ABANDONED OBJECT DETECTION USING OPERATOR-SPACE PURSUIT

Lucas A. Thomaz[†], Allan F da Silva[†], Eduardo A. B da Silva[†], Sergio L. Netto[†], Xiao Bian[‡], Hamid Krim[‡]

[†]PEE/COPPE, Federal University of Rio de Janeiro, Brazil,
[‡]North Carolina State University, USA.
{*lucas.thomaz, allan.freitas, eduardo, sergioln*}*@smt.ufrj.br;* {*xbian, ahk*}*@ncsu.edu*

ABSTRACT

This work presents a framework to be used in the detection of abandoned objects and other video events in a cluttered environment with a moving camera. In the proposed method a target video, that may have features we would like to detect, is compared with a pre-acquired reference video, which is assumed to have no objects nor video events of interest. The comparison is carried out by way of the achieved optimized operators, generated from the reference video, that produce Gaussian outputs when applied to it. Any anomaly of interest in the target video leads to a non-Gaussian output. The method dispenses with the target and reference videos being either synchronized or precisely registered, being robust to rotations and translations between the frames. Experiments show its good performance in the proposed environment.

***Index Terms*—** Object Detection, Operator Space, Cluttered Environment, Moving Camera

1. INTRODUCTION

In surveillance systems it is often desirable to reduce the influence of the human factor. This reduction tends to increase the process efficiency and minimize employee expenses as well as labor risks. In such a scenario, automatic environment surveillance techniques are valuable tools.

In the context of automatic surveillance, the simpler solutions involve fixed cameras covering the area of interest. To address them there are well studied and established techniques such as background subtraction, Gaussian-mixtures, and several statistical approaches as described in [1–5]. If one wants to cover a large area, however, many fixed cameras are usually needed, which implies substantial increases in equipment cost.

For these cases, more effective solutions are ones employing a single moving camera. However, there are several challenges involved in using a moving camera, as the need of temporal alignment and geometric registration to compensate the movement. Several works in the literature [6–10] address this kind of problem. The challenges are even greater when the environment of interest is visually complex, as is the case of cluttered backgrounds.

In this paper we address the problem of video surveillance in cluttered environments using a moving camera. As in most of the works in the literature, the approach proposed in this paper is based on the comparison of two video sequences, one that is considered the reference, with no objects or events we would like to detect, and the other that is the target, which may present objects or video events of interest.

The key proposal here is to perform video comparison in an operator domain. In that alternative representation, reference frames are associated to perfectly Gaussian images. Applying the same operator to the corresponding target frames may lead or not to Gaussian images, indicating the absence or the presence, respectively, of any event of interest such as an abandoned object. Unlike most previous works, the method proposed here does not need any kind of previous geometrical registration or detection of salient points in the video sequences. These are commonly used to make the frames look similar and thus to highlight the region of interest in the images. Also, the proposed method aims to be used in visually complex environments, where performing background modeling and using homographies are difficult tasks. In this sense our method needs only reference and target videos to be roughly time-aligned.

The remaining of this paper is organized as follows: Section 2 explains the technique upon which the main idea of the algorithm is based. A system overview is shown in Section 3, whereas the complete algorithm is introduced in Section 4 in a step-by-step manner. Section 5 shows experimental results illustrating the capabilities of the proposed system and the conclusions are presented in Section 6.

2. OPTIMAL OPERATOR-SPACE PURSUIT

In this paper we will address the problem of abandoned object detection with moving cameras using the main idea presented in [11]. In that work the authors propose to describe image sequences through a formalism of fiber bundles and construction of an operator space H which is homeomorphic to the manifold of hidden states of image sequences. The operator H can be used to categorize the image sequences by first developing an algorithm to find the optimal low-dimensional space where the discriminating information is compactly stored.

Being B the base manifold of the image sequence, we assume that H lies in a low dimensional sub-space of B. It is then desired to solve the constrained dimension minimization problem described as

$$\min \dim(H) \;\text{ s.t. }\; \|h_i(x_i) - g\|_2 \leq C, \;\; h_i \in H, \qquad (1)$$

where x_i, $i = 1, \ldots, m$ are frames of a given image sequence, C a constant, g is a two-dimensional Gaussian function and $h_i(x_i)$ is the application of the ith operator over the ith frame of the sequence.

Performing this minimization corresponds to finding the lowest rank matrix $H = \left[h_1 \cdots h_m\right]$ under the constraints of equation (1). However, this minimization is an NP-hard problem as shown in [12]. In [11] this problem is substituted by a constrained nuclear norm minimization, where equation (1) is replaced, in the Fourier domain, by

$$\min \|H\|_* \;\text{ s.t. }\; \|X_i h_i - g\|_2 \leq C, \;\; H = \left[h_1 \cdots h_m\right], \qquad (2)$$

where X_i, for $i = 1, 2, \ldots, m$, are diagonally structured matrices with Fourier coefficients of each frame on the diagonal [11] and once again C a constant and g is a two-dimensional Gaussian function.

One can use a more general form for equation (2):

$$\min \|H\|_* \text{ s.t. } \|A_i(H) - g\|_2 \leq C, \text{ for } i = 1, \ldots, m \quad (3)$$

where $A(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear operator.

For a matrix X, the singular-value threshold operator is defined as

$$D_\tau(X) = US_\tau(\Sigma)V^*, S_\tau(\Sigma) = \text{diag}\{(\sigma_i - \tau)_+\}, \quad (4)$$

where σ_i are the singular values of X and $u_+ = \max(0, u)$. This operator satisfies the following theorem, obtained from [12]:

Theorem 1: For each $\tau \geq 0$ and $Y \in \mathbb{R}^{m \times n}$, the singular-value threshold operator is the solution to

$$D_\tau(Y) = \arg\min_X \frac{1}{2}\|X - Y\|_F^2 + \tau\|X\|_*, \quad (5)$$

where $\|X\|_F$ is the Frobenius norm of X [13].

Reference [11] develops a modified version of the singular value thresholding algorithm adapted to equation (3). Using some techniques that may be found in [12] it is shown in [11] that the iteration that leads to the optimum H is given by

$$\begin{cases} H^k = D_\tau\left(\sum_i A_i^*(y_i^{k-1})\right) \\ \begin{bmatrix} y^k \\ s^k \end{bmatrix} = P\left(\begin{bmatrix} y^{k-1} \\ s^{k-1} \end{bmatrix} + \begin{bmatrix} b - A(X^k) \\ -\epsilon \end{bmatrix}\right) \end{cases}, \quad (6)$$

with b and ϵ constants, y^k being the projection of the frame x in a low-dimension subspace, and s^k are iterative regularization thresholds (as can be seen in [11]) and P being a projection operator given by

$$P(y, s) = \begin{cases} (y, s), & \|y\| \leq s \\ \frac{\|y\|+s}{2\|y\|}(y, s), & -\|y\| \leq s \leq \|y\| \\ (0, 0), & s \leq -\|y\| \end{cases}. \quad (7)$$

After the completion of the minimization process, H is a subspace spanned by matched-filters that can represent the images that form the original space of sequences. In this alternative representation, each filter is linked to an image of the sequence in the sense that if an image is used as input to the corresponding matched-filter the output would be the Gaussian function g.

Using the above logic the authors of [11] developed a metric to assess similarity between two video sequences. In order to do so, one of the sequences is chosen as the reference (X^q) and is used to generate the subspace of filters H^q. Then, each frame X_i^p of the other sequence is tested to measure its distance to the reference sequence, yielding the so-called *frame-to-sequence distance*

$$d(X_i^p, X^q) = \min_{j=1\ldots m} \|X_i^p h_j^q - g\|_2, \quad (8)$$

where g is the predicted Gaussian output to the matched input in the filter h^q.

In what follows we use the optimal subspace representation shown above to develop an abandoned-object detection method in videos from moving cameras. In the proposed framework, equation (8) is employed to measure the distance among frames from a reference and target videos.

3. OBJECT DETECTION: OPERATOR-SPACE PURSUIT APPROACH

As shown in the previous section the optimal sub-space representation of images proposed in [11] is a powerful tool for computing some form of a distance between two image sequences. In light of that, we propose to use such representation to compare the reference and target sequences obtained by an automatic video-surveillance system in order to detect some video event of interest. In fact, major differences between the two video sequences can be interpreted as an abandoned or missing object.

In that framework the reference sequence is used as a system input to form the optimal sub-space, generating the filters that will later be used to assess the similarity with the target-sequence frames. Although the process of finding the matched filters given by equations (6) and (7) can be computationally expensive, in a surveillance system the reference video may be available long before the target one. Then, the process of finding the optimal sub-space can usually be done as an off-line task. Also, in automatic surveillance systems, one can assume that the reference and target sequences are at least roughly temporally aligned. If this is the case, it is possible to simplify the measure of distance presented in equation (8), since there is no need for searching for the best correspondence in the whole reference sequence.

Another recurrent requirement in the scope of automatic surveillance is the geometric registration among the sequences. This is specially important when moving cameras are employed, since they are susceptible to vibration that is uncorrelated to the trajectory. Such vibration may generate frames where the position of the camera is not the same as it was in the reference video, creating differences in the frame view. In this paper we propose to avoid the need for registration by effecting comparisons of not only the expected output Gaussian function of a given filter, but also with shifted and rotated versions of this function, emulating small variations in the camera position, as detailed in the following section.

4. PROPOSED OBJECT-DETECTION ALGORITHM

In this section the proposed object-detection method is presented in a detailed step-by-step manner, as summarized in Figure 1.

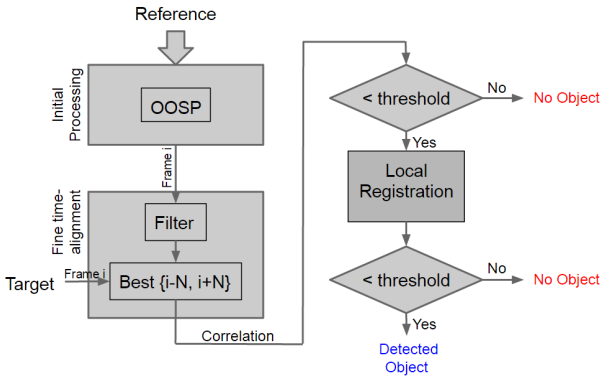


Fig. 1. Block diagram of the object-detection system using operator-space approach.

In a first stage, the reference video is used as input to the filter generator to determine the optimal operator space, as detailed in Section 2. In this process, the algorithm presented in [11] provides a

series of matched-filters that represent each of the frames in the reference video. As stated in the previous section, this entire task may be done off-line, prior to the acquisition of the target video.

To detect the video events of interest, for each incoming target frame, one must perform a similarity search in the operator domain with the reference video. A straightforward way to avoid an exhaustive search is to have the target and reference videos initially synchronized. However, this is not always possible or very time consuming, as the camera movement may differ during the recordings of the two reference and target sequences. In this case a compromise solution is to carry out, frame-by-frame, a fine temporal alignment of the reference and target videos. In order to do so, we search for the best reference-target frame match only within an N-frame vicinity around the one we are interested in. In our system, we employed $N = 10$ frames. We keep the larger similarity as the correct one and use its position as an offset for the starting point of the next-frame search and so on.

The comparison with the reference frame is carried out through the sequence of matched-filters generated in the first step of the algorithm. The target frame is filtered by the corresponding-frame filter and its output is compared with the expected Gaussian output. In this frame comparison, however, instead of making a simple subtraction, as proposed in [11] and expressed in equation (8), a different measure is considered. In this case, we use the maximum of the normalized cross-correlation function between the output generated by the target frame and the predicted Gaussian output. The advantages of such a correlation measure include its efficient computation in the frequency domain and its insensitivity to reasonably small spatial shifts between the target and reference videos, as illustrated in Figure 2.

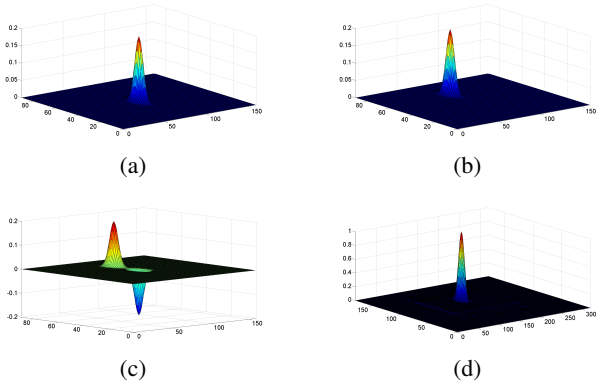


Fig. 2. Example of outputs of filter H and the correlation measure: (a) predicted Gaussian output; (b) obtained Gaussian-like output; (c) difference between outputs; (d) correlation between outputs.

If the similarity measure is below a given threshold value, then that frame possibly contains an abandoned object or an observable video event of interest. Low similarity values, however, may also be caused by frame mismatches due to geometric transformations that cannot be accounted for by a simple cross-correlation between the filter output and the predicted Gaussian function. Examples of such mismatches are different camera positions when acquiring the target and reference frames that can be modeled as a rotation around an arbitrary axis, as depicted in Figure 3.

It is well known that a rotation around an arbitrary axis can be modeled by a rotation followed by a translation [14]. The classical

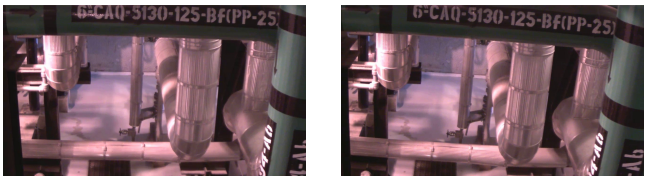


Fig. 3. Example of frame mismatch due to camera rotation: (a) reference frame; (b) corresponding target frame.

way to address this problem would be to perform a registration between the frames by finding keypoints common to both frames and computing the geometric transformations between them [8].

To avoid such a computationally-intensive strategy, one way to deal with the above mentioned problem would be to search through all small rotations and translations of the current frame in the cases that a low correlation value is obtained. In the present case, this search can be made in the matched-filter domain, exploring the shift- and rotation-invariant properties of the filter. In this way, if a filter h processes an image X yielding an image Y, the filter h rotated by θ, when applied to the image X rotated by θ, outputs the image Y rotated by θ but up to some border effects, as may be seen in figure 3. Such effects, however, reduce the normalized cross-correlation even in the case of accurate computation of the rotation, leading to false detections. To cope with this issue, these border effects can be minimized by applying a $K \times K$ Gaussian window to the center of the reference frames before computing the corresponding matched filters. In our implementation, we used $K = 21$ pixels. Next the normalized cross-correlation is computed between the filter's output to translated and rotated versions of a target-frame input and the desired Gaussian function. In this search, we chose a maximum rotation angle of 3.5° with steps of 0.25° and a maximum translation movement of 20 pixels. The highest correlation value for all these rotation and translation values indicates the proper rotation-translation combination for the given frame, and these values are used to initialize the search for the subsequent target frame.

After this process, another Gaussian window is applied to the whole reference image to construct a final filter (Figure 4a). The same window is applied to a shifted and rotated version of the target frame (Figure 4b), according to the best result in the last step, and the final correlation is assessed. If this value is lower than a certain preset threshold, it is considered to be an abandoned (or missing) object or an video event in the frame. Otherwise it is considered to be a similar frame with no observable events, as represented in Figure 1.

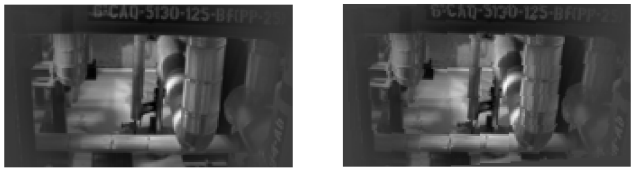
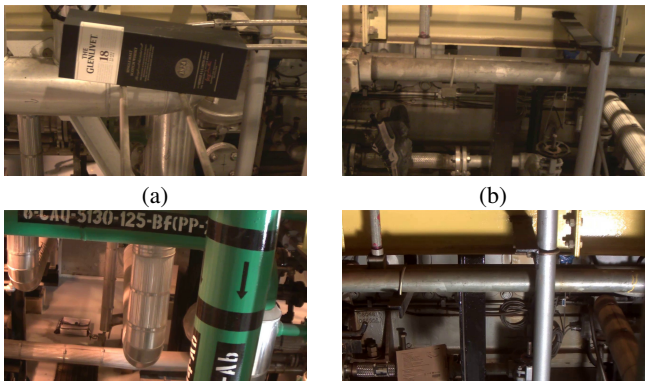


Fig. 4. Example of final frame comparison: (a) windowed reference frame; (b) registered target frame.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed system, 11 videos from the VDAO database [15] (available at [16]) were employed. That database contains several different objects in a cluttered industrial environment recorded by a camera moving horizontally along a linear track. There are objects of different shapes and sizes, as exemplified in Figure 5. There are also videos with varying degrees of illumination, making it possible to test the algorithm in a wide variety of scenarios.



(a) (b) (c) (d)

Fig. 5. Examples of detected objects in a cluttered background from VDAO database [15]: (a) blue box; (b) shoe; (c) camera box; (d) brown box.

The thresholds explained in Section 4 were set depending on the size of the smaller object to be detected. In the end, a value of 0.6 was sufficient to detect larger objects, 0.8 for medium objects, and 0.95 for very small objects. Tests were made for different object sizes and shapes and the different illumination levels covered in the VDAO database. In some of the VDAO videos employed, there are intervals where the reference and target videos poorly match due to camera rotation and translation between the frames, as seen in Figure 3. There is also a great deal of camera shake due to imperfections on the track. These characteristics allow the algorithm to be tested in situations where registration and salient-point detection are required.

Figure 6 depicts a plot of the similarity measure of the video sequence containing objects of different sizes. The regions of low correlation indicate properly the presence of a spurious object, as described above.

The experiments were designed to detect the target frames with a given abandoned object. In that sense, the detection performance was assessed by the number of true-positive detections (frames with object properly detected), number of true-negative detections (frames without objects correctly undetected), number of false-positive detections (frames without objects incorrectly detected) and number of false negative detections (frames with objects improperly undetected). Table 1 show the results of the object detection for 15 videos from the VDAO database, with about 350 seconds (8300 frames) per video in average.

In the performed experiments, it is easy to observe that most of the frames were correctly categorized. In the case of the frames that were considered false positives, the most common cause is a mismatch between the frames due to rotation and vertical translation with larger amplitudes than those predicted in the implementation of the system. Also the main cause of frames being considered false

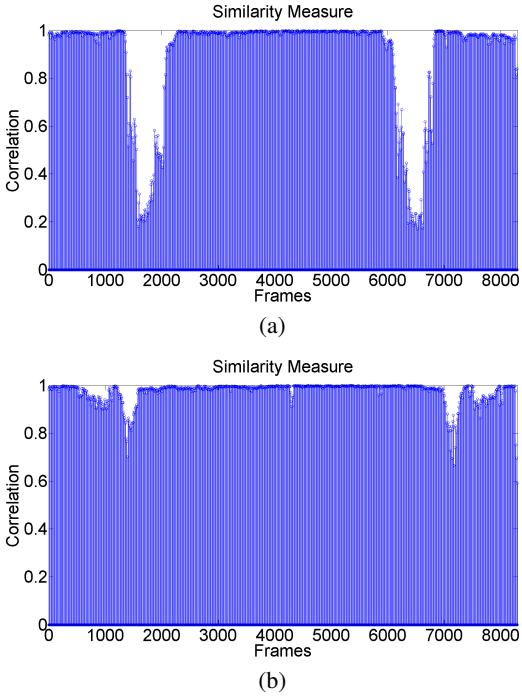


Fig. 6. Framewise similarity measure between two sets of reference and target videos from VDAO database [15]: (a) large object (blue box - Figure 5a); (b) small object (shoe - Figure 5b). Low correlation values indicate the presence of the abandoned object in the target video.

Table 1. Experimental results (frames) for proposed object-detection system.

	Positive	**Negative**
True	23540/26900 (87.51%)	86430/91000 (94.98%)
False	4570/91000(5.02%)	3360/26900 (12.49%)

negatives is when the object appears only partially in the frame, reducing its effective size. This occurs in the cases of partial occlusion of the object and also when the object is entering or leaving the frame. It is important to mention that at a higher level, all abandoned objects were properly detected in all VDAO videos considered.

6. CONCLUSIONS

This paper presented a new method to detect abandoned objects and video events in a cluttered environment without the need of previous registration or fine temporal alignment. The method is based on an optimized sub-space representation of frames that allow the comparison between images to be more robust. The method is able to cope with visually complex environments without the use of feature based registration, that is not a very robust procedure in this kind of environment. The results presented in this work are promising, showing that the method may be viable to implementation in automatic surveillance systems using moving cameras.

ACKNOWLEDGEMENTS

This work was partially funded by FAPERJ, ARO Grant # W911NF-04-D-0003-0022 and CNPq/MCTI Grant # 0890952735408505.

7. REFERENCES

[1] Venkatesh Saligrama, Janusz Konrad, and Pierre-Marc Jodoin, "Video anomaly identification," *IEEE Signal Processing Magazine*, vol. 27, pp. 18–33, September 2010.

[2] Badri Narayan Subudhi, Pradipta Kumar Nanda, and Ashish Ghosh, "A change information based fast algorithm for video object detection and tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 993–1004, July 2011.

[3] YingLi Tian, Rogerio Feris, Haowei Liu, Arun Humpapur, and Ming-Ting Sun, "Robust detection of abandoned and removed objects in complex surveillance videos," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, no. 5, pp. 565–576, September 2011.

[4] Li Cheng, Minglun Gong, Dale Schuurmans, and Terry Caelli, "Real-time discriminative background subtraction," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1401–1414, May 2011.

[5] Fatih Porikli, Yuri Ivanov, and Tetsuji Haga, "Robust abandoned object detection using dual foregrounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–11, January 2008.

[6] Eric Jardim, Xiao Bian, Eduardo A.B. da Silva, Sergio L. Netto, and Hamid Krim, "On the detection of abandoned objects with a moving camera using robust subspace recovery and sparse representation," *Accepted for presentation at IEEE International Conference on Acoustics, Speech, and Signal Processing, Brisbane, Australia*, April 2015.

[7] Joan Serat, Ferran Diego, F. Lumbreras, and J.M. Álvares, "Alignment of videos recorded from moving vehicles," in *14th International Conference on Image Analysis and Processing*, 2007, pp. 512–517.

[8] Hui Kong, Jean-Yves Audibert, and Jean Ponce, "Detecting abandoned objects with a moving camera," *IEEE Transactions on Image Processing*, vol. 2201-2210, pp. 803–806, August 2010.

[9] Ferram Diego, Daniel Ponsa, Joan Serrat, and Antonio M. López, "Video alignment for change detection," *IEEE Transactions on Image Processing*, vol. 20, pp. 1858–1869, July 2011.

[10] Soumyabrata Dey, Vladimir Reilly, Imran Saleemi, and Mubarak Shah, "Detection of independently moving objects in non-planar scenes via multi-frame monocular epipolar constraint," in *Computer Vision*, vol. 7576 of *Lecture Notes in Computer Science*, pp. 860–873. Springer Berlin Heidelberg, 2012.

[11] Xiao Bian and Hamid Krim, "Optimal operator space pursuit: A framework for video sequence data analysis," in *Computer Vision*, vol. 7725 of *Lecture Notes in Computer Science*, pp. 760–769. Springer Berlin Heidelberg, 2013.

[12] Zuowei Shen Jian-Feng Cai, Emmanuel J. Candès, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, March 2010.

[13] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, John Hopkins, Baltimore, 3rd edition, 1996.

[14] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, 2nd edition, 2004.

[15] Allan F. da Silva, Lucas A. Thomaz, Gustavo Carvalho, Mateus T. Nakahata, Eric Jardim, José F. L. de Oliveira, Eduardo A.B. da Silva, Sergio L. Netto, Gustavo Freitas, and Ramon R. Costa, "An annotated video database for abandoned-object detection in a cluttered environment," in *International Telecommunications Symposium*, 2014, pp. 1–5.

[16] "VDAO - Video Database of Abandoned Objects in a Cluttered Industrial Environment," [Online], Available at `http://www.smt.ufrj.br/~tvdigital/database/objects`.