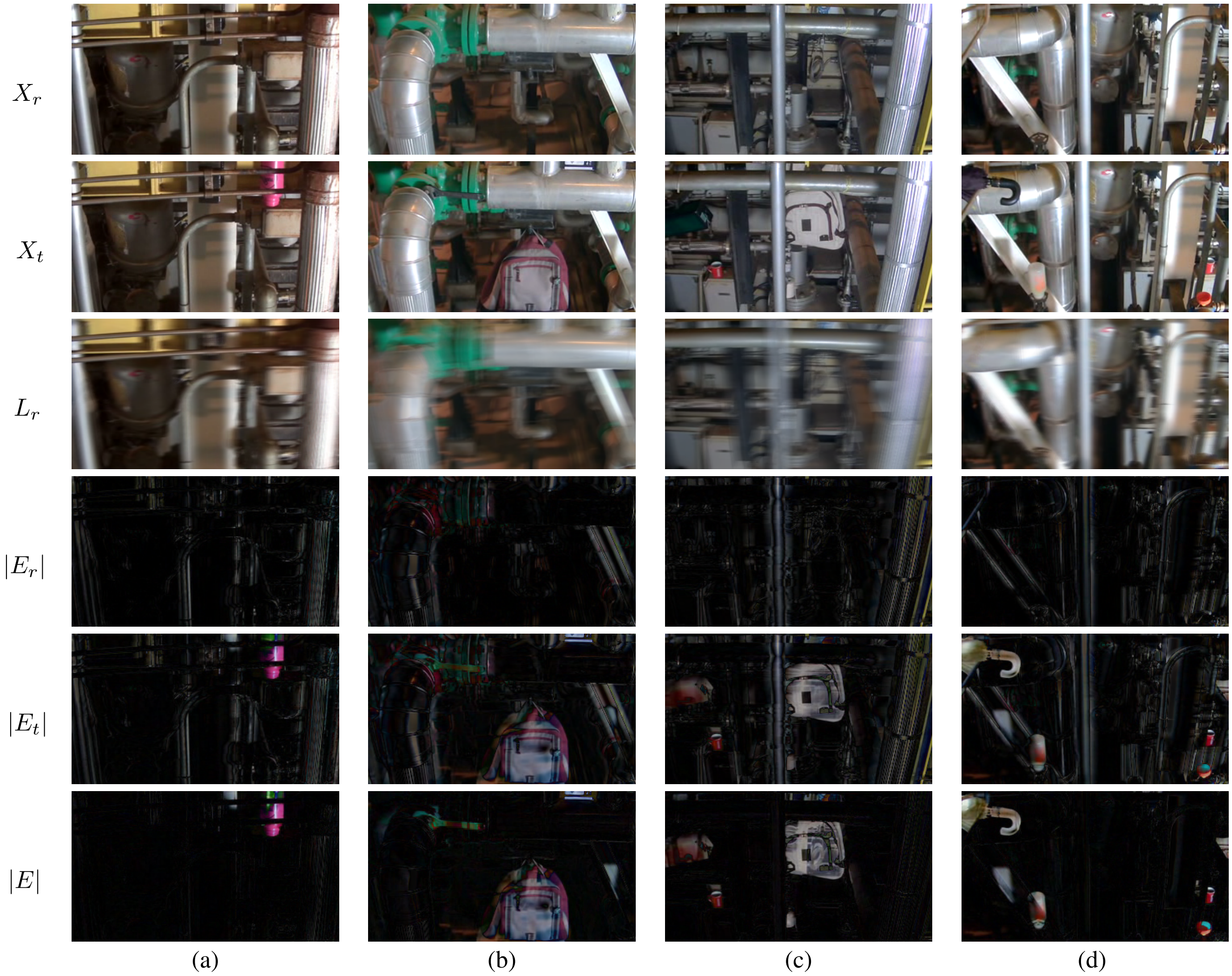


Fig. 2. Experimental results (single frames of matrices X_r, X_t, L_r, E_r, E_t, and E) using proposed low-rank representation for 4 different abandoned-object scenarios: (a) pink bottle; (b) backpack + wrench + box; (c) backpack + green box + mug + string roll; (d) umbrella + bottle + bottle cap + mug.

False negatives may occur if the abandoned object has the same pixel intensities as the background, as seen in the middle of the wrench in the second experiment (Fig. 2b). Most of these artifacts, however, can be removed by a simple post-processing such as median filtering.

It is important to note that the proposed method can even detect the object shadows, as verified on the left of each bottle in the first and fourth experiments (Fig. 2a and Fig. 2d).

6. CONCLUSIONS

We have presented a new approach for detecting changes in moving camera captured videos sequences by applying sparse representations based on the RoSuRe technique. We use the RoSuRe algorithm three times. In the first pass, the reference video is decomposed using the RoSuRe algorithm. In the second pass, the target sequence is decomposed based on the low-rank reference component determined in the first pass. In the third and last pass, the sparse error of the first pass is used as the low-rank component to represent the second-pass sparse error. Following this strategy, the sparse error of this last pass tends to contain only the abandoned objects, being free from the high frequency components inherent to the two video sequences. The results are promising and the method performs very well for all the examples, with the advantage of alleviating the need of geometric registration, needing only a very loose video synchronization. The presented results are very encouraging and promise a true viability of a deployable system of change detection methods for moving cameras using the proposed framework.

ACKNOWLEDGEMENTS

This work was partially funded by the grants DE-NA0002576, W911NF-04-D-0003, and CNPq/MCTI 0890952735408505.

7. REFERENCES

[1] H. Kong, J.-Y. Audibert, and J. Ponce, “Detecting abandoned objects with a moving camera,” *IEEE Trans. Image Processing*, vol. 19, no. 8, pp. 2201-2210, Aug. 2010.

[2] T. Zhou and D. Tao, “Greedy bilateral sketch, completion & smoothing”, *in Proc. Int. Conf. Artificial Intelligence and Statistics*, pp. 650-658, 2013.

[3] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210-227, Feb. 2009.

[4] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, “Sparse representation for computer vision and pattern recognition,” *Proc. IEEE*, vol. 98, no. 6, pp. 1031-1044, June 2010.

[5] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, “Toward a practical face recognition system: Robust alignment and illumination by sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372-386, Feb. 2012.

[6] X. Cui, J. Huang, S. Zhang, and D. N. Metaxas, “Background subtraction using low rank and group sparsity constraints,” *Proc. European Conf. Comp. Vision 2012*, LNCS vol. 7572, pp. 612-625, Oct. 2012,

[7] X. Bian and H. Krim, “Robust subspace recovery via bi-sparsity pursuit,” arXiv:1403.8067v2, Apr. 2014.

[8] A. F. da Silva, L. A. Thomaz, G. Carvalho, M. T. Nakahata, E. Jardim, J. F. L. de Oliveira, E. A. B. da Silva, S. L. Netto, G. Freitas, and R. R. Costa, “An annotated video database for abandoned-object detection in a cluttered environment,” *Proc. Int. Telecommunications Symp.*, São Paulo, Brazil, Aug. 2014.

[9] VDAO - Video Database of Abandoned Objects in a Cluttered Industrial Environment [Online]. Available: http://www.smt.ufrj.br/∼tvdigital/database/objects

[10] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis,” *J. ACM*, vol. 58, no. 3, pp. 11:1-37, June 2011.

[11] M. Soltanolkotabi and E. J. Candès, “A geometric analysis of subspace clustering with outliers,” *The Annals of Statistics*, vol. 40, no. 4, pp. 2195-2238, 2012.

[12] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, “Robust recovery of subspace structures by low-rank representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171-184, Jan. 2013.

[13] Y. Sheikh, O. Javed, and T. Kanade, “Background subtraction for freely moving cameras,” *Proc. IEEE Int. Conf. Comput.Vision*, pp. 1219-1225, 2009.

[14] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao, “Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization,” *Neural Information Processing Systems*, 2009.

[15] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, “Fast l_1-minimization algorithms for robust face recognition,” *IEEE Trans. Image Processing*, vol. 22, no. 8, pp. 3234-3246, Aug. 2013.