A CONSISTENT TWO-LEVEL METRIC FOR EVALUATION OF AUTOMATED ABANDONED OBJECT DETECTION METHODS

Patrick Krusch, Erik Bochinski, Volker Eiselein, Thomas Sikora

Technische Universität Berlin
Communication Systems Group
{*krusch,bochinski,eiselein,sikora*}@nue.tu-berlin.de

ABSTRACT

Scientific interest in automated abandoned object detection algorithms using visual information is high and many related systems have been published in recent years. However, most evaluation techniques rely only on statistical evaluation on the object level. Therefore and due to benchmarks with commonly only few abandoned objects and a non-standardized evaluation procedure, an objective performance comparison between different methods is generally hard. We propose a new evaluation metric which is focused on an end-user application case and an evaluation protocol which eliminates uncertainties in previous performance assessments. Using two variants of an abandoned object detection method, we show the features of the novel metric on multiple datasets proving its advantages over previously used measures.

Index Terms— metric, evaluation, abandoned object detection, video surveillance, security

1. INTRODUCTION AND RELATED WORK

Automated detection of abandoned objects is a high-interest topic among visual analytics experts in order to detect potential threats in public places. With an increased need of security applications, related research is driven by both the rising numbers of CCTV cameras installed and a paradigm shift from passive surveillance to active analysis with operator support by semi-automatic, intelligent systems. Typical problems for automated detection of abandoned objects are, e.g. the variety of scenes (e.g. changing outdoor weather and lighting conditions), reflections, occlusions due to, e.g. crowdiness of public places, camera motion and so on.

Many methods addressing the task of automated detection of abandoned objects have been published in the literature. Most of them are based on the distinction between foreground and background in an image, such as, e.g. the popular method proposed in [1] where Gaussian mixture models (GMM) are used to describe the scene background on a per-pixel basis. Different modes allow for multiple background representations, e.g. in case of changing backgrounds. While this simple method is fast and reliable in many cases, further extensions are required for the classification of static objects, e.g. dual background models which learn the background at multiple rates and thus allow both feature extraction in more detail and to extract more specific knowledge about new or removed objects [2, 3]. Further extensions in order to classify false alarms by passers-by include the usage of human detectors [4, 5] or mechanisms of detecting slight pixel-motion by non-static objects [6]. In [7], a pan-tilt-zoom camera is used to zoom on static objects and classify false alarms using a convolutional neural network. [8] uses a dual-background model with temporal consistency constraints and back-tracing in order to find the owner of an abandoned object.

For the evaluation of the metric proposed in this work, we use a finite-state machine proposed in [9] with a splitting-mode extension in order to adaptively parametrize the GMMs used for background subtraction [10]. In order to emphasize the features of the proposed metric, two variants of this system are used which include the integration of a person detection method based on a deformable parts model [11] as well as filtering objects based on their expected size.

1.1. State-of-the-Art for Evaluation of Abandoned Object Detection Algorithms

Despite increasing interest by researchers in recent years, the evaluation protocol of abandoned object detection systems is not standardized. Commonly used benchmarks involve, e.g. the realistic i-LIDS AVSS 2007 [12] and PETS 2006 [13] datasets recorded in public train stations. The recent ABODA dataset [8] shows lab environments but poses very specific challenges such as lighting changes or night scenarios.

A major issue for a common evaluation on these datasets are different task descriptions. In PETS 2006 and AVSS, spatial and temporal rules are to be obeyed in order to detect when an object owner leaves the object or the scene for more than 60 seconds. Only in this case, an alarm shall be raised. For evaluation, AVSS does include ground truth (GT) time stamps but no indication where the alarm should be raised, i.e. no bounding boxes of abandoned objects. For PETS 2006, only one floor point per object and the frame number of alarms is available as GT.

For ABODA, no explicit evaluation protocol is defined, thus leaving the details of a performance assessment mostly to the user of the dataset. As for PETS, neither the position of objects to be identified nor the related time stamps are defined as a publicly available ground truth. In some cases, this leads to uncertainties about the objects to be detected, e.g. in video 11, there are two objects left behind but apparently, only one has been used in the evaluation of [8]. It is uncertain which one shall be identified. Additionally, the videos are generally too short to apply the 60 seconds rule from the other datasets which is a problem for a common evaluation protocol.

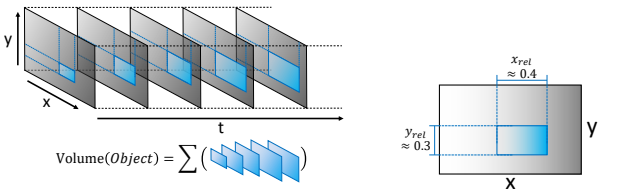
Another, general critique of all the datasets mentioned is that they contain only very few abandoned objects. Therefore, an often-performed statistical event-level analysis on a per-video basis is very coarse and leads to similar values for different systems (as will be shown in the evaluation). In this work, we aim at removing these uncertainties by defining a consistent and use-case oriented evaluation protocol for abandoned object detection.

The coarse evaluation results can be accounted for to some extent by using background subtraction datasets, e.g. [14] which, however, have a focus on separation of background and foreground and do not consider problems of abandonedness of objects. For these datasets, statistical evaluation in terms of correctly / wrongly detected pixels is very common and useful (e.g. [15]). This inspired us to propose a novel evaluation procedure for abandoned object detection.

The rest of the paper is organized as follows: Section 2 describes the guidelines for a novel, consistent evaluation protocol and use-case oriented metrics leading to the performance measures proposed in this paper. This is the main contribution of our work. Section 3 describes two abandoned object detection systems which are evaluated in Section 4 in order to show the usefulness of the proposed measures. Chapter 5 concludes the paper.

2. PROPOSED METRICS AND EVALUATION PROTOCOL

An important part of the metrics proposed in this paper is based on a pixel-level evaluation which can be described as the spatio-temporal accuracy for the alarms generated by the system. An abandoned object should normally be found as soon as possible after it has been left in the scene (*frame-wise temporal accuracy*) until it may be removed from the scene (as e.g. in AVSS). Also, the object should be segmented completely (*pixel-wise spatial accuracy*). Depending on the application case, an end user might prioritize one of these requirements over the other, e.g. when a potential threat shall be identified as soon as possible, spatial resolution is less important than temporal resolution. Such requirements, however, are not regarded by common evaluation protocols where it is generally unspecified if the object shall just be found to some extent at some time or with a certain accuracy.



(a) marked object in a video sequence (b) marked object in one frame

Fig. 1: Proposed object volume: The relative size of an object (width x_{rel} / height y_{rel}) is computed in all frames. The sum over all marked frames gives the final volume of the object.

Apart from the underlying pixel and frame-wise event segmentation, the system's performance as a whole is described in statistical terms of correctly identified events. Related metrics are given in the object-level part of the proposed metrics.

2.1. Spatio-Temporal Accuracy on Pixel Level

In order to assess the spatial and temporal precision of the detections received, we propose treating objects as normalized spatio-temporal volumes as shown in Figure 1. While in this work we use rectangular regions of interest, other shapes are thinkable for datasets with exact pixel-wise annotation.

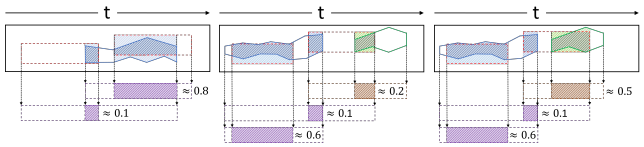
We define true positive (TP), false positive (FP) and false negative (FN) volumes which are more expressive than binary hits / misses. This results in the following volumes per video:

$$\begin{aligned} V_{TP} &= (\bigcup V_{det}) \cap (\bigcup V_{gt}), \\ V_{FP} &= (\bigcup V_{det}) \setminus (\bigcup V_{gt}), \\ V_{FN} &= (\bigcup V_{gt}) \setminus (\bigcup V_{det}) \end{aligned} \tag{1}$$

with $\bigcup V_{det}, \bigcup V_{gt}$ as the union of all detection / ground truth volumes. Statistical measures like **precision**, **recall** or **F-measure** can now be computed in the traditional manner as well as V_{log}, a measure for the **normalized False-Positive volume** over the whole video (higher values are better):

$$V_{log} = -log(\frac{V_{FP}}{V_{video}}) = -log(\frac{V_{FP}}{\#frames}) \tag{2}$$

Using these statistical measures gives a precise insight in the quality of the retrieved detections because the quality of the detections is expressed on a continuous scale. Two problems are solved: Firstly, the metric is finer grained, which is important for analyzing methods on the typically small datasets consisting of only a few sequences with one or two static objects. Secondly, methods with a high spatio-temporal accuracy can be identified and preferred over those which find a correct detection only after a longer training time, fail to detect it completely or generate many false pixel classifications.



(a) $\sigma_t = 0.1 \rightarrow$ 2/0/0 $\sigma_t = 0.5 \rightarrow$ 1/0/0 (b) $\sigma_t = 0.1 \rightarrow$ 2/0/1 $\sigma_t = 0.5 \rightarrow$ 1/1/0 (c) $\sigma_t = 0.1 \rightarrow$ 2/0/1 $\sigma_t = 0.5 \rightarrow$ 2/0/0

Fig. 2: Temporal intersection over union (IOU): Results are given as TP / FP / OS. Red dashed boxes show ground truth (GT) objects, blue and green regions represent detections. Marked regions indicate a fulfilled spatial IOU criterion.

2.2. Statistical Evaluation on Object Level

Especially for end-users, an analysis of the detections on a per-object basis in order to evaluate and fine-tune a system with respect to the generated alarms is needed. This also allows an effective transfer of the new volume-based evaluations to traditional metrics from the literature.

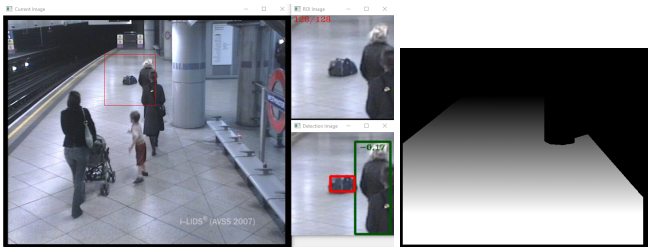
Objects are considered true positive if the spatial and temporal intersection-over-union (IOU) between a detection marked in $[t_{det,beg}; t_{det,end}]$ and an associated ground truth volume from $[t_{GT,beg}; t_{GT,end}]$ exceeds the respective thresholds σ_s, σ_t. Using $t_{beg} = \max(t_{det,beg}, t_{GT,beg})$ and $t_{end} = \min(t_{det,end}, t_{GT,end})$, this can be defined as

$$\frac{\sum_{i=t_{beg}}^{t_{end}} \mathcal{H}_0(\frac{A_{det,i} \cap A_{GT,i}}{A_{det,i} \cup A_{GT,i}} - \sigma_s)}{\max(t_{det,end}, t_{GT,end}) - \min(t_{det,beg}, t_{GT,beg}) + 1} > \sigma_t \quad (3)$$

with $A_{det,i}, A_{GT,i}$ as the spatial regions of interest in the i^{th} frame and the Heaviside function $\mathcal{H}_0$ with $\mathcal{H}_0(0) = 0$. Equation (3) thus allows parametrizing both spatial and temporal overlap required for a detection match to the ground truth and enables adjustments according to the user needs, e.g. prioritizing temporal accuracy over spatial precision.

As an additional measure, we introduce **oversegmentation** (OS) which describes true positives to an already associated GT object. The case of an object not being recognized as a whole but rather as multiple smaller detections has not been described in previous abandoned object detection metrics. One could consider such detections both false positives or matching but we believe that for the special nature of the task of abandoned object detection, they should be reported separately from completely false alarms.

Figure 2 explains visually the TP/FP/OS metric in the spatial and temporal domain where σ_t influences the results based on the marked portion of the temporal IOU. A low σ_t allows even small temporal IOUs to be counted as true positives or oversegmentation, while this can be restrained by a medium σ_t.



(a) (b)

Fig. 3: (a) Person detector running on sequence AVSS_AB_Medium. (b) Mask used for determining the size filter constraint.

2.3. Proposed Evaluation Protocol

In order to have a consistent evaluation protocol for all datasets, we consider an object abandoned when the owner starts walking away from it (first step completed). This is especially useful for short videos such as in the ABODA dataset.

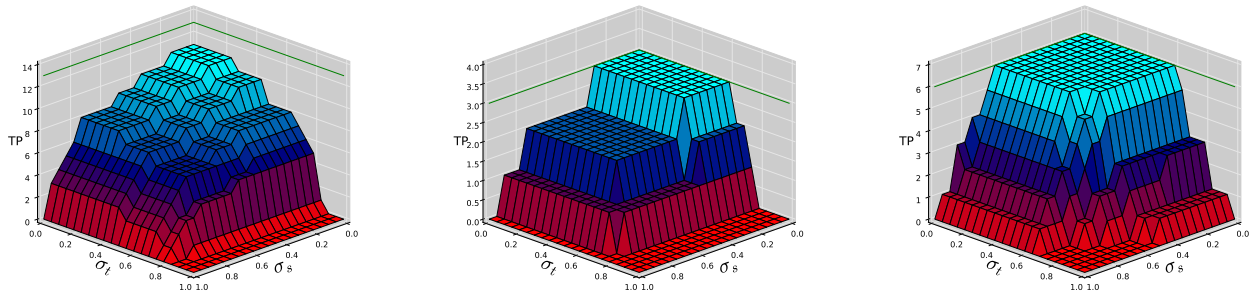
Due to the variable σ_s, σ_t in the proposed method, no assumptions on spatial or temporal overlaps required for a correct match are needed. These two points in our work ensure a significantly better comparability between different algorithms than previous performance protocols.

3. STATIC OBJECT DETECTION SYSTEM USED IN EVALUATION

For evaluation of the new metrics we use a system which combines a finite-state machine [9] with a splitting-mode extension to adaptively parametrize the GMMs used for background subtraction [10] (baseline). In order to reduce false positive detections (e.g. due to lighting conditions or bystanders), a two-step filtering approach (TSF) is used as described below. Baseline and TSF will be compared using the proposed performance metrics.

Person Filtering accounts for detections caused by humans remaining almost static for a longer period of time. We use OpenCV's implementation of the well-known discriminatively trained deformable parts models (DPM)[11] with the pre-trained model from VOC2007 to find humans around the detections in an image (see Figure 3 (a)). The resulting pedestrian detections are post-processed by non-maxima suppression and min-score thresholding. If a pedestrian detection overlaps at least to 50% with an object bounding box, this detection is considered a false alarm. It is removed if at least half of the frames were marked as a false alarm.

Size Filtering is additionally used to filter unexpectedly small detections. The size of objects of interest such as, e.g. backpacks, can be estimated in relation to a approximate depth map of the scene (see Figure 3 (b) for an exemplary map for AVSS) and improbable candidates are removed.



(a) ABODA (b) AVSS_2007 (c) PETS_2006

Fig. 4: True positive results for different σ_s and σ_t (two step filter). True positives in ground truth indicated by green line.

Dataset	AVSS_2007							PETS_2006			ABODA		
Method	[3]	[6]	[8]	[16]	[17]	BL	TSF	[8]	BL	TSF	[8]	BL	TSF
Prc	0.6	0.6	1.0	1.0	1.0	0.43	1.0	1.0	0.67	1.0	0.67	0.42	0.91
Rec	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.92	0.85	0.85
F1	0.75	0.75	1.0	1.0	1.0	0.6	1.0	1.0	0.8	1.0	0.77	0.56	0.88

Table 1: Traditional evaluation using object-based precision, recall and F-measure. BL: baseline, TSF: two-step-filter.

4. EVALUATION USING THE NEW METRICS

Table 1 shows an evaluation using the traditional event-based precision / recall measures from the literature on PETS 2006 (view 3), AVSS 2007 (AB videos) and ABODA. For a correctly matched detection, we assumed a minimal overlap of 1 frame and 1 pixel ($\sigma_s = \sigma_t = 0$) but the overlap needed is not specified in the related publications. It can, however, be seen that the results for AVSS and PETS using traditional metrics do not allow a clear ranking between different methods.

Table 2 shows the results using the proposed volume-based metrics. Unfortunately, without the source code, these cannot be computed for previously published methods. The additional filtering in TSF improves the precision and V_{log} values and therefore also the F-measure.

On the object level (Table 3), TSF improves the results over the baseline method because false positives are removed in all datasets. The performance is slightly better than [8] with one more true positive missed but less false positives.

Figure 4 shows combined results for different values of these parameters and the two-step filter approach. Results for PETS vary little with both σ values increasing. True positives for AVSS are segmented with less spatial but higher temporal accuracy (no detections only for $\sigma_t \geq 0.8$). Results for ABODA vary more but are good even for medium σ_s, σ_t.

The source code and ground truth needed to compute the proposed performance measures have been published [1] in order to encourage further developments and usage in the community.

[1] http://www.nue.tu-berlin.de/c2lm_script/

Dataset	Precision	Recall	F-measure	V_{log}
Baseline (AVSS)	0.61	0.72	0.66	6.16
TSF (AVSS)	0.65	0.72	0.69	6.36
Baseline (PETS)	0.51	0.76	0.61	6.68
TSF (PETS)	0.67	0.76	0.71	7.35
Baseline (ABODA)	0.83	0.63	0.72	7.94
TSF (ABODA)	0.95	0.63	0.76	9.35

Table 2: Results for all datasets (each over all videos) with proposed volume-based precision / recall / F-measure and V_{log} for baseline method and two-step-filter (TSF).

	AVSS_2007					PETS_2006						
Video	Easy	Med.	Hard	$\sum$		S1-T1	S2-T3	S4-T5	S5-T1	S6-T3	S7-T6	$\sum$
GT	1	1	1	3		1	1	1	1	1	1	6
Baseline	1/0	1/2	1/2	3/4		1/0	1/1	1/2/1	1/0	1/0	1/0	6/3/1
TSF	**1/0**	**1/0**	**1/0**	**3/0**		**1/0**	**1/0**	**1/0**	**1/0**	**1/0**	**1/0**	**6/0**
[8]	**1/0**	**1/0**	**1/0**	**3/0**		**1/0**	**1/0**	**1/0**	**1/0**	**1/0**	**1/0**	**6/0**

	ABODA											
Video	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	$\sum$
GT	1	1	1	1	1	2	1	1	1	1	2	13
Baseline	1/0	1/0	1/0	1/0	1/2	1/0	1/5	1/0	1/0	1/1	1/7	11/15
TSF	**1/0**	**1/0**	**1/0**	**1/0**	**1/0**	1/0	**1/1**	**1/0**	**1/0**	**1/0**	**1/0**	**11/1**
[8]	**1/0**	**1/0**	**1/0**	**1/0**	1/1	**2/0**	**1/1**	1/1	**1/0**	**1/0**	1/3	12/6

Table 3: Proposed object-based results using $\sigma_s = \sigma_t = 0$ [TP / FP / (OS)]. GT: Ground truth, TSF: two-step-filter.

5. CONCLUSION

In this paper, a new evaluation process and consistent metrics for abandoned object detection systems have been proposed in order to ensure a transparent, objective and reproducible performance assessment. By addressing the evaluation on both the pixel- and object-based level, user priorities in the evaluation can be addressed and especially through the volume-based metrics, a clear ranking is enabled even for datasets with few videos and a small number of objects.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Communitys FP7 and BMBF-VIP+ under grant agreement number 607480 (LASIE) and 03VP01940 (SiGroViD).

7. REFERENCES

[1] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 246–252, June 1999.

[2] A. Singh, S. Sawan, M. Hanmandlu, V. K. Madasu, and B. C. Lovell, "An abandoned object detection system based on dual background segmentation," *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 352–357, September 2009.

[3] F. Porikli, Y. Ivanov, and T. Haga, "Robust abandoned object detection using dual foregrounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, August 2008.

[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*, vol. 41, no. 5, pp. 565–576, 2011.

[5] L. G. Sole, A. S. Sonawane, S. R. Shinde, and V. M. Mane, "Video analytics for abandoned object detection and its evaluation on atom and arm processor," *IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–6, December 2013.

[6] A. K. Singh and A. Agrawal, "An interactive framework for abandoned and removed object detection in video," *Annual IEEE India Conference (INDICON)*, pp. 1–6, December 2013.

[7] T. Ogawa, D. Hiraoka, S. i. Ito, M. Ito, and M. Fukumi, "Improvement in detection of abandoned object by pan-tilt camera," *IEEE International Conference on Knowledge and Smart Technology*, pp. 152–157, February 2016.

[8] K. Lin, S.-C. Chen, C.-S. Chen, D.-T. Lin, and Y.-P. Hung, "Abandoned object detection video temporal consistency modeling and back-tracing verification for visual surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1359–1370, 2015.

[9] R. Heras Evangelio and T. Sikora, "Static object detection based on a dual background model and a finite-state machine," *EURASIP Journal on Image and Video Processing*, vol. 2011, no. 3, December 2011.

[10] R. Heras Evangelio, M. Pätzold, I. Keller, and T. Sikora, "Adaptively splitted gmm with feedback improvement for the task of background subtraction," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 5, pp. 863–874, 2014.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[12] "Avss2007 dataset," `http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html`.

[13] "Pets2006 dataset," `http://www.cvg.reading.ac.uk/PETS2006/data.html`.

[14] "Change detection dataset," `http://changedetection.net/`.

[15] Mustafa Karaman, Lutz Goldmann, Da Yu, and Thomas Sikora, "Comparison of static background segmentation methods," in *Visual Communications and Image Processing 2005*. International Society for Optics and Photonics, 2005, pp. 596069–596069.

[16] S. Bhinge, Y. Levin-Schwartz, G.-S. Fu, B. Pesquet-Popescu, and T. Adalı, "A data-driven solution for abandoned object detection: Advantages of multiple types of diversity," *IEEE Global Conference on Signal and Information Processing*, pp. 1347–1351, December 2015.

[17] Wahyono, A. Filonenko, and K.-H. Jo, "Detecting abandoned objects in crowded scenes of surveillance videos using adaptive dual background model," *8th International Conference on Human System Interactions (HSI)*, pp. 224–227, June 2015.