Detection of Abandoned Objects Using Robust Subspace Recovery with Intrinsic Video Alignment

Lucas A. Thomaz, Allan F. da Silva, Eduardo A. B. da Silva, Sergio L. Netto
PEE - COPPE
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
Email: {lucas.thomaz, allan.silva, eduardo, sergioln}@smt.ufrj.br

Hamid Krim
ECE
North Carolina State University
North Carolina, USA
Email: ahk@ncsu.edu

***Abstract*—The detection of abandoned objects in videos from moving cameras is of great importance to automatic surveillance systems that monitor large and visually complex areas. This paper proposes a new method based on sparse decompositions to identify video anomalies associated with abandoned objects. The proposed scheme inherently incorporates synchronization between the reference (anomaly-free) and target (under analysis) sequences thus reducing the implementation complexity of the overall surveillance system. Results indicate that the proposed video-processing scheme can lead to 95% complexity reduction while maintaining excellent detection capability of foreground objects.**

I. Introduction

In present days video surveillance systems are almost everywhere, as most public places have one or more fixed cameras performing security tasks. According to a report from Transparency Market Research from 2016 [1], for instance, the video surveillance equipment market worldwide is expected to reach around U$43 billions by 2019. All this surveillance infrastructure generates a huge amount of video data that humans most likely will not be able to analyze properly. To deal with this issue several algorithms have been designed to process the video streams automatically and extract significant and reliable information [2], [3].

In some of these applications the camera may be too expensive to be attached to a single place overlooking a specific section of the environment. In this case a possible solution is to install the camera on a mobile platform that covers a wide surveillance area. Some solutions have been proposed in the last few years to cope with the problem of detecting abandoned objects in such scenarios: in [4] a camera mounted on a car performs the detection of objects on streets; in [5] a camera mounted on a robot platform detects abandoned objects in an industrial environment; and in [6] a camera placed on a train detects the presence of objects on the rails.

This paper deals with a class of such anomaly detection methods, that is based on sparse representations of the reference (certified anomaly-free) and target (to be processed) video sequences. One successful example of such methods is the one in [5], that uses robust subspace recovery [7], [8] to detect abandoned objects in a cluttered industrial plant using a single camera mounted on a robotic platform that follows a predefined path. Although presenting good detection performance, the method in [5] requires the previous temporal alignment of the reference and target videos. The present paper then introduces a novel method where the video synchronization procedure is inherent to the sparse representation, greatly reducing the overall computational complexity of the entire anomaly-detection procedure.

To introduce the proposed techniques, the remainder of this work is organized as follows: Section II reviews the method in [5] and Section III describes the proposed method in which temporal alignment can be obtained from the sparse decomposition used for anomaly detection. In Section IV the experimental results are presented and discussed, and in Section V the authors' conclusions are provided.

II. Moving-Camera RoSuRe

In this section we describe the work in [5] to establish the grounds for the proposed contributions. In that work a sparse decomposition technique, referred to as Robust Subspace Recovery (RoSuRe) [7], [8], represents a data matrix $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{LW} + \mathbf{E}, \tag{1}$$

where $\mathbf{L}$ is a low-rank representation of $\mathbf{X}$ that is also a union of low-rank subspaces, matrix $\mathbf{W}$ is a block-diagonal sparse matrix that specifies how the subspaces of $\mathbf{L}$ should be combined to approximate $\mathbf{X}$, and $\mathbf{E}$ is a residue matrix. In this framework $\mathbf{L}$ can be non-trivially represented by its subspaces, such that $\mathbf{LW} = \mathbf{L}$, with $\mathbf{W}_{ii} = 0$. A fundamental feature of the RoSuRe algorithm is that both $\mathbf{W}$ and $\mathbf{E}$ matrices are sparse. To obtain such representation one performs the following minimization using convex optimization techniques [9]:

$$\min_{\mathbf{W},\mathbf{E}} ||\mathbf{W}||_1 + \lambda||\mathbf{E}||_1,\ s.t.\ \mathbf{X} = \mathbf{L} + \mathbf{E},\ \mathbf{LW} = \mathbf{L},\ \mathbf{W}_{ii} = 0. \tag{2}$$

In [5] an anomaly-detection algorithm for moving cameras was developed based on the RoSuRe scheme. In such approach one assumes that the surveillance camera moves slowly enough so that consecutive frames of the video share a low-rank representation. Therefore, the so-called moving camera RoSuRe (mcRoSuRe) algorithm starts by decomposing the reference-video matrix $\mathbf{X}_r$ using Eq. (1), such that

$$\mathbf{X}_r = \mathbf{L}_r\mathbf{W}_r + \mathbf{E}_r, \tag{3}$$

$$\mathbf{E}_r = \mathbf{X}_r - \mathbf{L}_r, \tag{4}$$

where $\mathbf{L}_r$ is the low-rank representation of $\mathbf{X}_r$ and $\mathbf{E}_r$ is its sparse complement. If we consider that the low-rank representation of the target sequence $\mathbf{X}_t$ is also $\mathbf{L}_r$, then it is possible to rewrite $\mathbf{X}_t$ using a formulation analogous to Eq. (1), yielding

$$\mathbf{X}_t = \mathbf{L}_r\mathbf{W}_t + \mathbf{E}_t, \tag{5}$$

978-1-4673-6853-7/17/$31.00 ©2017 IEEE

with $\mathbf{W}_t$ and $\mathbf{E}_t$ both being sparse matrices. The step-by-step optimization algorithm used to perform this decomposition is detailed in [5]. In this decomposition all the information of $\mathbf{X}_t$ that could not be represented from $\mathbf{L}_r\mathbf{W}_t$ is cast upon $\mathbf{E}_t$. This is just the case of the anomalies in the target sequence ($\mathbf{X}_t$), that are not present in neither the reference matrix $\mathbf{X}_r$ nor in $\mathbf{L_r}$. However, the information in $\mathbf{E}_t$ is not restricted to only the anomalies in the target video. As $\mathbf{L}_r$ is a low-rank matrix it is not able to represent neither high-frequency components of the target sequence ($\mathbf{X}_t$) nor the components caused by the absence of registration between the two video sequences. Both of these terms, however, are supposed to be common between the reference and target videos and therefore can be extracted from $\mathbf{E}_t$ using the similar decomposition

$$\mathbf{E}_t = \mathbf{E}_r\mathbf{W}_e + \mathbf{E}_e. \tag{6}$$

If this final decomposition is performed correctly matrix $\mathbf{E}_e$ will contain only the desired anomalies in the target video.

A summarized version of the mcRoSuRe algorithm is presented in Algorithm 1.

Algorithm 1 Moving Camera RoSuRe

Require: $\mathbf{X}_r$, $\mathbf{X}_t$

$\min_{\mathbf{W}_r,\mathbf{E}_r} ||\mathbf{W}_r||_1 + \lambda||\mathbf{E}_r||_1$, s.t. $\mathbf{X_r} = \mathbf{L_r} + \mathbf{E_r}$, $\mathbf{L}_r\mathbf{W}_r = \mathbf{L}_r$, $\mathbf{W}_{r_{ii}} = 0$

$\min_{\mathbf{W}_t,\mathbf{E}_t} ||\mathbf{W}_t||_1 + \lambda||\mathbf{E}_t||_1$, s.t. $\mathbf{L_r}\mathbf{W_t} = \mathbf{X_t} - \mathbf{E_t}$

$\min_{\mathbf{W}_e,\mathbf{E}_e} ||\mathbf{W}_e||_1 + \lambda||\mathbf{E}_e||_1$, s.t. $\mathbf{E_r}\mathbf{W_e} = \mathbf{E_t} - \mathbf{E_e}$

As can be verified in [5], this algorithm has a good performance when the anomalies to be detected using a moving camera are abandoned objects in visually complex environments. Although this algorithm in principle does not require time alignment between the reference and target videos, its optimization steps are computationally very expensive. Due to this fact, experiments in [5] could only be run in very small video sequences, e.g., of dimensions 320×180 pixels, with 70-frame long reference and 50-frame long target videos.

III. Target localization using the mcRoSuRe-TA algorithm

In this section we propose the mcRoSuRe-TA (mcRoSuRe - Temporal Alignment) algorithm, which is a modification of the mcRoSuRe algorithm that inherently incorporates the temporal sequence alignment between the reference and target videos. The mcRoSuRe-TA has the advantage of saving a lot of computation by reducing the unnecessarily large search space in the optimization steps within the standard mcRoSuRe approach.

An example of a mcRoSuRe $\mathbf{W}_t$ matrix (see Eq. (5)) can be seen in Fig. 1. By analyzing its structure it is possible to see that these sparse matrices bear some information that can be useful for aligning the reference and target videos - the significant $\mathbf{W}_t$ coefficients (white) indicate which frames of the low rank reference video $\mathbf{L_r}$ influence each frame of the reference video $\mathbf{X}_t$ in the sparse model of Eq. (5). However, since the white stripe in $\mathbf{W}_t$ is quite wide, such information is not very helpful. This is so because a wide white stripe means that one frame in $\mathbf{X}_t$ corresponds to too many frames in $\mathbf{L_r}$, which implies that proper frame alignment requires a search in a space that is still too large.

Figure 1. $\mathbf{W}_t$ matrix from the mcRoSuRe method. The brighter pixels denote higher values in the matrix. The vertical dimension corresponds to the target video and the horizontal dimension to the reference video.

In the experiments performed in this paper, short target videos, containing only a small amount of frames that may contain abandoned objects (e.g., 200 frames) should be temporally aligned with long reference videos (of about 6000 frames). In order to use the $\mathbf{W}_t$ matrix for alignment purposes it is necessary to find first the region of the reference video that has greater correlation with the target video being processed.

As pointed out above, to reduce the search space one should have a $\mathbf{W}_t$ matrix with a narrow white stripe. The main reason for the wide white stripe in $\mathbf{W}_t$ depicted in Fig. 1 is that $\mathbf{L_r}$ is low rank, which makes the frame correspondences between $\mathbf{X}_t$ and $\mathbf{L_r}$ somewhat fuzzy. Therefore, we propose to use the original reference data matrix $\mathbf{X}_r$ instead of the low-rank matrix $\mathbf{L}_r$ to reconstruct the target video data matrix $\mathbf{X}_t$, thus modifying Eq. (5) to

$$\mathbf{X}_t = \mathbf{X}_r\mathbf{W}_t + \mathbf{E}_t. \tag{7}$$

The advantages brought by this modification are twofold: the first concerns the alignment precision and the second the algorithm complexity.

The improvement in alignment precision can be better explained as follows: as the matrix $\mathbf{L}_r$ is a low-rank representation, it does not contain the high-frequency components of the reference video $\mathbf{X_r}$. Thus, when one tries to use $\mathbf{W}_t$ generated by Eq. (5) for this purpose the result is not precise enough since $\mathbf{X}_t$ still has high-frequency components that are not well mapped by $\mathbf{L}_r$. As can be seen in Fig. 2 matrix $\mathbf{W}_t$ generated with Eq. (5) has a wide white stripe, that is, it does not show precisely which frames in $\mathbf{L_r}$ correspond to the proper target video frames. On the other hand, matrix $\mathbf{W}_t$ generated with Eq. (7) presents a more sharp delimitation of this region, as illustrated in the same figure.

(a) Matrix $\mathbf{W}_t$ generated with Eq. (5).

(b) Matrix $\mathbf{W}_t$ generated with Eq. (7).

Figure 2. Matrices $\mathbf{W}_t$ generated with: (a) Eq. (5); (b) Eq. (7). Brighter pixels denote higher values in the matrices. The vertical dimension corresponds to the target video frames and the horizontal dimension to the reference video frames. In (b) the region of the reference video that corresponds to the given target video section is more sharply delimited.

The lower complexity for the resulting algorithm comes first from the fact that for calculating the $\mathbf{W}_t$ in the mcRoSuRe algorithm (Eq. (5)) one must generate $\mathbf{L}_r$, while if one uses

$\mathbf{X}_r$ instead of $\mathbf{L}_r$ to decompose $\mathbf{X}_t$ one saves that initial computation. Another important factor is that the matrix $\mathbf{W}_t$ from Eq. (7) has a thinner white stripe than the one from Eq. (5), which implies a more precise mapping for the reference frames used to reconstruct the target-video frames.

In addition, by using the proposed mcRoSuRe-TA technique to align the reference and target video sequences one can perform the anomaly detection using only the reference video excerpt that corresponds to the target video sequence under analysis. To do so it suffices to crop the reference video frames using the $\mathbf{W}_t$ as a guide, as detailed in Fig. 3. In such scenario, the selected frames compose a new reference excerpt that we refer to as $\mathbf{X'}_r$ that has much less columns than $\mathbf{X_r}$, which greatly reduces the resulting computational complexity.

(a) (b)

Figure 3. By using $\mathbf{W}_t$ in Eq. 7 one can select frames of the reference video that correspond to the target video frames and create a smaller reference matrix $\mathbf{X'_r}$ that contains only the relevant reference frames for processing the target frames. Using such a smaller reference sequence saves a lot of computation. (a) Represents the $\mathbf{W}_t$ obtained using the whole reference matrix $\mathbf{X}_r$ (b) Represents the $\mathbf{W}_t$ obtained using the cropped version of the reference matrix $\mathbf{X'}_r$

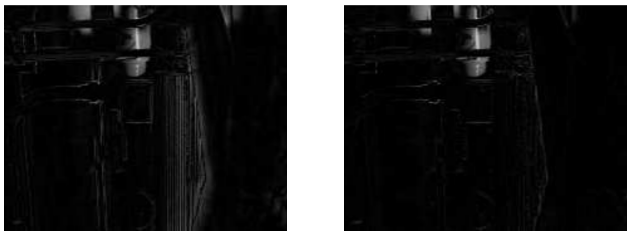
After composing matrix $\mathbf{X'}_r$ and performing again the decomposition in Eq. (7), one obtains a matrix $\mathbf{W'}_t$ and a residual $\mathbf{E'}_t$. The resulting $\mathbf{W'}_t$ will look somewhat like Fig. 3(b). As in the original mcRoSuRe scheme, one then has to perform the decomposition of $\mathbf{E'}_t$ obtained from Eq. (7) with $\mathbf{E'}_r$ matrix (using Eq. (1) with $\mathbf{X'}_r$ matrix). Therefore, in the mcRoSuRe-TA scheme, besides the step in Eq. (7), three more steps are needed

$$\mathbf{X'}_r = \mathbf{L'}_r\mathbf{W'}_r + \mathbf{E'}_r, \tag{8}$$

$$\mathbf{X}_t = \mathbf{X'}_r\mathbf{W'}_t + \mathbf{E'}_t, \tag{9}$$

$$\mathbf{E'}_t = \mathbf{E'}_r\mathbf{W'}_e + \mathbf{E'}_e. \tag{10}$$

One example of the resulting residue matrix $\mathbf{E'}_t$ side by side with $\mathbf{E'}_e$ is shown in Fig. 4.



(a) Residue matrix $\mathbf{E'}_t$. (b) Residue matrix $\mathbf{E'}_e$.

Figure 4. There are fewer undesired artifacts in the residue matrix $\mathbf{E'}_e$ when compared with $\mathbf{E'}_t$.

Using the mcRoSuRe-TA algorithm it is possible to work the full abandoned object detection framework, obviating an external temporal alignment step, as described in Algorithm 2.

Algorithm 2 Object Detection Using Proposed mcRoSuRe-TA

$\min_{\mathbf{W}_t,\mathbf{E}_t} ||\mathbf{W}_t||_1 + \lambda||\mathbf{E}_t||_1$, s.t. $\mathbf{X}_\mathrm{t} = \mathbf{X}_\mathrm{r}\mathbf{W}_\mathrm{t} + \mathbf{E}_\mathrm{t}$,
Crop reference frames of interest based on $\mathbf{W}_t$ matrix. Create $\mathbf{X'}_r$.
$\min_{\mathbf{W'}_r,\mathbf{E'}_r} ||\mathbf{W'}r||_1 + \lambda||\mathbf{E'}r||_1$, s.t. $\mathbf{X'}_\mathrm{r} = \mathbf{L'}_\mathrm{r} + \mathbf{E'}_\mathrm{r}$,
$\mathbf{L'}_r\mathbf{W'}_r = \mathbf{L'}_r$, $\mathbf{W'}_{r_{ii}} = 0$
$\min_{\mathbf{W'}_t,\mathbf{E'}_t} ||\mathbf{W'}t||_1 + \lambda||\mathbf{E'}t||_1$, s.t. $\mathbf{X'}_\mathrm{r}\mathbf{W'}_\mathrm{t} = \mathbf{X}_\mathrm{t} - \mathbf{E'}_\mathrm{t}$
$\min_{\mathbf{W}_e,\mathbf{E}_e} ||\mathbf{W}_e||_1 + \lambda||\mathbf{E}_e||_1$, s.t. $\mathbf{E'}_\mathrm{r}\mathbf{W}_\mathrm{e} = \mathbf{E'}_\mathrm{t} - \mathbf{E'}_\mathrm{e}$

IV. Experimental Results

The proposed algorithm was tested using the VDAO database of abandoned objects (described in [10] and available at [11]). This database consists of several videos recorded in a visually cluttered, industrial environment, with a camera mounted on a moving platform that follows a linear back and forth path. Each video contains several back and forth passes of the camera. In the database several different objects are placed along the path, with every object configuration being recorded in two different lighting conditions. Every video in the database that has been recorded with an abandoned object configuration has a corresponding reference video with no abandoned objects. Also, every target video has at least one abandoned object.

Although the VDAO database videos are in color and with HD resolution, in the performed experiments they were downsampled to 320×180 resolution and converted to grayscale. Also, in the experiments only one forth pass of each video has been used, which corresponded to an average of 5000 frames per reference video. The target videos are segments of 200 frames that contain at least one abandoned object. All experiments were performed with the same machine configuration: an Intel i7-3630QM with 2.4GHz and 16GB of RAM, running MATLAB© 2012b.

Fig. 5 shows the comparison between the results from the mcRoSuRe [5] and proposed mcRoSuRe-TA algorithms. In the mcRoSuRe results a temporal pre-alignment step has been performed manually, that is, the reference video was selected in a way to ensure it would at least contain the same region shown in target video. In contrast, for mcRoSuRe-TA the alignment is performed automatically using matrix $\mathbf{W'}_t$ as explained above. Some sample frames of the intermediate steps of the algorithm are shown in Fig. 5 for comparative purposes. In each column a video containing one single object of the VDAO [11] database is represented, with each row representing a step of the respective algorithm.

These results show that the performance of the proposed mcRoSuRe-TA algorithm is at least as good as that of the mcRoSuRe algorithm. Both methods are able to detect correctly the abandoned objects in the given cluttered scenario. In addition, for both algorithms false negative detections are very rare and mostly caused by strong similarities between the intensity values of the object and the background in the reference video. Also, likewise the mcRoSuRe, most of the false positive detections in mcRoSuRe-TA are caused by camera jitter or lighting differences between the reference and target videos.

To assess the computational performance of both methods two different experiments were run. The first one used the same experimental setup of [5], where small 50-frame target videos were processed with reference videos only a few frames larger (70 frames, chosen manually, as explained above). In these conditions, both algorithm run in equivalent times. The second experiment used larger reference (900 to 5000 frames) and tar-

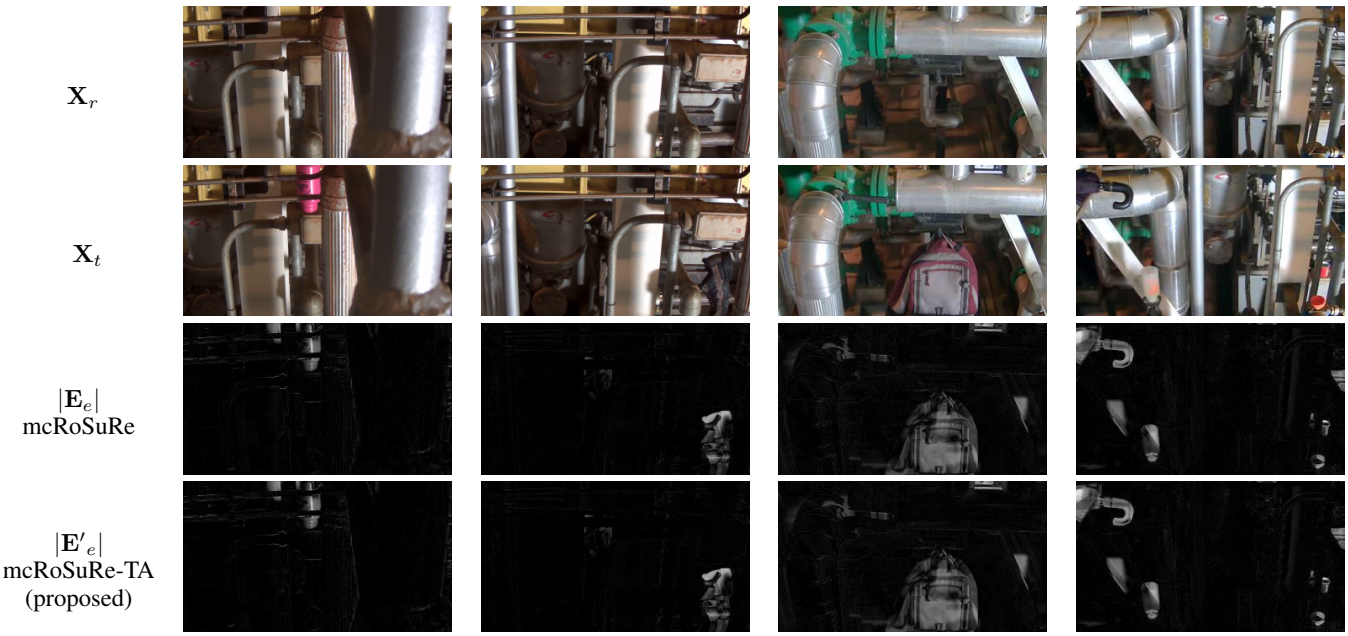


Figure 5. Comparative results between the two algorithms (single frames of matrices X_r, X_t, and E of both methods) for 4 different abandoned-object videos from VDAO [10] database: (a) pink bottle; (b) shoe (c) backpack + wrench + box; (d) umbrella + bottle + bottle cap + mug. The similar performance of both methods is clear from these experiments.

get (100 to 200 frames) video sequences. In this new and more realistic scenario the mcRoSuRe-TA algorithm was able to run at least 20 times faster than the mcRoSuRe, yet maintaining the same qualitative detection results. This happened because mcRoSuRe-TA could process the whole reference video much more efficiently, using the whole reference only in its first steps to obtain the matrix $\mathbf{W}_t$ (see Figs. 2 and 3). After this step the proposed scheme uses only the smaller reference videos, which provides significant savings in computation.

V. Conclusion

This paper proposed the mcRoSuRe-TA algorithm, an algorithm for detecting anomalies in a cluttered environment using a moving camera. The proposed scheme is a modification to the mcRoSuRe algorithm [5] that embeds a form of time alignment between the reference and target videos, restricting the search space for every optimization step and, therefore, reducing the resulting computational complexity. Unlike its standard counterpart mcRoSuRe, the proposed mcRoSuRe-TA can run using the full reference video without compromising the overall computational complexity. Results obtained using the VDAO database are encouraging as the method is able to detect effectively the abandoned objects with very few false detections.

Acknowledgment

This work was partially funded by CNPq and CAPES (Process Number 88881.135449/2016-01).

References

[1] Transparency Market Research,"Video Surveillance and VSaaS Market - Global Industry Analysis, Size, Share, Growth, Trends and Forecast 2016 - 2024", TMS Analisys, Albany, USA. April, 2016.

[2] L. A. Thomaz, A. F. da Silva, E. A. B. da Silva, S. L. Netto, X. Bian and H. Krim, "Abandoned object detection using operator-space pursuit," *Proc. International Conference on Image Processing*, Quebec City, Canada, pp. 1980-1984, Sept. 2015.

[3] C. Cuevas, R. Martínez and N. García, "Detection of stationary foreground objects: A survey," *Computer Vision and Image Understanding*, vol. 152, pp. 41-57, Nov. 2016.

[4] H. Kong, J.-Y. Audibert and J. Ponce, "Detecting abandoned objects with a moving camera, *IEEE Transactions on Image Processing*, vol. 19, no. 8, pp. 2201–2210, Aug. 2010,

[5] E. Jardim, X. Bian, E. A. B. da Silva, S. L. Netto and H. Krim, "On the detection of abandoned objects with a moving camera using robust subspace recovery and sparse representation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Australia, pp. 1295-1299, Apr. 2015.

[6] H. Mukojima, D. Deguchi,Y. Kawanishi, I. Ide, H. Murase, M. Ukai, N. Nagamine and R. Nakasone, "Moving camera background-subtraction for obstacle detection on railway tracks," *IEEE International Conference on Image Processing*, Phoenix, USA, pp. 3967-3971, Sept. 2016.

[7] X.Bian and H. Krim, "Robust subspace recovery via dual sparsity pursuit," http://arxiv.org/abs/1403.8067, Mar. 2014.

[8] X. Bian and H. Krim, "BI-sparsity pursuit for robust subspace recovery," *IEEE International Conference on Image Processing*, Quebec City, Canada, pp. 3535-3539, Sept. 2015.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, USA, 2004.

[10] 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. International Telecommunications Symposium*, Sao Paulo, Brazil, Aug. 2014.

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