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. crowdingness 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.


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 abandonment 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.

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:

\[
\begin{align*}
V_{TP} & = (\bigcup V_{det}) \cap (\bigcup V_{gt}), \\
V_{FP} & = (\bigcup V_{det}) \setminus (\bigcup V_{gt}), \\
V_{FN} & = (\bigcup V_{gt}) \setminus (\bigcup V_{det})
\end{align*}
\]

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):

\[
V_{log} = -\log\left(\frac{V_{FP}}{V_{video}}\right) = -\log\left(\frac{V_{FP}}{\#frames}\right) \quad (2)
\]

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.
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 evaluation 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 $[t_{\text{det,beg}}, t_{\text{det,end}}]$ and an associated ground-truth volume from $[t_{\text{GT,beg}}, t_{\text{GT,end}}]$ exceeds the respective thresholds $\sigma_s, \sigma_t$. Using $t_{\text{beg}} = \max(t_{\text{det,beg}}, t_{\text{GT,beg}})$ and $t_{\text{end}} = \min(t_{\text{det,end}}, t_{\text{GT,end}})$, this can be defined as

$$\sum_{i=1}^{t_{\text{end}}} H_0\left(\frac{A_{\text{det,beg} \cup A_{\text{GT,beg}}}}{A_{\text{det,beg} \cap A_{\text{GT,beg}}}} - \sigma_s\right) \geq \max(t_{\text{det,end}}, t_{\text{GT,end}}) - \min(t_{\text{det,beg}}, t_{\text{GT,beg}}) + 1 > \sigma_t \tag{3}$$

with $A_{\text{det,beg}, \text{GT,beg}}$ as the spatial regions of interest in the $i^{th}$ frame and the Heaviside function $H_0$ with $H_0(0) = 0$. 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 $\sigma_t$ influences the results based on the marked portion of the temporal IOU. A low $\sigma_t$ allows even small temporal IOUs to be counted as true positives or oversegmentation, while this can be restrained by a medium $\sigma_t$.

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 $\sigma_s, \sigma_t$ 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 model (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.
Table 1: Traditional evaluation using object-based precision, recall and F-measure. BL: baseline, TSF: two-step-filter.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AVSS_2007</th>
<th>PETS_2006</th>
<th>ABODA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>TSF</td>
<td>BL</td>
</tr>
<tr>
<td>Rec</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F1</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
</tbody>
</table>

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).

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

Fig. 4: True positive results for different $\sigma_s$ and $\sigma_t$ (two step filter). True positives in ground truth indicated by green line.
7. REFERENCES


