Siamese Networks for Bounding-Box to Silhouette Annotation of Video Databases

Thadeu L. B. Dias[1], Luiz G. C. Tavares[1], Rafael Padilla[1], Allan F. da Silva[1], Lucas A. Thomaz[2], Sergio L. Netto[1], and Eduardo A. B. da Silva[1]

***Abstract*— Pixel-level ground truth masks for object detection databases are extremely useful in the context of machine learning, specially for convolutional neural network applications. However, the manual labeling process of such data demands a lot of effort and time, especially in videos, in which the labeling needs to be performed in each frame. Therefore, only bounding box annotations, that are much faster to perform, are present in most databases. In this work we propose a semi-automated approach to transform bounding-box annotations into silhouette annotations with a reduced processing time. We compute features of a siamese network in the region inside a bounding box and obtain the probability of a pixel to belong to the foreground, which is then refined by a post-processing step. We employ our methodology to the VDAO dataset, creating a new annotation that contains the silhouette of the objects. We estimate that our method results in a reduction of the annotation time by 90% in average, while providing an accurate silhouette for the objects.**

***Keywords*— siamese networks; anomaly detection; video annotation; deep learning.**

I. Introduction

An anomaly may be defined as a non-conforming pattern that varies depending on the scenario and circumstances. Automatic anomaly detection systems can be found in many applications including road inspection, waste sorting, supervision of public areas such as airports, train stations, shopping centers, etc [1]. In the computer vision context, objects added or removed from a video scene can be considered anomalies. However, considering the presence of an object as an anomaly depends on the context and is an important definition for any supervision system. Objects like guns and abandoned luggage, for instance, could be considered anomalies in public crowded places [2], [3], [4]. In industrial sites, such as offshore platforms, objects left in areas that may result in accidents or items capable of producing flames are critical anomalies that could lead to serious consequences if not timely detected [5]. Depending on the application, an immense amount of object classes could be classified as anomalies, preventing popular object detection systems to be used in such cases [6]. In order to overcome this issue, some anomaly detection systems are designed to compare a given input video to another reference video, representing normal environmental conditions [7].

Surveillance systems usually monitor large areas with multiple static or PTZ (Pan-Tilt-Zoom) cameras. Static cameras are limited to a much narrower view of the scene, while PTZ cameras, by zooming and rotating, can cover a wider area. The cost of expanding the monitored area increases if more cameras need to be applied [8], [9], [10]. An alternative for that is the use of moving cameras, which can cover larger areas and may lead to reduced installation cost. The drawback of applying moving cameras to detect anomalous objects or situations is the increasing complexity of the problem. Occlusions, camera jitters, shadows, and light variations limit the success of the same techniques applied in static cameras [11], [12], [13]. Works such as [13], [14], [15] address the anomaly detection problem with a moving camera by comparing a target video, possibly containing the undesired anomaly, to a reference video representing the normal conditions. After aligning both reference and target videos, different similarity measurements are used to detect any stray objects in the scenes.

The anomaly detection problem with a moving camera is well represented in the Video Database of Abandoned Objects (VDAO) [5]. The VDAO is a challenging dataset covering anomaly detection in large industrial environments with many adversities found in surveillance videos that hinder the automatic detection of anomalies. This dataset contains reference and target videos recorded in a cluttered environment with illumination changes. The target videos contain objects distributed in the scenario and, as the camera moves, the objects can be partially or totally occluded in some video scenes. The objects are annotated by their bounding boxes and due to the occlusions and shadows, the manual silhouette annotation is an even more challenging task than it usually is.

In applications where objects are partially occluded or have irregular shapes, however, the bounding box annotations do not represent the objects precisely. In this context, this paper proposes an annotation method to transform bounding box annotations into silhouette annotations when the target and reference videos are available. Our annotation method was applied in the VDAO due to its challenging real-life features but can be easily adapted to other databases. The main idea behind the proposed method is the use of a siamese neural network to extract deep features from both videos and compare the results at the pixel level to produce the object silhouette.

This paper is organized as follows: Section II briefly describes the VDAO database which is used as a case study for the proposed annotation method. Section III proposes a feature extraction method using siamese networks that provides a candidate for the silhouette segmentation, which is further refined using the algorithm detailed in Section IV. Section V describes the obtained silhouette-level VDAO database, and Section VI reports the final conclusions.

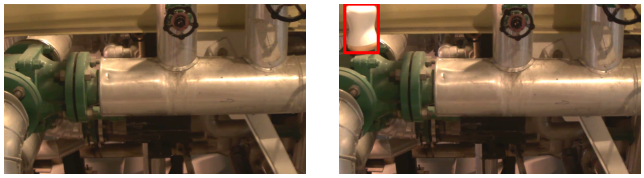
[1]PEE, COPPE, Universidade Federal do Rio de Janeiro, Brazil. [2]Instituto de Telecomunicações, Leiria, Portugal. E-mails: {thadeu.dias, luiz.tavares, rafael.padilla, allan.freitas, lucas.thomaz, sergioln, eduardo}@smt.ufrj.br.

II. The VDAO Database

The so-called VDAO database [5] is a collection of 77 videos recorded by a camera mounted on a robot moving in a back-and-forth trajectory along an industrial plant. The videos have resolution of 1280×720 pixels at 24 frames per second, showing a total of 24 objects placed in the cluttered scenario, simulating objects that do not belong to the environment in expected conditions. Similarly to other databases, reference and target videos are provided. The reference videos are those which do not contain any anomaly, in this case, represented by the abandoned objects, and the target videos contain a single or multiple abandoned objects. In this work, only the single-object videos were considered, and the objects presented in these videos belong to one of the following classes: brown box, camera box, dark-blue box, pink bottle, shoe, towel, white jar, black coat, and black backpack. Many realistic constrains, such as occlusion, light changes, and jitters, hinder the detection of the objects in the target videos, making the VDAO a very challenging dataset for the anomaly detection task.

Besides the 77 VDAO videos, an extra database containing short videos to be used as benchmark is also available. Such videos are 200-frame long and are part of the subset referred to as VDAO-200. This auxiliary testing database, available at [16], contains 59 single-object videos with different objects in different positions under two different illumination conditions. All 59 videos of the VDAO-200 testing database are short-duration patches of the single-object VDAO videos.

The VDAO also provides annotation files containing the position of the abandoned objects for each frame. Current annotations, however, only include the coordinates of bounding boxes encompassing the objects presented in the target videos, as seen in Fig. 1, as manual annotation at the silhouette level was considered unpractical at the time. In fact, even though popular silhouette-annotation tools, such as LabelMe [17] and COCO Annotator [18], could facilitate the process of silhouette annotation, one must still manually sketch the object shape at each video frame, making the task quite time consuming. Due to the industrial cluttered scenario and the variations of illumination present in the VDAO videos, the objects are constantly covered with shadows or are partially occluded, further complicating the annotation process.



(a) reference frame (b) target frame

Fig. 1: Example of a reference frame (a) and its aligned target frame (b). A red bounding box highlights a white jar in the target frame that does not appear in the reference frame.

III. Video Analysis

An efficient comparison of the reference and target videos requires a pre-alignment of these videos in order to remove spatio-temporal noise. Given a pair of reference and target videos, the alignment process searches for a reference frame R_i that best matches a given target frame T_j according to their Euclidean distance. For that, we limit the search to a neighborhood of only 11 reference frames centered at the best-match frame for the previous T_{j-1} target frame. The alignment reduces the occurrence of artifacts outside the object area, improving the efficiency of the silhouette annotation method.

By using the initial layers of a pre-trained deep convolutional neural network (CNN) as feature extractor, we propose to generate rich high-dimensional descriptors for fine regions of the frame space. Comparing pre-aligned reference and target descriptors, we can use a simple metric to detect the pixelwise frame regions that contain anomalies, as represented in Fig. 2.

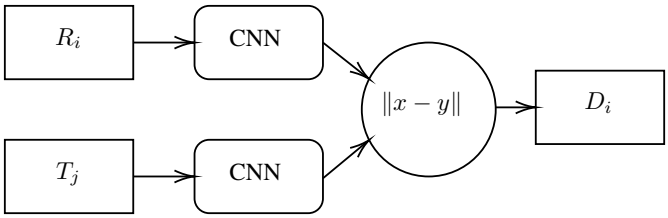


Fig. 2: Truncated twin ResNet-50 CNNs as feature transformers for comparing pre-aligned reference (R_i) and target (T_j) frames in order to detect video differences at the pixel level.

Let $F : \mathbb{R}^{3\times1280\times720} \rightarrow \mathbb{R}^{256\times320\times180}$ be the mapping from RGB space to feature space provided by the layer 'conv2_3' of a ResNet-50 [19] trained on the ILSVRC 2015 dataset, and the transformed reference and target features denoted by $\tilde{R}_i = F(R_i)$ and $\tilde{T}_i = F(T_j)$, respectively. A simple local estimate of the discrepancy between reference and target frames can be obtained through a Euclidean mapping for each descriptor such that

$$[D_i]_{xy} = \sqrt{\sum_c \left([\tilde{R}_i]_{cxy} - [\tilde{T}_j]_{cxy}\right)^2}. \quad (1)$$

An example of a generated difference map for a given frame pair is depicted in Fig. 3, where one may notice how the target-frame stray object (highlighted with a bounding box) yields pixels with a high difference value. Outside the bounding box, where the object is not present, one mainly finds very low difference values and some scattered high difference values that may be removed by additional image-processing techniques. Therefore, we can further refine the difference mapping by incorporating the bounding-box information and ignoring the regions outside the box. Motivated by that observation, we developed a semi-automated technique using morphological operations to obtain the silhouettes of the object from the difference mapping and provide a pixelwise annotation of the objects on a larger dataset, which is fully described in Section IV.

In order to further improve the separation between object and background and make the annotation process more consistent, we concatenate the difference maps inside the boxes for all frames in a video and normalize the difference values using the threshold value τ_D and sample standard deviation σ_D which were obtained from the Otsu method [20]. This results

in a normalized dynamic-range transformation such that

$$[\tilde{D}_i]_{xy} = f\left(\frac{[D_i]_{xy} - \tau_D}{\sigma_D}\right), \quad (2)$$

where the values $\tilde{D}_i$ represent the probability of a pixel to be considered an anomaly and $f(x)$ denotes the logistic function

$$f(x) = \frac{1}{1+e^{-x}}. \quad (3)$$

IV. Proposed Silhouette-Annotation Algorithm

Using the framework defined in Section III, we obtain for each pair of reference and target frames a normalized difference map indicating the probability of a pixel in the bounding box to belong or not to an abandoned object. However, when using a simple binarization procedure to create a mask of foreground and background pixels, one gets a large number of misclassified pixels, as shown in Figs. 4a and 4b.

Analyzing the misclassification cases obtained after the binarization of the difference map, one can see that two main issues arise from this procedure: small regions of false-positive pixels appear outside the object and false-negative regions form small holes inside the object. Such artifacts can be removed by the opening and closing morphological operations [21], respectively, if one chooses the correct structuring element, as illustrated in Figs. 4c and 4d.

In addition, when the object is too similar to the background, as for instance in shadowed regions, the detection algorithm is not able to properly identify all parts of the object as foreground. To cope with this issue, we consider that a pixel may be classified into three different groups: foreground, where the pixel certainly belongs to the abandoned object; background, where the pixel certainly does not belong to the abandoned object; and an undefined zone, where one is not sure to which class the pixel belongs. The undefined zone is formed by applying an erosion [21] on the foreground region followed by a dilation [21] on the background region, thus creating a transition region between foreground and background, as represented by the gray color in Fig. 4e. Based on all these aspects, we then propose the following post-processing steps:

Binarization: Comparison of $\tilde{D}_i$ in Eq. (2) against a threshold value t to form a binary classification mask. In this process, the threshold t is such that false-positive pixels appear far enough from the object and only in isolated cases in order to be removed by subsequent steps (Fig. 4b);

Opening: This operation is performed with a circular structuring element of diameter o to remove the maximum of false-positive regions left by the binarization step (Fig. 4c) without affecting the true-positive regions;

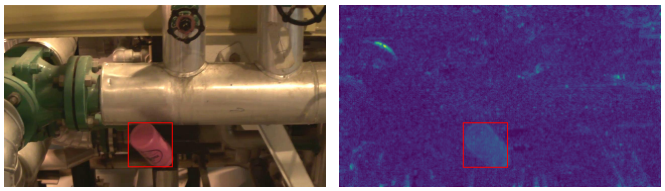
Closing: This operation employs a circular structuring element of diameter c to fill in the false-negative regions within the object (Fig. 4d);

Erosion and Dilation: These combined operations use a circular structuring element of diameter e to create a border around the object which may be considered as a neutral unidentified region (Fig. 4e).

One should note, however, that this procedure requires the proper configuration of the hyperparameters (t, o, c, e) to control the binarization and the morphological operations. These parameters are tuned for sets of contiguous frames, referred to as segments. For the first frame of a segment containing an abandoned object, we perform a manual search for the optimal parameters to annotate the frame, if possible narrowing the search around the parameters used in another segment. For the subsequent frames, one can reduce the annotation time by using information obtained during the annotation of previous annotated frames.

The whole annotation procedure goes on like this: First, we compute a transformation between the bounding boxes of the previous (annotated) frame and the current frame, using it to generate the annotation of the current frame by displacing the background, foreground, and undefined regions of the previous frame. If no false detection of background and foreground is observed, we use the displaced annotation of the previous frame as the annotation for the current frame. If a false detection is observed, we determine the difference map within the bounding box of the current frame and apply the post-processing procedure above with current (t, o, c, e) values. If it is not possible to correct the false detection, we assume that the video excerpt is too long and divide it into two smaller segments, each one requiring a different set of parameter values, which are then adjusted.

In this scheme, the videos are initially segmented so that the segments can be categorized in two types, Type A or Type B, depending whether a given object is entirely or partially visible in its frames, respectively (note that a partially visible object may be entering, leaving, or being partially occluded in a scene). During the annotation process, the segments are recursively divided in two smaller segments until all frames



(a) target frame (b) raw difference mapping

Fig. 3: Example of a target frame (a) with an abandoned object (identified by its red bounding box) and mapped difference frame (b), where the object silhouette clearly stands out.

(a) (b) (c) (d) (e) (f)

Fig. 4: Steps in the silhouette annotation of a shoe through morphology operations. In the proposed scheme, the difference map (a) is binarized (b) and then subjected to successive stages of opening (c), closing (d), erosion and dilation (e). In these figures, the white area represents the obtained silhouette and the gray area represents the undefined zone. The object extracted using the final silhouette can be seen in (f).

containing objects are annotated.

On Type A segments, the object shape does not vary drastically from frame to frame and the morphological operations tend to provide similar results on consecutive frames, which can then be easily annotated. On Type B segments, however, the morphological parameters usually require some adjustments in a frame by frame basis.

V. Silhouette-Annotation Results

We employ the proposed annotation method to obtain the object silhouettes in all 59 single-object VDAO videos and, consequently, in the VDAO-200 as well. The new database, henceforth called VDAO-AS, describes each object silhouette by a sequence of 1280×720 images, with each pixel indicating the presence of an object with a white color, the absence of an object with a black color, and the undefined zone is represented by the gray color.

In order to assist the annotation procedure, we developed a tool that implements the morphological post-processing steps described in Section IV and allows one to control its parameters and visualize the results, as depicted in Fig.5. The graphical user interface for the silhouette annotation was developed in Python using the PyQt library and OpenCV. Among its capabilities, the interface allows the user to create, split, or concatenate segments when the sequence, the difference map, and the bounding-box annotation are provided. It is also possible to set the morphological hyperparameters (t, o, c, e), for each video segment. For all the frames within a segment, the annotation tool initializes the silhouette as a translated version of the silhouette in the first frame of the segment.

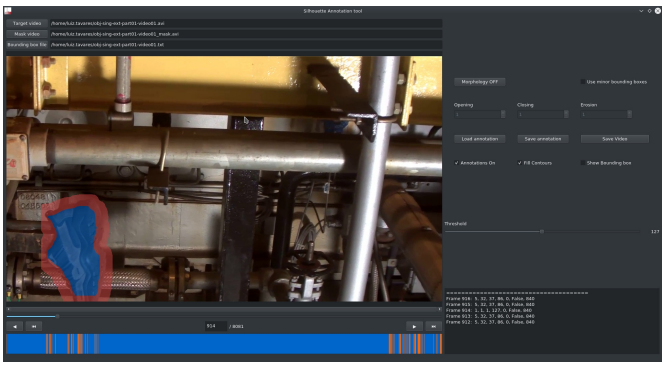
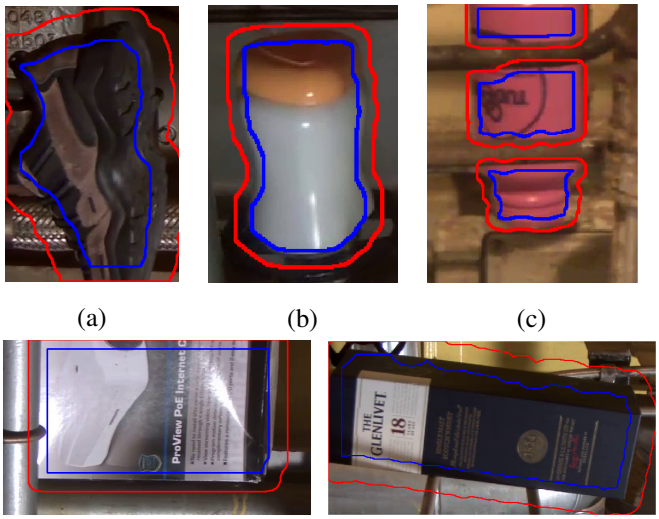


Fig. 5: Silhouette annotation tool used to create the pixel-level VDAO database. For a better visualization, the undefined zone is here shown in red and the annotated foreground in blue.

A. Silhouette VDAO Database

Examples of the VDAO-AS database with the silhouette of the foreground and undefined zones are shown in Fig. 6. Despite the existence of the undefined zone, one can see that the annotated foreground (blue line) is much more adapted to the actual shape of the abandoned object when compared to a bounding box. The undefined zone (red region) is supposed to be discarded when using the database to train supervised methods. This version of the database along with the new ground truth and the annotation tool is available at [22].



(a) (b) (c)

(d) (e)

Fig. 6: Examples of objects with their annotated silhouette (blue line) and undefined zone (red line): (a) shoe; (b) white jar; (c) pink bottle; (d) camera box; (e) dark-blue box.

Some hyperparameter statistics for the entire the annotation process are given in Table I, indicating the overall range and some insight into the sensitivity of each algorithm parameter.

When creating the VDAO-AS database, the amount of time required to annotate each video is related to the number of actions performed to tune the hyperparameters, which depends on the the quality of the difference map. If the difference map is able to differentiate the object from the background (high values inside and low values outside the object boundaries), the annotation process tends to be faster by using the same set of hyperparameters for a long video segment. If this is not the case, the hyperparameters must be adjusted more often, increasing the annotation time, and larger undefined zones may be necessary. In Table II it is possible to see the proportion of annotated frames which required the hyperparameters to be manually adjusted, and the ones in which the user simply verifies if the previous silhouette fits the object in the current frame. We estimate that a manual annotation for all frames should take over 10 h per video, whereas the proposed method reduced this time to an average of 1 h per video.

VI. Conclusion

We consider the use of CNN features to distinguish between foreground (anomaly related) and background pixels in a video-surveillance system. The proposed method can be used to refine a bounding-box annotation process into a silhouette one for any database which includes reference (anomaly free) and target videos, such as the VDAO dataset considered here. As a by-product of this work, we also developed an annotation tool with a graphical user interface to guide the annotation procedure. With such a tool, the annotation of each VDAO video took in average about 1 h, which represents a reduction of the annotation time by 90% when compared to the manual annotation. Overall, the resulting silhoutte-level VDAO (VDAO-AS) contains 130,408 frames with a frame-by-frame silhouette annotation using the proposed method.

TABELA I: Statistics on the annotated hyperparameter values for each object type: diameter of the circular structuring element in pixels for the morphological operations and binarization threshold value.

Object	Opening			Closing			Erosion/Dilation			Threshold		
	Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean
shoe	70	1	11.29	157	1	49.54	99	1	29.83	208	32	93.76
dark-blue box	69	1	9.32	190	1	44.14	173	1	27.35	173	5	90.37
camera box	69	1	15.31	197	1	55.35	113	1	27.47	195	6	82.04
white jar	36	1	7.41	199	1	26.46	61	1	18.57	201	8	81.22
brown box	99	1	8.10	182	1	37.17	99	1	29.55	251	6	92.10
pink bottle	50	1	12.63	199	1	20.92	79	1	24.47	173	35	103.85
towel	99	1	12.56	194	1	35.59	99	1	27.83	209	5	82.92
black coat	84	1	14.25	199	1	45.86	136	1	54.33	162	10	73.51
black backpack	151	1	14.87	199	1	26.46	199	1	51.30	138	7	94.66

TABELA II: Number of frames annotated with manual tuning of hyperparameters (labelled as "manual") and using previous frame silhouette (labelled as previous sillouette).

Object	Manual	Previous sillouette	Total
shoe	1019 (8.86%)	10479 (91.14%)	11498
darkBlueBox	441 (3.81%)	11144 (96.19%)	11585
cameraBox	660 (5.65%)	11022 (94.35%)	11682
whiteJar	484 (4.58%)	10086 (95.42%)	10570
brownBox	658 (4.61%)	13628 (95.39%)	14286
pinkBottle	482 (4.63%)	9923 (95.37%)	10405
toalha	1132 (6.34%)	16724 (93.66%)	17856
blackCoat	1078 (5.27%)	19363 (94.73%)	20441
blackBackpack	842 (3.81%)	21243 (96.19%)	22085

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES) - Finance Code 001, CNPq, and FAPERJ. This work was also partially funded by Programa Operacional Regional do Centro, project PlenoISLA POCI-01-0145-FEDER-028325 and by FCT/MCTES through national funds and when applicable co-funded by EU funds under the project UIDB/EEA/50008/2020. We also thank the support of NVIDIA Corporation. for donating the Titan Xp GPU used in this research.

REFERÊNCIAS

[1] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[2] D. M. Sheen, D. L. McMakin, and T. E. Hall, "Three-dimensional millimeter-wave imaging for concealed weapon detection," *IEEE Transactions on Microwave Theory and Techniques*, vol. 49, no. 9, pp. 1581–1592, 2001.

[3] T. D. Räty, "Survey on contemporary remote surveillance systems for public safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 5, pp. 493–515, 2010.

[4] Y. Tian, R. S. Feris, H. Liu, A. Hampapur, and M.-T. Sun, "Robust detection of abandoned and removed objects in complex surveillance videos," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 5, pp. 565–576, 2010.

[5] A. 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," in *International Telecommunications Symposium*, São Paulo, Brazil, Aug. 2014, pp. 1–5.

[6] A. F. Otoom, H. Gunes, and M. Piccardi, "Feature extraction techniques for abandoned object classification in video surveillance," in *IEEE International Conference on Image Processing*, San Diego, USA, Oct. 2008, pp. 1368–1371.

[7] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, June 2010, pp. 1975–1981.

[8] A. Dore, M. Soto, and C. S. Regazzoni, "Bayesian tracking for video analytics," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 46–55, 2010.

[9] B. N. Subudhi, P. K. Nanda, and A. 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, 2011.

[10] V. Saligrama, J. Konrad, and P.-M. Jodoin, "Video anomaly identification," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 18–33, 2010.

[11] L. A. Thomaz, A. F. da Silva, E. A. da Silva, S. L. Netto, and H. Krim, "Detection of abandoned objects using robust subspace recovery with intrinsic video alignment," in *IEEE International Symposium on Circuits and Systems*, Baltimore, USA, May 2017, pp. 1–4.

[12] W.-C. Hu, C.-H. Chen, T.-Y. Chen, D.-Y. Huang, and Z.-C. Wu, "Moving object detection and tracking from video captured by moving camera," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 164–180, 2015.

[13] E. Jardim, L. Thomaz, E. A. B. da Silva, and S. L. Netto, "Domain-transformable sparse representation for anomaly detection in moving-camera videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 1329–1343, 2020.

[14] 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, 2010.

[15] 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," in *IEEE International Conference on Image Processing*, Phoenix, USA, Sept. 2016, pp. 3967–3971.

[16] "VDAO-200: 200-frame excerpts from VDAO database," (Accessed: March 12, 2020). [Online]. Available: http://www.smt.ufrj.br/~tvdigital/database/research

[17] J. Yuen, B. Russell, C. Liu, and A. Torralba, "Labelme video: Building a video database with human annotations," in *IEEE International Conference on Computer Vision*, Kyoto, Japan, Oct. 2009, pp. 1451–1458.

[18] J. Brooks, "COCO Annotator," 2019, (Accessed: March 12, 2020). [Online]. Available: https://github.com/jsbroks/coco-annotator/

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv e-prints*, p. arXiv:1512.03385, Dec. 2015.

[20] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[21] J. Serra, *Image Analysis and Mathematical Morphology*. Orlando, USA: Academic Press, 1983.

[22] "VDAO-AS: Video database for abandoned-object detection with annotated silhouettes." [Online]. Available: http://www.smt.ufrj.br/~tvdigital/database/objects/page_02.html