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 \(\times\) 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.
using the augmented Lagrangian multiplier method with the augmented Lagrangian function [7]

$$\mathcal{L}(W, E, Y, \mu) = \|W\|_1 + \lambda \|E\|_1 + (LW - L, Y) + \frac{\mu}{2} \|LW - L\|^2, \quad \text{(3)}$$

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 $\tau_\alpha$ and letting $\hat{W}_k = I - W_k$, one can define the update steps of $W$ and $E$ such as [7]:

$$W_{k+1} = \arg \min_W \|W\|_1 + f(W, E)$$

$$\begin{align*}
&= \tau_{\frac{1}{\eta_1}} \left[ W_k - \frac{1}{\eta_1} \nabla_W f(W_k, E_k) \right] \\
&= \tau_{\frac{1}{\eta_1}} \left[ W_k + \frac{1}{\eta_1} L_{k+1}^T \left( L_{k+1} \hat{W}_k - \frac{Y_k}{\mu_k} \right) \right], \quad \text{(4)}
\end{align*}$$

$$E_{k+1} = \arg \min_E \lambda \|E\|_1 + f(W, E)$$

$$\begin{align*}
&= \tau_{\frac{1}{\eta_2}} \left[ E_k - \frac{1}{\eta_2} \nabla_E f(W_{k+1}, E_k) \right] \\
&= \tau_{\frac{1}{\eta_2}} \left[ E_k + \frac{1}{\eta_2} \left( L_{k+1} \hat{W}_{k+1} - \frac{Y_k}{\mu} \right) \hat{W}_{k+1}^T \right], \quad \text{(5)}
\end{align*}$$

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**

- $L_{k+1} = X - E_k$
- $W_{k+1} = \tau_{\frac{1}{\eta_1}} \left[ W_k + \frac{1}{\eta_1} L_{k+1}^T \left( L_{k+1} \hat{W}_k - \frac{Y_k}{\mu_k} \right) \right]$
- $(W_{k+1})_{ii} = 0$
- $\hat{W}_{k+1} = I - W_{k+1}$
- $E_{k+1} = \tau_{\frac{1}{\eta_2}} \left[ E_k + \frac{1}{\eta_2} \left( L_{k+1} \hat{W}_{k+1} - \frac{Y_k}{\mu} \right) \hat{W}_{k+1}^T \right]$
- $Y_{k+1} = Y_k + \mu_k \left( L_{k+1} \hat{W}_{k+1} - L_{k+1} \right)$
- $\mu_{k+1} = \rho \mu_k$

**end**

4. ROBUST SUBSPACE RECOVERY 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, $X_r = L_r \hat{W}_r + E_r$. $L_r$ 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

$L_{k+1} = X - E_k$

$W_{k+1} = \frac{\tau}{\eta_1} \left[ W_k - \frac{1}{\eta_1} L^T \left( LW_k - L'_{k+1} + \frac{Y_k}{\mu_k} \right) \right]$

$E_{k+1} = \frac{\tau}{\eta_2} \left[ E_k - \frac{1}{\eta_2} \left( LW_k - L'_{k+1} + \frac{Y_k}{\mu_k} \right) \right]$

$Y_{k+1} = Y_k + \mu_k \left( LW_k - L'_{k+1} \right)$

$\mu_{k+1} = \rho \mu_k$

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 $X_t = L_t W_t + E_t$, 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 $L_t$ determined in the sparse representation of $X_t$, the RoSuRe algorithm must be modified to work with a given fixed low-rank term $L_t$. To do so, the cost function in Eq. (2) should be modified to

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

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

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

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:

$$W_{k+1} = \arg \min_{W} ||W||_1 + g(W, E)$$

$$= \frac{\tau}{\eta_1} \left[ W_k - \frac{1}{\eta_1} \nabla_W g(W_k, E_k) \right]$$

$$= \frac{\tau}{\eta_1} \left[ W_k - \frac{1}{\eta_1} L_t^T \left( L_t W_k - X + E_k + \frac{Y_k}{\mu_k} \right) \right], \quad (8)$$

$$E_{k+1} = \arg \min_{E} \lambda \||E||_1 + g(W, E)$$

$$= \frac{\tau}{\eta_2} \left[ E_k - \frac{1}{\eta_2} \nabla_E g(W_{k+1}, E_k) \right]$$

$$= \frac{\tau}{\eta_2} \left[ E_k - \frac{1}{\eta_2} \left( L_t W_{k+1} - X + E_k + \frac{Y_k}{\mu_k} \right) \right], \quad (9)$$

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

$$X_r = L_r W_r + E_r. \quad (10)$$

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

$$X_t = L_r W_t + E_t. \quad (11)$$

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 $L_r W_t$. 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

$$E_t = E_r W + E. \quad (12)$$

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 $320 \times 180$ 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.
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


