Performance Evaluation in Video-Surveillance Systems: The EventVideo Project Evaluation Protocols

Juan C. SanMiguel, Álvaro García-Martín, and José M. Martínez

Abstract During recent years, automatic video-surveillance systems have experienced a great development driven by the growing need for security. Many approaches exist whose performance is not clear for a large variety of available scenarios. To precisely identify which ones operate better for each scenario, empirical performance evaluation has been widely used for determining their strengths and weaknesses through their results. This approach requires defining two aspects (usually named as the evaluation protocol): the dataset (representative sequences) and the metrics (performance estimators). Common empirical approaches use metrics based on ground-truth data that define an ideal result, but there are also some novel approaches that do not require such data. Furthermore, the existence of several metrics and the growing availability of video data increase the complexity of the protocol design as well as require us to automate the whole evaluation process. In this chapter, considering the main analysis stages of a typical video-surveillance system (video object segmentation, people detection, video object tracking and event recognition), we introduce their evaluation protocols within the scope of the EventVideo project.

1 Introduction

During recent years, automatic video-surveillance systems have experienced a great development driven by the need for security in private and public places. Many approaches are available whose effectiveness is not clear [10]. They deal with a huge variety of environments that might change over time (e.g., lighting conditions) or present a substantial difference (e.g., sunny or rainy day). Hence, the performance

J.C. SanMiguel (✉) · Á. García-Martín · J.M. Martínez
Video Processing and Understanding Lab, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain
e-mail: Juancarlos.SanMiguel@uam.es

Á. García-Martín
e-mail: Alvaro.Garcia@uam.es

J.M. Martínez
e-mail: JoseM.Martinez@uam.es

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_9, © Springer-Verlag Berlin Heidelberg 2013

of such systems can degrade significantly in these scenarios [17]. As these systems are composed of several analysis stages [35], a performance analysis for each one is required before examining the entire system. To precisely identify which approaches operate better in certain scenarios, performance evaluation has been proposed in the literature as a way to determine their strengths and weaknesses. The widely used empirical approach is based on evaluation through the analysis of the obtained results. For such analysis, two components have to be specified: the dataset (a set of sequences covering the situations that the algorithm might face being large enough to represent real world conditions) and the metrics (which allow us to quantify the performance of algorithms or systems). These two aspects are also known as the evaluation protocol [4, 22]. Traditional performance evaluation approaches use metrics based on ground-truth data that represents a manual annotation of the ideal result. The generation of ground-truth is usually a time consuming task and, therefore, limits the dataset size. Although there are other approaches not focused on ground-truth data [30, 38], most of the current literature assumes the availability of such data. Furthermore, the existence of several metrics increases the complexity of designing an evaluation protocol. Another point to be taken into account is the increasing quantity of video data available, which generates a new need to automate and optimize the whole evaluation process. In this chapter, we present the evaluation protocols (dataset and metrics) for the main analysis stages that compose a typical video-surveillance system (video object segmentation, people detection, video object tracking and event recognition) within the scope of the EventVideo project.[1]

The remainder of this chapter is organized as follows. First, the selected stages and evaluation scenarios of the EventVideo project are described in Sect. 2. Then, the related work on performance evaluation is discussed in Sect. 3. After that, Sect. 4 presents the evaluation protocols of the EventVideo project. Finally, Sect. 5 summarizes the chapter with some conclusions and future work.

2 Evaluation Scenarios

The EventVideo project considers the most common analysis stages of video-surveillance systems and evaluates them under different scenarios. In this section, we describe these stages and the classification criteria for the scenarios.

2.1 *Selected Analysis Stages*

The stages that compose a typical video-surveillance system are (see Fig. 1) [35]:

- *Video object segmentation*: extracts the foreground objects by applying analysis steps to the video sequence such as foreground analysis [5] and shadow removal [27]. Its output is a binary mask indicating the foreground objects.

[1] http://www-vpu.eps.uam.es/eventvideo/

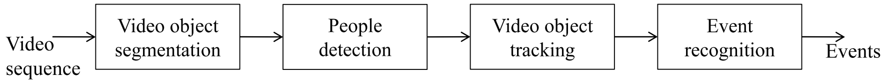


Fig. 1 Typical processing chain for a video-surveillance system

Table 1 Proposed classification for the evaluation scenarios

Scenario	Complexity	Density
S1	Low	Low
S2	High	Low
S3	Low	High
S4	High	High

- *People detection*: assigns a confidence of being people for each candidate region (that could either a frame region or a blob extracted from the foreground binary mask) by computing their similarity with a trained person model [13]. Its outputs are the score (confidence) and location of each analyzed candidate.
- *Video object tracking*: consists on locating the objects of interest (i.e., targets) in the sequence frames [22]. Its output is the location of each tracked target.
- *Event recognition*: detects events using the output of the previous stages [1]. An event is defined as an action performed by one or multiple persons (e.g., walking, handshaking). For each detection, the output includes a descriptor with its spatio-temporal location (frame span and position) and score (detection confidence).

2.2 Scenario Classification

For each stage of the video-surveillance system, the evaluation process should consider different scenarios to appropriately represent real world conditions. For understanding the limitations of current approaches, each scenario is classified according to two criteria: complexity and density. The former describes whether the visual data represents situations that can be easily characterized or not. For example, video object segmentation is an (relatively) easy task for static cameras and scene backgrounds but its complexity highly increases when dealing with moving cameras or motion in the background. The latter considers the number of moving objects in the sequence. Independently of the stage, an increasing number of objects affects its performance. This criterion is particularly interesting in video-surveillance where crowded places are common scenarios (e.g., airports, mass sport events). For example, abandoned object detection presents variable difficulty depending on the moving people density (fewer people, less complexity). Finally, we consider two levels for each criterion (low and high) to define four evaluation scenarios (see Table 1). Sample frames of the evaluation scenarios are depicted in Fig. 2.

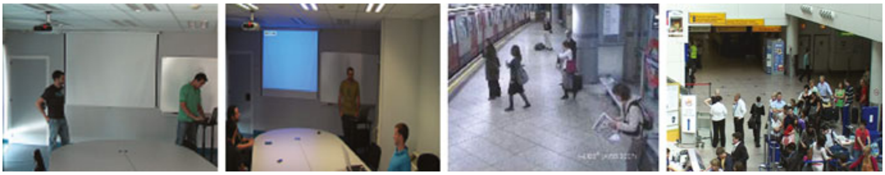


Fig. 2 Sample frames of the evaluation scenarios for event recognition. (*From left to right*): simple event *standing* (S1), complex event *UseObject* (S2), *abandonedObject* detection (S3) and complex event *bag stealing* (S4)

3 Related Work

In this section, we briefly review the state of art for each selected stage with respect to its datasets (see Table 2) and metrics, the two components of evaluation protocol.

3.1 *Video Object Segmentation*

Video object segmentation also known as foreground/background detection is a critical task in video-surveillance that presents many challenges related with, among others, shadows, camouflage, static objects and background motion [5]. For evaluating the existing approaches under such conditions, several datasets are available:

- VSSN2006:[2] provided within the VSSN Workshop 2006, this dataset consists of 14 sequences with artificial foreground objects introduced into real backgrounds for representing illumination changes, shadows and background motion (ground-truth data is provided for 10 sequences at pixel-level for every frame).
- IPPR06:[3] the IPPR contest motion segmentation dataset includes three different sequences of walking persons (with ground-truth at pixel-level for every frame) that model shadows, illumination changes and image noise.
- CVSG:[4] this dataset [34] consists of 14 sequences that represent the critical segmentation factors for foreground (appearance, size, velocity) and background (appearance, motion, multimodality) by artificially combining real foreground objects and backgrounds (with ground-truth at pixel-level for every frame).
- SABS:[5] this dataset [5] is an artificial dataset that represents nine common challenges of background subtraction for video-surveillance. It consists on nine sequences with isolated challenges which are divided into training and test data (with ground-truth at pixel-level for every frame).

[2] http://imagelab.ing.unimore.it/vssn06/

[3] http://media.ee.ntu.edu.tw/Archer_contest/

[4] http://www-vpu.eps.uam.es/DS/CVSG/

[5] http://www.vis.uni-stuttgart.de/index.php?id=sabs

Table 2 Categorization of existing datasets according to the scenarios of Table 1

	Covered scenario			
	S1	S2	S3	S4
Video object segmentation				
VSSN2006	X	X		
IPPR06	X			
CVSG	X	X		
SABS	X	X		
CDW2012	X	X		
People detection				
ETHZ			X	
TUD-Pedestrians			X	
DCII			X	
Caltech Pedestrian			X	
PDds	X	X	X	
Video object tracking				
PETS	X	X	X	
VISOR	X			
EPFL	X	X		
SOVTds	X		X	
Event detection				
CAVIAR	X	X		
ETISEO	X	X	X	
PETS 2006	X	X		
PETS 2007	X	X		X
I-LIDS	X		X	
VISOR	X	X		
CANDELA	X	X		
CANTATA	X			
ASODds	X	X		
EDds	X		X	

- CDW2012:[6] the IEEE Workshop on Change Detection 2012 proposed a rigorous benchmarking effort for representing well-known segmentation challenges captured in indoor and outdoor settings. In total, it has 31 sequences grouped into six categories (with ground-truth at pixel-level for every frame).

For ground-truth based metrics, video object segmentation can be evaluated at the lowest semantic level, that is, pixel-level, or at higher semantic levels, that is, region-

[6] http://www.changedetection.net

level, object-level, etc. In the literature, the pixel-level evaluation strategy is the most popular [5, 18]. It considers foreground detection as a binary classification of each pixel, resulting in a segmentation mask. The accuracy of this classification is expressed by means of recall (R), precision (P) and their harmonic mean, the F-score (F):

$$P = TP/(TP + FP), \tag{1}$$

$$R = TP/(TP + FN), \tag{2}$$

$$F = 2 \cdot P \cdot R/(P + R), \tag{3}$$

where TP, FP and FN indicate, respectively, the number of correct detections, false alarms and missed detections at pixel-level. For high-level evaluation, [7] used the center of the segmented objects whereas [24] focused on the splits and merges of foreground regions for composing the objects. In addition, [8] introduced spatio-temporal metrics derived from geometrical properties of the segmented objects.

Although non ground-truth based metrics are less popular, according to [29], they can be roughly classified into region (study the segmented regions), model (use available object models) or assisted (use complementary algorithms). Among them, the most relevant is [14] that defined the motion and color contrast along the boundaries of object regions and its adaptation for video object segmentation [29].

3.2 People Detection

The complexity of people detection is mainly related with the difficulty of modeling persons because of their huge variability in appearance, poses, movements, points of views and object-person interactions. This complexity is even higher in crowded video-surveillance scenarios which often include multiple persons, occlusions and background variability. Several datasets are available for its evaluation:

- ETHZ:[7] this dataset [15] consists of four stereo-sequences recorded in a real street walking scenario. For each one, it provides the sequences for both cameras, the camera calibration, the precomputed depth maps using the stereo images, and the ground-truth annotations (at bounding box level).
- TUD-Pedestrians:[8] this dataset [2] consists of 250 images (311 fully visible people) and two complex sequences (highly overlapped people showing significant variation in clothing and articulation), including the bounding box ground-truth.
- DCII:[9] the Daimler Mono Pedestrian Detection Benchmark Data Set II [13] consist of a sequence captured from a moving vehicle in a 27-minute drive through urban traffic and its associated ground-truth at bounding box level.

[7] http://www.vision.ee.ethz.ch/~aess/iccv2007/

[8] http://www.d2.mpi-inf.mpg.de/andriluka_cvpr08

[9] http://www.gavrila.net/

- Caltech Pedestrian Dataset:[10] this dataset [11] consists of approximately 10 hours of video (∼250000 frames divided into clips of 135 minutes) taken from a vehicle driving in an urban environment. In total, around 350000 bounding boxes and 2300 unique pedestrians were annotated. The annotation includes temporal correspondence between bounding boxes and detailed occlusion labels.
- PDds:[11] the PDds corpus [16] consists of 90 sequences for evaluation in video-surveillance covering the most common challenges with variable complexity. For each person, ground-truth is provided for each frame at bounding box level.

Regarding the metrics, people detection performance can be evaluated using ground-truth data at two levels: sequence sub-unit (frame, window, etc) or global sequence. Sub-unit performance is usually measured in terms of Detection Error Tradeoff (DET) [9, 12] or Receiver Operating Characteristics (ROC) [13, 23] curves. Global sequence performance is estimated through Precision-Recall (PR) curves [2, 21, 37]. The first level gives information of the classification stage, while the second one provides the overall system performance. In both cases the detector's output is a confidence score for each person detection, where larger values indicate higher confidence. Both evaluation methods compute progressively the respective parameters such as the number of false positives, Recall rate or Precision rate iterating from the lowest possible score to the highest possible score. Each score threshold iteration provides a point on the curve. On one hand, ROC curves represent the fraction of matched annotations with the detections (true positive rate, *TPR*, Recall or Sensitivity) vs. the fraction of wrong detections out of the negatives (non-people image samples) (false positive rate, *FPR* or 1-Specificity). On the other hand, PR curves represent also the *TPR* but in this case vs. the proportion of positive detections that are true positives (positive predictive value, *PPV* or Precision).

3.3 Video Object Tracking

Video object tracking is a complicated task due to high variability of the data to analyze as well as the many steps involved in the tracking process (feature extraction, target representation and propagation of the target model over time). For evaluating performance of tracking algorithms, several datasets are available:

- PETS:[12] the PETS Workshop series have been releasing a tracking-related dataset almost every year since 2000. As the dataset sizes are large and they cover real situations, these datasets are widely used in the research community. Among the existing datasets, the most important ones related to tracking are the PETS2000 (outdoor people and vehicle tracking for single camera), PETS2001 (outdoor people and vehicle tracking for single camera using two synchronized views) and

[10] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

[11] http://www-vpu.eps.uam.es/DS/PDds/

[12] http://www.cvg.cs.rdg.ac.uk/slides/pets.html

PETS2009 (outdoor people tracking in crowded environments with multicamera setup). However, PETS datasets have two limitations: there is no ground-truth available and the challenges proposed are focused on event recognition (i.e., without describing the specific tracking problems for each video).

- VISOR:[13] this video repository has been conceived as a support tool for different video-surveillance projects [36]. Related to tracking, it includes six sequences (without ground-truth data) covering common problems such as occlusions, scale changes and complex movements.
- EPFL:[14] this dataset is oriented to multicamera settings for outdoor and indoor video-surveillance. It contains five scenarios with around 30 sequences showing occlusions and scale changes. Although camera calibration is provided for all the scenarios, ground-truth data is only available for some sequences.
- SOVTds:[15] this dataset is provides an extensive coverage of the common tracking-related problems in video-surveillance. For each problem, it is designed with four complexity levels including both real and synthetic sequences carefully selected from other datasets (related and non-related with video tracking). It contains 125 sequences and the associated ground-truth for every frame.

For video object tracking evaluation, metrics based on ground-truth can be divided into frame or sequence level. Frame-level considers the information within the frame being similar to an estimation of classification performance. Hence, standard Precision and Recall (Eqs. 1 and 2) are used for computing the spatial similarity between estimations and ground-truth locations of targets at pixel [25] or object-level [3]. Sequence-level measures the accuracy of the target trajectories such as the temporal accumulation of frame-level pixel accuracy [25] or the trajectory fragmentation [19] (i.e., the number of generated segments).

Approaches for tracking evaluation without ground-truth can be grouped into trajectory-based, feature-based and hybrid categories [30]. Trajectory-based approaches analyze the generated trajectories in which the time-reversibility of object motion is commonly used [38]. Feature-based approaches analyze target feature variation [30] or compute statistics for checking model consistency such as the covariance of the target state [26]. Finally, hybrid category describes the combinations of the previous approaches such as the use of the time-reversibility and the covariance analysis [32].

3.4 Event Recognition

As event recognition considers all the outputs of the stages that compose the video-surveillance system and therefore, its performance is influenced by all the factors affecting each stage. For evaluating its performance, several datasets are available:

[13] http://www.openvisor.org/

[14] http://cvlab.epfl.ch/data/pom/

[15] http://www-vpu.eps.uam.es/DS/SOVTds

- CAVIAR:[16] this dataset includes 17 sequences of human activities for indoor video-surveillance. It covers several events (with ground-truth data) such as people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place.
- ETISEO:[17] this dataset [25] contains 86 indoor and outdoor video-surveillance sequences (corridors, streets, building entries, subway,...) with different types of complexity levels. Several events are annotated considering person-object interactions as well as person movement.
- PETS 2006:[18] this dataset is focused on multicamera sequences for *abandoned luggage* detection with increasing scene complexity in terms of nearby people. It contains 28 sequences (∼1–2 minutes long) with 24 annotated events.
- PETS 2007:[19] this dataset considers the events *loitering*, *stolen luggage* and *abandoned luggage* in a crowded scenario. A four-camera setting is employed to record, 32 sequences (∼2–3 minutes long) containing 36 events in total.
- I-LIDS:[20] this dataset has three sequences (∼3.5 minutes long) for abandoned object detection at an underground station classified into three complexity levels (easy, medium, and hard), which are defined considering the crowd density.
- ViSOR:[21] this dataset is classified in different categories including outdoor and indoor events (human actions, traffic monitoring, cast shadows,...). A total of 140 sequences with variable length is available for events related with human-object interactions (*abandoned object*, *Leave car, Enter Car,...*).
- CANDELA:[22] this dataset contains 16 indoor sequences (∼30 secs long) for *abandoned object*, including interactions between object owners. Despite the simplicity of the scenario, the low resolution and the relatively small size of objects present challenges for detecting the events.
- CANTATA:[23] this dataset is focused on *abandoned* and *stolen object*s in non-crowded outdoor scenarios. A total of 31 sequences (∼2 minutes long) are available from two different views (leaving and removing objects in the sequences).
- ASODds:[24] this dataset provide a representative test-set for discriminating previously detected stationary regions in video-surveillance systems able to detect abandoned and stolen objects. Annotations of both events are also provided. Sequences (over 100) have been extracted from related public datasets.
- EDds:[25] this dataset contains 17 sequences (∼3–4 minutes long) focused on human-related events for indoor video-surveillance considering interactions be-

[16] http://homepages.inf.ed.ac.uk/rbf/CAVIAR

[17] http://www-sop.inria.fr/orion/ETISEO/intro_presentation.htm

[18] http://www.cvg.rdg.ac.uk/PETS2006/data.html

[19] http://www.cvg.rdg.ac.uk/PETS2007/data.html

[20] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html

[21] http://www.openvisor.org/

[22] http://www.multitel.be/~va/candela/abandon.html

[23] http://www.multitel.be/~va/cantata/LeftObject/

[24] http://www-vpu.eps.uam.es/DS/ASODds

[25] http://www-vpu.eps.uam.es/DS/EDds

Table 3 Critical factors in video object segmentation specified in the CVSG dataset

Foreground		Background	Camera
Single objects	Groups		
Textural complexity, apparent velocity, object structure, uncovered extent, object size	Largest difference, object interactions	Textural complexity, multimodality	Motion

tween persons and environmental objects and activities without involving physical contact. In particular, two activities (*HandUp* and *Walking*) and three person-object interactions (*Leave*, *Get* and *Use object*) have been annotated.

For event recognition, the common evaluation scheme is to optimally determine the match between ground-truth annotations the event detections. This one-to-one mapping can be done temporally or spatio-temporally [25]. The former only considers the duration of the detection and the annotation whereas the latter extends it by including a constraint for similar spatial locations. Moreover, an additional constraint can be imposed considering the confidence of the detected event [31].

4 Evaluation Protocols

In this section, we introduce the proposed protocols for performance evaluation of the selected video-surveillance stages within the scope of the EventVideo project.

4.1 Video Object Segmentation

4.1.1 Selected Dataset

For this stage, the Chroma Video Segmentation Ground-truth (CVSG) dataset [34] is selected as it covers the main problems of video object segmentation. It consists of a set of video sequences obtained according to a thorough study of the critical factors affecting segmentation performance (summarized in Table 3). As specific values of these factors can significantly increase or decrease the complexity of the segmentation task (and therefore, the expected algorithm accuracy), they are convenient for designing multiple sequences with variable complexity. Foreground objects have been recorded in a chroma studio, in order to automatically obtain pixel-level high quality segmentation masks with different foreground factors. Then, real scene backgrounds are also recorded with different camera and background factors. Finally, the resulting corpus consists on the composition of the foreground and background sequences obtaining a total of 14 sequences ($\sim$7000 frames). Some examples are shown in Fig. 3. As it can be observed, they present low density scenarios with variable complexity thus covering the S1–S2 scenarios defined in Table 1.

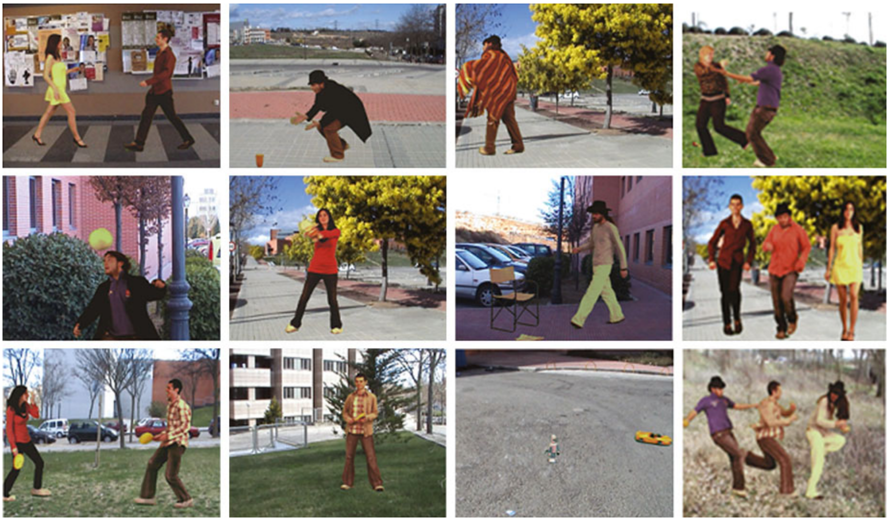


Fig. 3 Sample frames for the sequences of the CVSG dataset

4.1.2 Metrics Based on Ground-Truth Data

As a first approach, we have selected the pixel-wise evaluation based on ground-truth data [18]. In order to evaluate and compare the segmentation techniques, we have selected the precision and recall measures for foreground ($P1$, $R1$) and background ($P0$, $R0$) detection:

$$P0 = TN/(TN+FN), \qquad R0 = TN/(TN+FP), \tag{4}$$

$$P1 = TP/(TP+FP), \qquad R1 = TP/(TP+FN), \tag{5}$$

where TP indicates the number of foreground pixels correctly detected, TN the number of background ones correctly detected, FP the number of foreground pixels wrongly detected as background and FN the number of background ones wrongly detected as foreground. Additionally, the F-Score measure has been selected to combine P and R measures for foreground ($F1$) and background ($F0$) results:

$$F0 = 2 \cdot P0 \cdot R0/(P0+R0), \tag{6}$$

$$F1 = 2 \cdot P1 \cdot R1/(P1+R1). \tag{7}$$

In order to achieve the objective of evaluating and finding the optimal parameters of the algorithms, it have been maximized the average of the F-score measures for foreground and background, $F0$ and $F1$.

4.1.3 Metrics not Based on Ground-Truth Data

We also evaluate segmentation performance without ground-truth data by means of the color-based metric $DC1$ proposed by [29]. It relies on comparing the boundaries

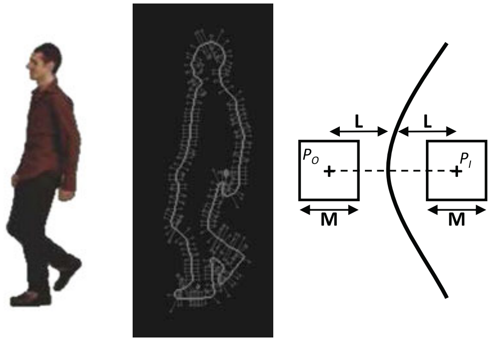


Fig. 4 Boundary-based contrast scheme proposed by [29]. (**a**) Segmented object, (**b**) its boundary with the normal lines and (**c**) a zoom on a boundary pixel location

of the segmented objects against the color boundaries extracted from each frame. The scheme is depicted in Fig. 4. For each boundary pixel, a normal line of length $2L+1$ is defined and the color differences between the initial (P_I) and ending (P_O) points of this line are obtained in a $M \times M$ patch as follows:

$$CD(t;i) = \frac{\|P_O^i(t) - P_I^i(t)\|}{\sqrt{3 \cdot 255^2}}, \tag{8}$$

where $P_O^i(t)$ and $P_I^i(t)$ are the mean colors of the $M \times M$ patches centered atP_I and P_O points (using the RGB color space quantified into 256 levels) extracted from each ith boundary pixel of the foreground region at time t. $CD(t;i)$ ranges from 0 to 1 if both points belong to, respectively, the same or different color regions.

Then, the evaluation of the foreground segmentation for each region, O_j is performed and combined for multiple foreground regions as follows:

$$DC1_{O_j}(j) = \frac{1}{K_t}\sum_{i=1}^{K_t} CD(t;i,j), \tag{9}$$

$$DC1(t) = \min_j\big(DC1_{O_j}(t)\big), \tag{10}$$

where K_t is the number of boundary pixels, $CD(t;i,j)$ is the color difference of the ith boundary pixel of the jth analyzed foreground region. Its value ranges from 0 (lowest segmentation quality) to 1 (highest segmentation quality). Finally, the mean of $DC1(t)$ is taken over all the sequence frames to get an evaluation score.

4.2 People Detection

4.2.1 Selected Dataset

For this stage, the Person Detection dataset (PDds) [16] is selected as it covers the main problems affecting people detection in video-surveillance. It consists of a set of sequences with different levels of complexity and their associated ground-truth

Table 4 Critical factors in people detection

Background		Classification	
Textural complexity	Variability	Appearance variability	People-object interactions
Low, medium, high	Lighting changes, view changes, multimodal	Pose variations, different clothes, carry objects	Objects, people, objects & people

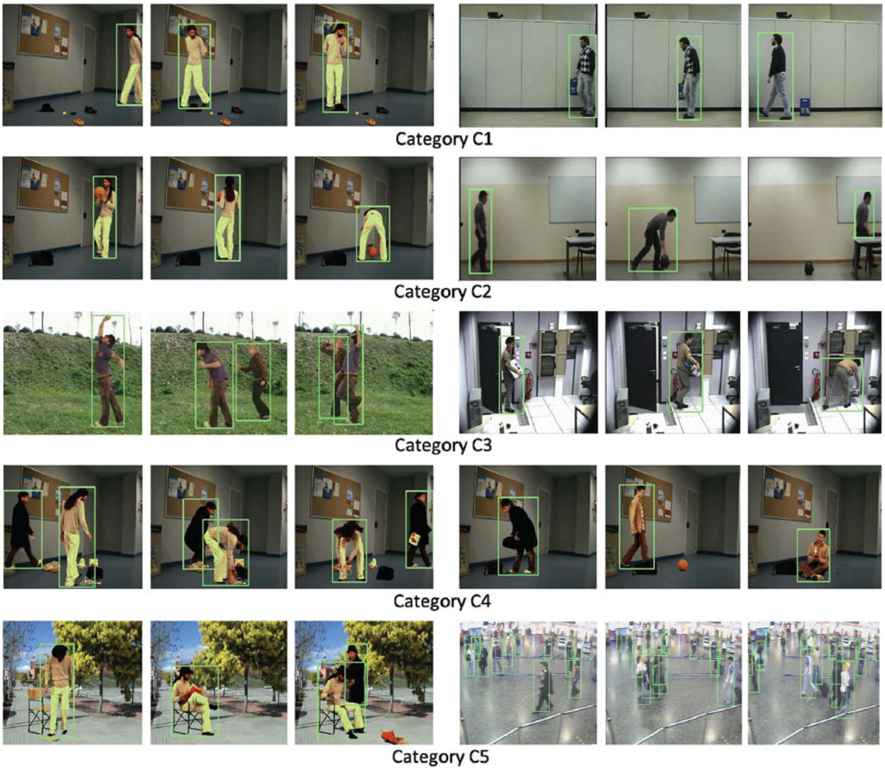
Table 5 Description of the PDds dataset and their associated critical factors

Sequence	Category	Subcategory	Background		Classification	
			Textural complexity	Variability	Appearance variability	People/object interactions
1–4	C1	C1-a	Low	Low	Low	Low
5–6	C1	C1-b	Low	Medium	Low	Low
7–8	C2	C2-a	Low	Low	Medium	Low
9–10	C2	C2-b	Low	Low	Medium	Medium
11–12	C2	C2-c	Low	Medium	Low	Medium
13	C3	C3-a	Medium	Medium	Medium	Low
14–16	C3	C3-b	Medium	Medium	Medium	Medium
17–18	C4	C4-a	Low	Low	Medium	High
19–20	C4	C4-b	Low	Low	High	Medium
21	C4	C4-c	Low	Low	High	High
22–24	C5	C5-a	Medium	High	Medium	High
25	C5	C5-b	Medium	High	High	Medium
26	C5	C5-c	High	High	Medium	High
27–33	C5	C5-d	High	High	High	Low
34–65	C5	C5-e	High	High	High	Medium
66–90	C5	C5-f	High	High	High	High

(bounding box annotations for each frame). Sequences have been classified into different complexity categories depending on previously identified critical factors for people detection performance. Table 4 summarizes such factors and Table 5 lists the video sequences and their complexity. Sample frames are shown in Fig. 5. The resulting corpus contains 91 sequences (∼28000 frames) exceeding other public pedestrian datasets in the amount of data and its complexity variability. As it can be observed, they present low density scenarios with variable complexity thus covering the S1, S2 and S3 scenarios defined in Table 1.

4.2.2 Metrics Based on Ground-Truth Data

For evaluating people detection performance based on ground-truth, we aim to compare the overall performance of different detection systems, so we have chosen the



Category C1

Category C2

Category C3

Category C4

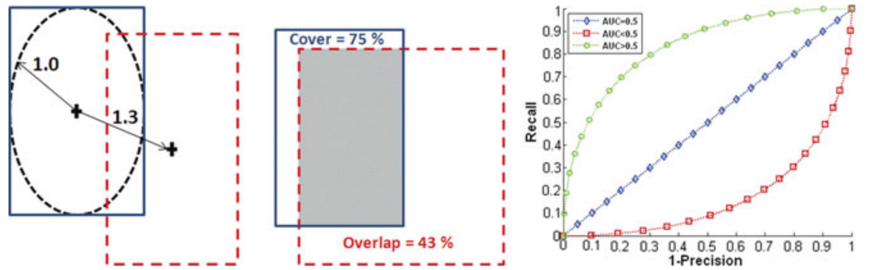
Category C5

Fig. 5 Sample frames for the categories of the PDds dataset

PR evaluation method (see Sect. 3.2). For each value of the detection confidence, PR curves compute Precision and Recall as shown in Eqs. 1 and 2.

In order to evaluate not only the (binary) yes/no detection but also the precise pedestrians locations and extents, we use three criteria, defined by [20], that allow comparing hypotheses at different scales: the relative distance, cover, and overlap. The relative distance dr measures the distance between the bounding box centers in relation to the size of the annotated bounding box (see Fig. 6a). Cover and overlap measure how much of the annotated bounding box is covered by the detection hypothesis and vice versa (see Fig. 6b). A detection is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25 % of the true object size) and cover and overlap are both above 50 %. Only one hypothesis per object is accepted as correct, so any additional hypothesis on the same object is considered as a false positive.

We usually use the integrated Average Precision (AP) to summarize the overall performance, represented geometrically as the area under the PR curve (AUC-PR), in order to express more clearly the results we have chosen the representation Recall vs 1-Precision (see Fig. 6c). In addition, focusing on the people detection evaluation in video security systems, we want also to evaluate the detector at the operating point, that is, at the predefined optimal decision threshold for each algorithm. Thus,



(a) Relative distance criterion for comparing bounding boxes.

(b) Cover and overlap criteria for comparing bounding boxes.

(c) (1-Precision)-Recall curve and area under the curve.

Fig. 6 Performance evaluation metrics for people detection

Table 6 Complexity factors for the video tracking dataset

Problem	Criteria (factors)
Complex movement	The target changes its speed (pixels/frame) abruptly in consecutive frames
Gradual illumination	The average intensity of an area changes gradually with time until a maximum intensity difference is reached
Abrupt illumination	The average intensity of an area changes abruptly with respect to its surroundings (maximum intensity difference)
Noise	It includes natural (snow) or white Gaussian noise which is manually added with varying deviation value
Occlusion	Objects in the scene occlude a percentage of the target
Scale changes	The target changes its size with a maximum relative change regarding its original size
Similar objects	An object with similar color to the target appears in the neighborhood of the target

we can compare the final operational performance and not just its overall performance.

4.3 Video Object Tracking

4.3.1 Selected Dataset

For this stage, the Single Object Video Tracking dataset (SOVTds) is selected to evaluate single-object tracking algorithms for video-surveillance. SOVTds covers seven common tracking problems in video-surveillance by identifying its critical factors (see Table 6). Then, it organizes the sequences into four situations: synthetic, real laboratory, simple real and complex real data. For the first two situations, the

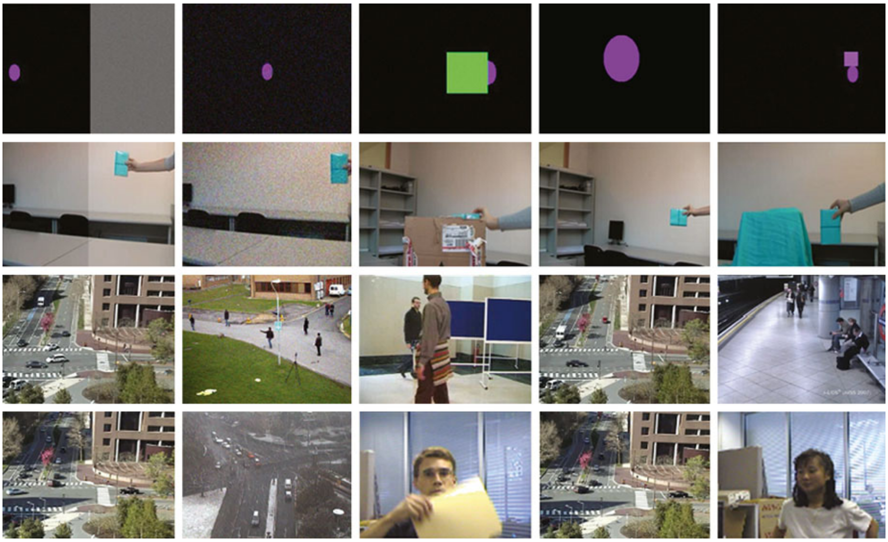


Fig. 7 Sample frames for the situations of the proposed dataset (*from top row to bottom row*): synthetic, laboratory, Simple real and Complex real. Samples of some tracking problems are also presented for each column (*from left to right*): abrupt illumination change, noise, occlusion, scale change and (color-based) similar objects

sequences were recorded trying to isolate the tracking problems whereas the last two situations contain carefully selected clips from existing datasets. In total, the corpus has 125 sequences (∼23000 frames). Sample frames are shown in Fig. 7. Moreover, the complexity of the tracking problems is estimated for each sequence through the factors. As this dataset represents simple and complex problems in nonhighly crowded situations, it covers the S1, S2 and S3 scenarios defined in Table 1.

4.3.2 Metrics Based on Ground-Truth Data

In order to evaluate the tracking accuracy, the SFDA (Sequence Frame Detection Accuracy) metric was chosen which calculates for each frame the spatial overlap between the estimated target location and the ground-truth annotation.

$$SDFA = \frac{\sum_{t=1}^{N_{\text{frames}}} FDA(t)}{\sum_{t=1}^{N_{\text{frames}}} \exists(N_{GT}^t + N_P^t)} \tag{11}$$

$$FDA(t) = \frac{OverlapRatio}{\frac{N_{GT}^t + N_P^t}{2}} \tag{12}$$

where N_{frames} is the number of frames, N_{GT}^t and N_P^t represent the number of ground-truth and estimated locations in the th frame, $\exists(\cdot)$ indicates if ground-truth or estimation data exist for the th frame and *OverlapRatio* is the pixel-level spatial overlap between both locations divided by their area sum.

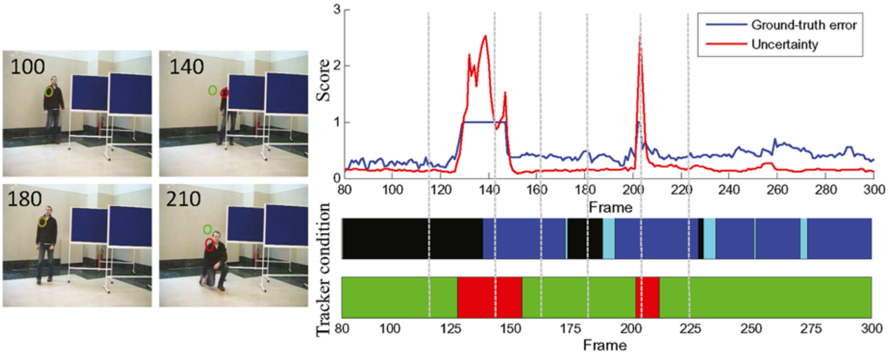
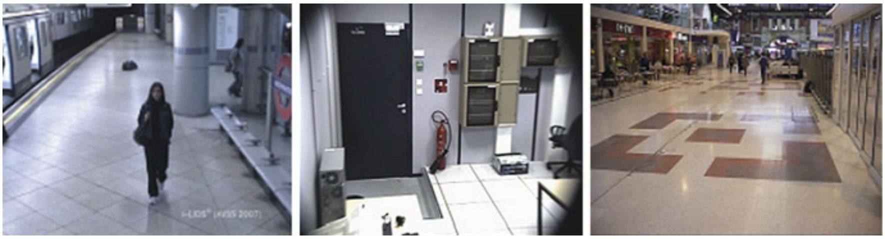


Fig. 8 Tracking results, tracker condition estimation and temporal segmentation for target H5 (occlusion_1 sequence; frames shown are 100, 140, 180 and 210) [32]. Tracking results and ground-truth annotations are represented as *green* and *red* ellipses, respectively. (*Green*: successful tracking; *Red*: unsuccessful tracking; *Black*: scanning; *Cyan*: locking in; *Blue*: locked on.)

4.3.3 Metrics not Based on Ground-Truth Data

For estimating tracking performance without ground-truth data, we use [32] which is based on estimating the uncertainty of the tracking algorithm (i.e., tracker) and then, analyzing its values to decide whether it is successful or not. Such uncertainty, S_t, can be used as indicator of periods of unstable output data (e.g., wrong target estimation) allowing the tracker evaluation. It can be measured by analyzing the state-space representation of particle-filter based approaches [22] or by adapting the output of deterministic trackers such as for Mean-shift tracking [33].

Then, we identify when the tracker is stable (i.e., following the target) by detecting changes of S_t within a window of length λ. We compute two relative variations of uncertainty for the change of $S_{t-\lambda}$ with respect to S_t and vice versa, using two lengths for short and long term changes (λ_1 and λ_2) as defined in [32]. The former change indicates low-to-high uncertainty changes whereas the latter represents high-to-low uncertainty changes. As a result, four signals are computed by combining the two variations and the two lengths. Then, changes on the four signals are detected by using a three-threshold scheme and combined in a finite-state machine for estimating the tracker condition: focused on the target, scanning the video frame or locking on the target after a failure [32]. Finally, we use time-reversed analysis to check the tracker recovery when it focuses on an object after failure (transition from third to first tracker condition) as it might be on a distractor (background objects with features similar to those of the target). A tracker in reverse direction from this recovery instant until a reference point (the last time instant when the tracker was successful) [32] and the spatial overlap between the reverse and the forward trackers (the one to evaluate) is computed for determining if the tracker has recovered or not. Figure 8 shows an example of tracker condition and successful estimation.



Category 1 Category 2 Category 3

Fig. 9 Sample frames for the available categories in the ASODds dataset

Table 7 ASODds dataset description

Category	Number of annotations (blobs)				Complexity
	Annotated sequences		Real sequences		
	Abandoned	Stolen	Abandoned	Stolen	
C1	771	442	756	863	Low
C2	666	316	794	397	Medium
C3	595	174	852	660	High
All	2032	932	2402	1920	

4.4 Event Detection

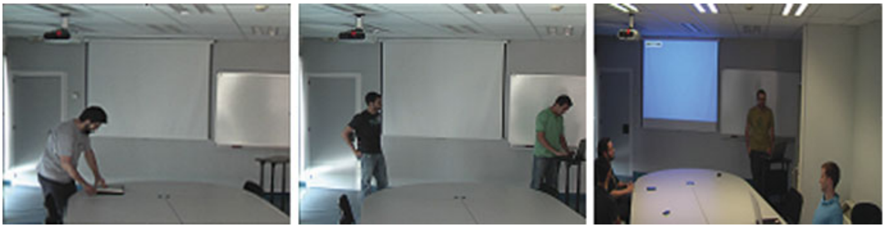
4.4.1 Selected Datasets

For event detection, two datasets have been selected: the Abandoned and Stolen Discrimination dataset (ASODds) and the Event Detection dataset (EDds).

Abandoned and Stolen Object Discrimination Dataset—ASODds The ASODds dataset [6] consists of two annotation sets of the foreground binary masks for abandoned and stolen objects. The first one has been obtained by manually annotating the objects of interest in the video sequence (annotated data). The second one represents real data has been obtained by running [28] over the sequences to get inaccurate masks (real data). Then, the sequences have been grouped into three categories according to a subjective estimation of the background complexity that consists on the presence of edges, multiple textures, lighting changes, reflections, shadows and objects belonging to the background. Currently, three categories have been defined considering low (C1), medium (C2) and high (C3) background complexity. According to the criteria proposed in Sect. 2, the categories C1 and C2 present low complexity and few number of objects (situation S1) whereas the C3 covers low complex and crowded scenarios (situation S3). Sample frames of such categories are shown in Fig. 9 and a summary of the annotated events in the dataset and the associated complexity of each category is available in Table 7.

Table 8 EDds dataset description. The complexity estimation codes are Low (L), Medium (M), High (H) and Very High (V). The events are Leave-object (LEA), Get-object (GET), Use-object (USE), Hand Up (HUP) and Walking (WLK)

Sc1	Events occurrences					Complexity estimation			
	Iterations			Activities		S1	S2	S3	S4
	LEA	GET	USE	HUP	WLK				
1	18	13	9	9	54	M	L	M	M
2	7	7	10	14	44	M	M	M	H
3	14	14	22	20	10	V	H	V	V



Category 1 Category 2 Category 3

Fig. 10 Available categories in the EDds dataset

Event Detection Dataset—EDds Currently, the dataset EDds [31] contains 17 sequences recorded using a stationary camera at resolution of 320 × 240 at 12 fps. It is focused on two types of human-related events: interactions and activities. In particular, two activities (*HandUp* and *Walking*) and three human-object interactions (*Leave*, *Get* and *Use object*) have been annotated. Moreover, all the test sequences have been grouped into three categories according to a subjective estimation of the analysis complexity according to the criteria defined in the previous subsections for the foreground, tracking, feature and event stages that compose a typical event detection system. A summary of the annotated events in the dataset and the associated complexity of each category is available in the Table 8. Sample frames of such categories are shown in the Fig. 10.

4.4.2 Metrics Based on Ground-Truth Data

For matching event annotations and detections, we use the following conditions:

$$Match\big(E^{GT}, E^{D}\big) = \begin{cases} 1 & \text{if } score > \rho \quad \wedge \\ & |T^{D}_{\text{start}} - T^{GT}_{\text{start}}| < \tau_1 \quad \wedge \\ & |T^{D}_{\text{end}} - T^{GT}_{\text{end}}| < \tau_2 \quad \wedge \\ & \frac{2|A^{GT} \cap A^{D}|}{|A^{GT}| + |A^{D}|} > \sigma \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Table 9 Classification of datasets according to criteria defined in Sect. 2.2. The (–) indicates that the dataset partially fulfills the requirements of such criterion

		Density	
		Low	High
Complexity	Low	CVSG, PDds, SOVTds, ASODds, EDds	PDds (–), ASODds (–)
	High	PDds (–), SOVTds (–), EDds (–)	

where E^{GT} and E^{D} are the annotated and detected events; score is the detection probability; (T^{D}_{start}; T^{D}_{end}) and (T^{GT}_{start}; T^{GT}_{end}) are the frame intervals of the annotated (GT) and detected (D) events; A^{GT} and A^{D} represent the average area (in pixels) of each event; $|A^{GT} \cap A^{D}|$ is their average spatial overlap (in pixels); ρ, τ_1, τ_2 and σ are positive thresholds (heuristically set to the values $\rho = 0.75$, $\tau_1 = \tau_2 = 100$, and $\sigma = 0.5$).

Then, we use the Precision (P) and Recall (R) measures for evaluating the performance of the matching process. Precision is the ratio between the correct and the total number of detections. Recall is the ratio between the correct detections and the total number of annotations. We also use the F-score measure, β, to combine Precision and Recall as shown in Eqs. 1 and 2.

5 Conclusions

In this chapter, we have presented the material for performance evaluation within the EventVideo project. In particular, we have selected some stages: video object segmentation, people detection, video object tracking and event detection. Then, we have described the employed datasets and protocols for their evaluation in Sect. 4 (CVSG, PDds, SOVTds, ASODds y EDds; all of them available at http://www-vpu.eps.uam.es/webvpu/en/recursos-publicos/datasets/).

In addition, a novel methodology that does not follow the traditional ground-truth based approach has been presented in Sects. 4.1.3 and 4.3.3 for, respectively, the video object segmentation and tracking stages. Moreover, according to the scenario classification of Sect. 2.2 (with the variables complexity and density), the datasets used in the EventVideo project are categorized as listed in Table 9.

As future work, the selected datasets will be used for comparing the most recent approaches for evaluating the current status of the state-of-the-art (and which of the criteria in Table 9 could be considered as achieved). Moreover, we will consider the extension of the datasets to cover the highest levels of the defined situations and the inclusion of additional information to help visual analysis (such as depth and laser).

Acknowledgements Work supported by the Spanish Government under Project TEC2011-25995 EventVideo.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: a review. ACM Comput. Surv. **43**(3), 16:1–16:43 (2011)
2. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, pp. 1–8 (2008)
3. Bashir, F., Porikli, F.: Performance evaluation of object detection and tracking systems. In: Proc. IEEE Int. Workshop Perform. Eval. Track. Surveill, New York, USA, pp. 7–14 (2006)
4. Baumann, A., Boltz, M., Ebling, J., Koenig, M., Loos, H.S., Merkel, M., Niem, W., Warzelham, J.K., Yu, J.: A review and comparison of measures for automatic video surveillance systems. EURASIP J. Image Video Process., 1–30 (2008)
5. Brutzer, S., Heferlin, B., Heidemann, G.: Evaluation of background subtraction techniques for video surveillance. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, pp. 1937–1944 (2011)
6. Campos, L.C., SanMiguel, J.C., Martínez, J.M.: Discrimination of abandoned and stolen object based on active contours. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal based Surveillance, Klagenfurt, Austria, pp. 101–106 (2011)
7. Charles, J.J., Kuncheva, L.I., Wells, B., Lim, I.S.: An evaluation measure of image segmentation based on object centres. In: Proc. of the International Conference on Image Analysis and Recognition, pp. 283–294 (2006)
8. Correia, P., Pereira, F.: Objective evaluation of video segmentation quality. IEEE Trans. Image Process. **12**(2), 186–200 (2003)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 886–893 (2005)
10. Dee, H., Velastin, S.: How close are we to solving the problem of automated visual surveillance? Mach. Vis. Appl. **19**(5), 329–343 (2008)
11. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 304–311 (2009)
12. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
13. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. **31**(12), 2179–2195 (2009)
14. Erdem, C., Sankur, E., Tekalp, A.: Performance measures for video object segmentation and tracking. IEEE Trans. Image Process. **13**(7), 937–951 (2004)
15. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: IEEE Int. Conf. on Computer Vision, Rio de Janeiro, Brazil, pp. 1–8 (2007)
16. García-Martín, A., Martínez, J.M., Bescós, J.: A corpus for benchmarking of people detection algorithms. Pattern Recognit. Lett. **33**(2), 152–156 (2012)
17. Greoris, B., Bremond, F., Thonnat, M.: Real-time control of video surveillance systems with program supervision techniques. Mach. Vis. Appl. **18**(3), 185–205 (2007)
18. Herrero, S., Bescós, J.: Background subtraction techniques: systematic evaluation and comparative analysis. In: Proc. of the Advanced Concepts for Intelligent Vision Systems, Bordeaux, France, pp. 33–42 (2009)
19. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 319–336 (2009)
20. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 878–885 (2005)
21. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. Int. J. Comput. Vis. **77**(1–3), 259–289 (2008)

22. Maggio, E., Cavallaro, A.: Video Tracking: Theory and Practice. Wiley, New York (2011)
23. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1863–1868 (2006)
24. Nascimento, J., Marques, J.: Performance evaluation of object detection algorithms for video surveillance. IEEE Trans. Multimed. **8**(4), 761–774 (2006)
25. Nghiem, A.-T., Bremond, F., Thonnat, M., Valentin, V.: Etiseo, performance evaluation for video surveillance systems. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal based Surveillance, London, UK (2007)
26. Nickels, K., Hutch, S.: Estimating uncertainty in ssd-based feature tracking. Image Vis. Comput. **20**(1), 47–58 (2002)
27. Prati, A., Mikic, I., Trivedi, M., Cucchiara, R.: Detecting moving shadows: algorithms and evaluation. IEEE Trans. Pattern Anal. Mach. Intell. **25**(7), 918–923 (2003)
28. SanMiguel, J.C., Martínez, J.M.: Robust unattended and stolen object detection by fusing simple algorithms. In: Proc. of IEEE Int. Conf. on Advanced Video and Signal based Surveillance, Santa Fe, USA, pp. 18–25 (2008)
29. SanMiguel, J.C., Martínez, J.M.: On the evaluation of background subtraction algorithms without ground-truth. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, Boston, USA, pp. 180–187 (2010)
30. SanMiguel, J.C., Cavallaro, A., Martínez, J.M.: Evaluation of on-line quality estimators for object tracking. In: Proc. of the IEEE Int. Conf. on Image Processing, Hong Kong, China, pp. 825–828 (2010)
31. SanMiguel, J.C., Escudero-Viñolo, M., Martínez, J.M., Bescós, J.: Real-time single-view video event recognition in controlled environments. In: Proc. of the Int. Workshop on Content-Based Multimedia Indexing, Madrid, Spain, pp. 91–96 (2011)
32. SanMiguel, J.C., Cavallaro, A., Martínez, J.M.: Adaptive online performance evaluation of video trackers. IEEE Trans. Image Process. **21**(5), 2812–2823 (2012)
33. SanMiguel, J.C., Cavallaro, A., Martínez, J.M.: Standalone evaluation of deterministic video tracking. In: IEEE Int. Conference on Image Processing, Orlando, FL, USA, pp. 1353–1356 (2012)
34. Tiburzi, F., Escudero, M., Bescós, J., Martínez, J.M.: A ground-truth for motion-based video-object segmentation. In: Proc. of IEEE Int. Conf. on Image Processing, San Diego, USA, pp. 17–20 (2008)
35. Valera, M., Velastin, S.: Intelligent distributed surveillance systems: a review. IEE Proc., Vis. Image Signal Process. **152**(2), 192–204 (2005)
36. Vezzani, R., Cucchiara, R.: Video surveillance online repository (ViSOR): an integrated framework. Multimed. Tools Appl. **50**(2), 359–380 (2010)
37. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 794–801 (2009)
38. Wu, H., Sankaranarayanan, A., Chellappa, R.: Online empirical evaluation of tracking algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1443–1458 (2010)