Computer Vision and Image Understanding 152 (2016) 103–117

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA

Carlos Cuevas*, Eva María Yáñez, Narciso García

Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid (UPM), ES-28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 9 December 2015
Revised 18 April 2016
Accepted 11 August 2016
Available online 22 August 2016

Keywords:
Database
Dataset
Ground-truth
Moving object detection
Foreground detection
Background subtraction
Challenges
Stationary foreground
Moving camera

ABSTRACT

A public, complete, compact, and well structured database is proposed, which allows to test moving object detection strategies. The database is composed of many real indoor and outdoor sequences organized in different categories, each of one covering a specific challenge. In contrast to other databases, the proposed one is fully annotated at both pixel and object levels. Therefore, it is suitable for strategies exclusively focused on the detection of moving objects and also for those that integrate tracking algorithms in their detection approaches. Additionally, it contains sequences recorded with static and moving cameras and it also provides information about the moving objects remaining temporally static.

To test its usefulness, the database has been used to assess the quality of some outstanding moving object detection methods.

1. Introduction

The quantity and variety of electronic devices endowed with a camera have grown tremendously in the past few years (e.g. smartphones, digital cameras, video game consoles, intelligent vehicles, or augmented reality glasses). As a consequence of this growth, the demand for new computer vision applications to perform high-level analysis tasks (e.g. action recognition, object classification, people counting or virtualization) is rapidly increasing (Shapiro, 2013).

As a first stage, many of these applications include moving object detection algorithms (Chiranjeevi and Sengupta, 2012), the results of which influence the quality of the later stages. Therefore, the proper functioning of the detection module is critical for all these applications and several algorithms to efficiently detect moving objects have been developed and published up to date (Radke et al., 2005; Sobral and Vacavant, 2014).

In spite of the huge importance of moving object detection strategies, there are only a few labeled databases (i.e. with ground-truth data) for the assessment of their quality (Haines and Xiang, 2014).

On the one hand, some databases are fully labeled. However, they do not use real videos but synthetic sequences (Bourdis et al., 2011; Brutzer et al., 2011). Therefore, they are not able to represent some complex realistic conditions such as illumination changes or shadows cast by moving objects.

On the other hand, databases using natural video are not fully labeled because it is a labor-intensive task. They contain a reduced set of labeled images per sequence (sometimes only one image) (Toyama et al., 1999; Yin et al., 2007), or only some areas of interest in the images are labeled (Goyette et al., 2012). In addition, all of these datasets provide pixel-level labels but do not include object-level labels. Consequently, since many sequences contain multiple moving objects with complex and overlapping trajectories, it is not possible to maintain the identities of the objects throughout the sequences, which could be useful for the testing of the performance of strategies that combine detection and tracking algorithms (Creusot, 2014; Cuevas et al., 2010; Gallego et al., 2012).

Additionally, many recent moving object detection strategies are aimed at solving specific problems or challenges, such as the achievement of results robust to shadows cast by moving objects (Amato et al., 2014; Horprasert et al., 1999) or the detection of moving objects remaining temporally static (Baxter et al., 2015; Maddalena and Petrosino, 2013). However, although the existing databases are commonly structured according to different challenges, they do not separate adequately the challenges, making

* Corresponding author.
E-mail addresses: ccr@gti.ssr.upm.es (C. Cuevas), eyg@gti.ssr.upm.es (E.M. Yáñez), narciso@gti.ssr.upm.es (N. García).

http://dx.doi.org/10.1016/j.cviu.2016.08.005
1077-3142/

it difficult to test the quality of such strategies (e.g. a set of sequences with objects causing shadows but never remaining static would be desirable). Moreover, the length of the sequences is usually much higher than required, which also complicates the process of evaluation of the algorithms.

In this paper, we propose a complete, useful and publicly available[1] database named LASIESTA (Labeled and Annotated Sequences for Integral Evaluation of SegmenTation Algorithms), which has been designed to avoid the drawbacks of the aforementioned databases.

LASIESTA is composed of a large amount of real sequences recorded in a great variety of indoor and outdoor scenarios. All the sequences are completely labeled at both pixel and object levels. They are structured into well separated categories, each one oriented to a specific challenge (e.g. occlusions, shadows, illumination changes, camouflage, etc.). Additionally, it features sequences with camera motion, which also allows the evaluation of strategies that work on sequences recorded with cameras that are not completely static (Kong et al., 2010; Ye et al., 2011). Finally, another noteworthy feature is that the LASIESTA database in the only one including information about the moving objects remaining temporally static, which makes it adequate for assessing the quality of approaches focused on the detection of abandoned objects (Baxter et al., 2015; Maddalena and Petrosino, 2013).

Thanks to the mentioned features, LASIESTA is the most suitable database for assessing the quality of strategies for moving object detection.

The structure of this paper is as follows. First, Section 2 presents the state of the art concerning databases focused on the evaluation of image and video processing algorithms. Then, in Section 3, the most relevant databases to evaluate strategies for moving object detection are analyzed. The main challenges for these detection strategies are described in Section 4. The characteristics and content of the proposed database are discussed in Section 5. In Section 6, some outstanding methods for detecting moving objects are evaluated using the proposed dataset. Finally, Section 7 summarizes the main conclusions.

2. State of the art

Some of the most relevant databases proposed over the past years to evaluate the performance of very different types of video and image processing applications have been reviewed. The main characteristics of these databases are summarized in Table 1 (sorted in order of publication). According to the criteria considered (e.g., the nature of its content or the type of applications they evaluate), these databases can be classified in the ways described below.

2.1. According to the nature of their content

Image databases: Are formed by sets of independent images. Some examples of these databases are the ImageNet database (Deng et al., 2009) and the Berkeley database (Martin et al., 2001). The first one can be used for several tasks (e.g., object recognition, or object clustering). The second one allows the evaluation of image segmentation algorithms.

Video databases: Are formed by sets of sequences of images. For example, BEHAVE dataset (Blunsden and Fisher, 2010) (analysis of multi-person behavioral interactions) and HMDB database (Kuehne et al., 2011) (human motion recognition).

Multicamera databases: Are formed by sets of images or video sequences recorded simultaneously with multiple cameras and showing the same scenario from different points of view. Some examples of these databases are MuHaVi (Singh et al., 2010) and Videoweb (Denina et al., 2011).

3D databases: These databases contain images with depth information obtained from time of flight cameras (Lai et al., 2011), laser scanners (Spinello and Arras, 2011; Spinello et al., 2010), or stereoscopic images (Goldmann et al., 2010; Urvoy et al., 2012).

2.2. According to the way in which they have been created

Real databases: Their images or videos have been obtained from the real world (they do not contain virtual elements). Some examples of these databases are SCface (Grgic et al., 2011), which contains images to test face recognition algorithms, or VIRAT (Oh et al., 2011), which allows to assess the performance of event recognition methods.

Virtual databases: Their images or videos have been created synthetically (partial or completely). The main advantage of these datasets is that they do not require to manually annotate or segment their images, since this information is defined before generating the images or the videos (Brutzer et al., 2011). However, the way in which these databases are created makes it difficult to simulate certain circumstances that are common in real (non-virtual) content, such as shadows and highlights cast by moving objects, illumination changes, or realistic dynamic backgrounds. Examples of virtual databases are AICD (Bourdis et al., 2011), OVVV (Taylor et al., 2007), and ViHaSi (Ragheb et al., 2008).

2.3. According to their annotations

With bounding boxes: In addition to the videos, these databases contain a bounding box (or an ortoedro for 3D data) associated to each moving object. Some examples are CAVIAR (Fisher et al., 2005), Caltech pedestrian dataset (Dollár et al., 2009), and BEHAVE (Blunsden and Fisher, 2010).

With pixel-level masks: Some databases include pixel-wise foreground labels that indicate an status for each image pixel (e.g., background, moving object, shadow, etc.). Although these databases offer a more detailed information than those using bounding boxes, they are the less common in the literature. This is because obtaining pixel-level masks is a labor-intensive task. Some examples are SABS (Brutzer et al., 2011), Wallflower (Toyama et al., 1999), ChangeDetection (Goyette et al., 2012), and STAR (Li et al., 2004).

With other types of annotations: In some databases, such as i LIDS Team (2006), some representative events in the videos appear labeled (e.g., a person entering into a room). Other databases, such as medical ones (Müller et al., 2004), include text with the name of the diseases in the images.

2.4. According to the characteristics of the scenarios

Indoor: Images or videos captured in indoor scenarios (e.g., CAVIAR (Fisher et al., 2005) and CITIC (Fernandez-Sanchez et al., 2013a; 2013b)).

Outdoor: Images or videos captured in outdoor scenarios (e.g., VIRAT Video Dataset (Oh et al., 2011)).

Indoor/Outdoor: Images or videos captured in both indoor and outdoor scenarios (e.g., ChangeDetection (Goyette et al., 2012) or VPU-Lab database (García-Martín et al., 2012)).

2.5. According to the type of application to evaluate

The databases can be also classified depending on the type of application for which they have been developed. In contrast to the above described classifications, this one results in a wide variety

[1] http://www.gti.ssr.upm.es/data/LASIESTA

Table 1
Main databases proposed over the last years.

Name	Nature of content	Type of content	Annotations	Full annotated	Indoor/ outdoor	Application
Hull (1994)	Images	Real	Other	–	–	Text recognition
Visible Human 1998 (Spitzer and Whitlock, 1998)	Images	Virtual	Other	Yes	–	Biomedical
Wallflower 1999 (Toyama et al., 1999)	Videos	Real	Masks	No	Both	Moving object detection
CMU-Pittsburgh 2000 (Kanade et al., 2000)	Videos/multicamera	Real	Other	–	Indoor	Facial recognition
Berkeley 2001 (Martin et al., 2001)	Images	Real	Masks	Yes	Both	Semantic classification
STAR 2004 (Li et al., 2004)	Videos	Real	Masks	No	Both	Moving object detection
Medical 2004 (Müller et al., 2004)	Images	Real	Other	–	–	Biomedical
KTH Action 2004 (Schuldt et al., 2004)	Videos	Real	Other	–	Both	Gesture recognition
CAVIAR 2005 (Fisher et al., 2005)	Videos	Real	Bounding boxes	No	Indoor	Tracking/activity recognition
i LIDS Team (2006)	Videos/multicamera	Real	Bounding boxes	No	Both	Tracking/event detection
HumanEva 2006 (Sigal and Black, 2006)	Videos/multicamera	Real	Other	–	Indoor	Gesture recognition
PETS 2000–2006 PET	Videos/multicamera	Real	Bounding boxes	No	Both	Tracking/event detection
OVVV 2007 (Taylor et al., 2007)	Videos/multicamera	Virtual	Masks	Yes	Outdoor	Moving object detection/tracking
ETISEO 2007 (Nghiem et al., 2007)	Videos	Real	Bounding boxes	Yes	Both	Tracking/event detection
Microsoft Cambridge 2007 (Yin et al., 2007)	Videos	Real	Masks	No	Indoor	Moving object detection
ViHaSi 2008 (Ragheb et al., 2008)	Videos/multicamera	Virtual	Other	–	Indoor	Human activity recognition
ImageNet 2009 (Deng et al., 2009)	Images	Real	Other	–	Both	Semantic classification
Caltech 2009 (Dollár et al., 2009)	Videos	Real	Bounding boxes	Yes	Outdoor	Tracking
Comprehensive 2010 (Goldmann et al., 2010)	3D	Real	Other	–	Indoor	Video quality assessment
BEHAVE 2010 (Blunsden and Fisher, 2010)	Videos	Real	Bounding boxes	Yes	Both	Human activity recognition
MuHaVi 2010 (Singh et al., 2010)	Videos/multicamera	Real	Other	–	Indoor	Human activity recognition
VOC 2010 (Everingham et al., 2010)	Images	Real	Bounding boxes	Yes	Both	Semantic classification
TRECVID 2010 (Over et al., 2011)	Videos	Real	Other	–	Both	Event detection
RGB-D object 2011 (Lai et al., 2011)	3D	Real	Other	–	Indoor	Semantic classification
AICD 2011 (Bourdis et al., 2011)	Images	Virtual	Other	Yes	Outdoor	Moving object detection
HMDB 2011 (Kuehne et al., 2011)	Videos	Real	Other	–	Both	Human activity recognition
Videoweb 2011 (Denina et al., 2011)	Videos/multicamera	Real	Other	–	Outdoor	Human activity recognition
Scface 2011 (Grgic et al., 2011)	Videos/multicamera	Real	Other	–	Indoor	Face recognition
VIRAT 2011 (Oh et al., 2011)	Videos	Real	Bounding boxes	Yes	Outdoor	Tracking/Event detection
SABS 2011 (Brutzer et al., 2011)	Videos	Virtual	Masks	Yes	Outdoor	Moving object detection/tracking
CASIA 2011 (Liu et al., 2011)	Images	Real	Other	–	–	Text recognition
NAMA3DS1-COSPAD1 2012 Urvoy et al. (2012)	3D	Real	Other	–	Both	Video quality assessment
ChangeDetection 2012 (Goyette et al., 2012)	Videos	Real	Masks	No	Both	Moving object detection
VPU-Lab 2012 (García-Martín et al., 2012)	Videos	Real	Bounding boxes	Yes	Both	Tracking/people detection
CITIC 2013 (Fernandez-Sanchez et al., 2013a; 2013b)	Videos	Real	Masks	No	Indoor	Moving object detection

of possibilities. The most relevant applications are: Semantic classification (Everingham et al., 2010; Martin et al., 2001), text recognition (Hull, 1994), biomedicine (Müller et al., 2004; Spitzer and Whitlock, 1998), event detection (Desurmont et al., 2005; Kolekar et al., 2008; Oh et al., 2011; Over et al., 2011; Qian et al., 2012; Wali and Alimi, 2010), moving object detection (Goyette et al., 2012; Toyama et al., 1999; Yin et al., 2007), tracking (Dollár et al., 2009; Nghiem et al., 2007), gesture recognition (Schuldt et al., 2004; Sigal and Black, 2006), human activity recognition (Blunsden and Fisher, 2010; Fisher et al., 2005; Kuehne et al., 2011), and facial recognition (Grgic et al., 2011; Kanade et al., 2000).

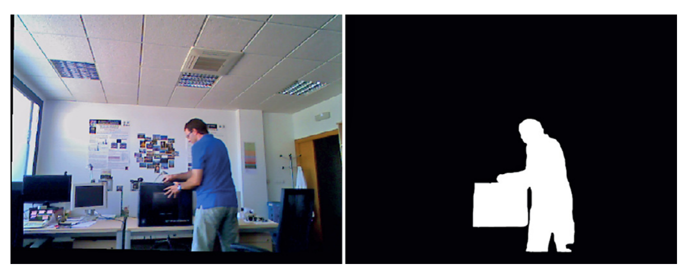

Fig. 1. Example from the CITIC dataset. Left: original image. Right: ground-truth mask.

3. Databases for evaluating motion detection strategies

This section summarizes the most relevant databases that, similarly to the proposed one, are focused to the evaluation of moving object detection algorithms (Subsection 3.1). Additionally, some other popular databases that, despite of having been mainly designed for tracking strategies, are typically used to assess the quality of motion detection approaches, are also presented (Subsection 3.2).

3.1. Databases for moving object detection strategies

To evaluate adequately the strategies for moving object detection, datasets with ground-truth masks at pixel level are required. The creation of this kind of ground-truth is a very laborious task. Consequently, the existing databases are incomplete (they only provide ground-truth masks for some images or for some regions of interest in the images) or are composed of sequences with virtual content (which makes it easier to label all the images, but hinders the representation of realistic situations). Next, some of the most relevant databases for moving object detection are described. Some of the main characteristics of these databases are shown in Table 2.

3.1.1. CITIC dataset

It consists of two sets of sequences to evaluate background subtraction algorithms based on both depth and color information. The first set (Fernandez-Sanchez et al., 2013b) contains 4 sequences recorded with stereo cameras. The second set (Fernandez-Sanchez et al., 2013a) is formed by 4 videos recorded with the Kinect sensor from Microsoft. The database also provides some binary ground-truth masks (between 3 and 7 labeled images) for each one of its sequences. Fig. 1 illustrates one of these masks and its corresponding original image.

Table 2
Databases to assess the quality of motion detection strategies.

Name	Num. sequences	Num. images	Resolution ($H \times W$)	Annotated images
CITIC 2013 (Fernandez-Sanchez et al., 2013a; 2013b)	8	4593	480×640[a]	0.89%
ChangeDetection 2012 (Goyette et al., 2012)	49	143,578	240×320[a]	100%
SABS 2011 (Brutzer et al., 2011)	9	13,409	600×800	52, 2%
Microsoft Cambridge 2007 (Yin et al., 2007)	28	6373	240×320	9, 63%
OVVV 2007 (Taylor et al., 2007)	undef	undef	undef	100%
Wallflower 1999 (Toyama et al., 1999)	7	16,158	120×160	0.04%
STAR 2004 (Li et al., 2004)	9	18,447	128×160[a]	0.98%

[a] The resolution shown is the standard, but there are some sequences with other resolutions.

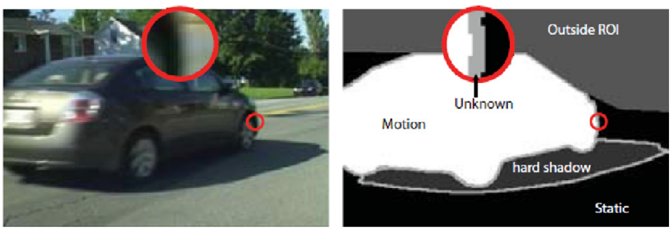

Fig. 2. Left: image from the ChangeDetection database. Right: corresponding ground-truth mask.

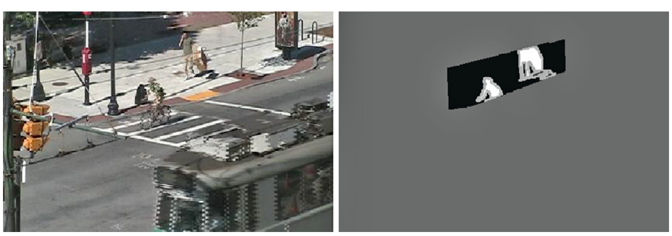

Fig. 3. Low quality image from ChangeDetection (left) and restricted ground-truth mask (right).

3.1.2. ChangeDetection

The first version of the ChangeDetection database (Goyette et al., 2012) was published in 2012. It contains 31 video sequences classified in 6 categories that address different challenges for moving object detection: basic sequences, sequences with dynamic background, sequences with camera jitter, sequences with intermittent object motion, and sequences captured with infra-red cameras. Later, in 2014, 18 new sequences distributed in 4 new categories were added.

In the ground-truth masks provided by this dataset, the pixels appear classified in 5 different labels: static (background pixels), motion (foreground pixels), unknown (pixels around the contour of the segmented moving objects, where it cannot be insured whether the pixels belong to the background or to the foreground), hard shadow (pixels belonging to shadows cast by moving objects), and outside ROI (pixels in the areas that have not been annotated). Fig. 2 shows an example of an image from this database and its corresponding ground-truth mask.

With the exception of the proposed database, the ChangeDetection database is the only one that offers lots of fully annotated real images. Consequently, in spite of its recent publication, it has been used by several authors to test their algorithms (Kim et al., 2013; Kryjak et al., 2013; Seidel et al., 2013).

However, the resolution of many of its segmented objects is too low and some of its videos are of dubious quality (they appear to have been post-processed, often with a low quality de-interlacing algorithm that leaves ghosts) (Haines and Xiang, 2014). The left image in Fig. 3 illustrates a low quality image from ChangeDetection in which the interlacing is clearly observed.

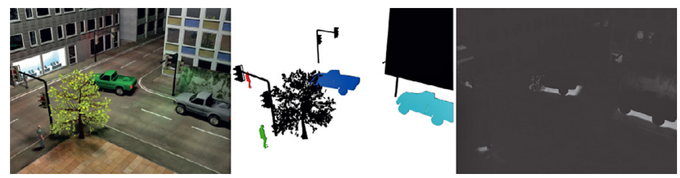

Fig. 4. Left: image from the SABS dataset. Middle: labeled ground-truth mask. Right: shadow mask.

Furthermore, the categories in this dataset do not separate adequately the typical challenges in moving object detection, which makes it difficult to test the quality of algorithms focused to specific challenges. Additionally, the sequences are unnecessarily long, which complicates separating some challenges from other. In some sequences, to try to focus on particular challenges, the analysis are restricted to very small regions that have been arbitrarily chosen (see the right image in Fig. 3). Nevertheless, the application of these restrictions is not enough to isolate many of the challenges.

3.1.3. SABS

The Stuttgart Artificial Background Subtraction (SABS) dataset (Brutzer et al., 2011) is another popular database that has been used by many authors to test their algorithms (Chang et al., 2012; Saptharishi et al., 2012; Shimada et al., 2013). It is composed of synthetic sequences to evaluate 7 different challenge situations: gradual and sudden illumination changes, dynamic background, camouflage, shadows, bootstrapping, and video noise. Each sequence contains 801 frames for training (except the bootstrapping one) and 600 frames for testing (except two of them, which have 1400 testing images).

All the testing images have a ground-truth mask associated, where each moving object has a different label. Therefore, this database not only allows to test moving object detection methods but also tracking algorithms, which require to distinguish between the different moving objects in the sequences. Additionally, masks with information about the shadows cast by the moving objects are also included in the dataset.

To emulate the noise introduced by camera sensors in real data, the sequences have been recorded by adding Gaussian noise to the pixels values. Moreover, they have been created using a raytracing technology with global illumination to try and simulate scenes with realistic lighting.

Fig. 4 presents an image from SABS dataset, its label mask, and its associated mask of shadows.

Besides the obvious disadvantage of not using real images, another major disadvantage of this database is that all the sequences show the same scenario and, moreover, all of them have been captured from the same point of view.

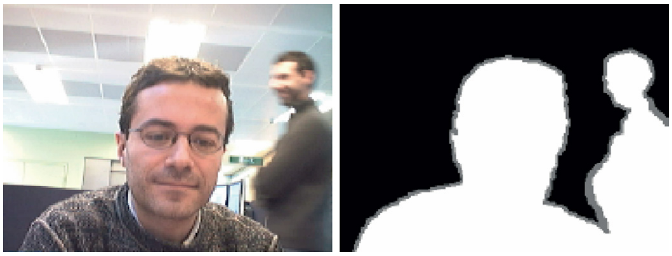

Fig. 5. Left: original image from the Microsoft Cambridge database. Right: corresponding labeled ground-truth.

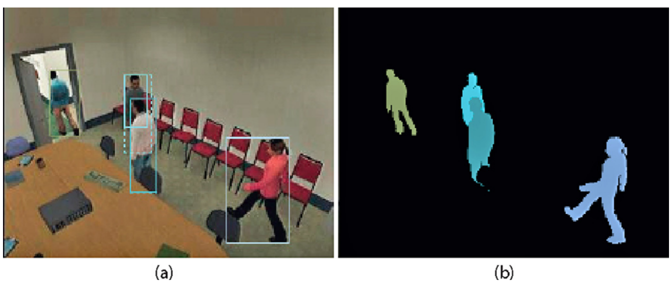

Fig. 6. Left: synthetic image with bounding boxes from OVVV. Right: corresponding pixel-wise mask.

Fig. 7. Left: original image from the Wallflower dataset. Right: associated ground-truth mask.

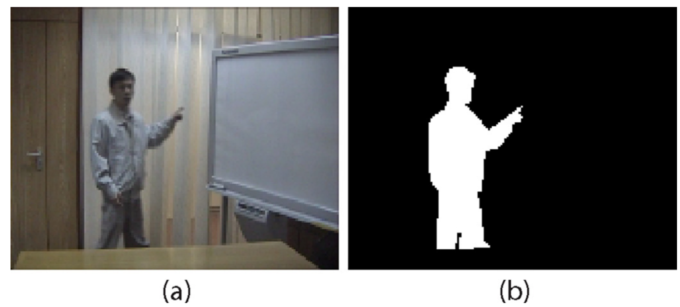

Fig. 8. Left: frame from the STAR database. Right: corresponding ground-truth mask.

3.1.4. Microsoft Cambridge

The Microsoft Cambridge database (Yin et al., 2007) is composed of 28 video-chat type sequences. These sequences have been pixel-wise labeled every fifth or tenth frame into three possible states: foreground, background and uncertainty (areas where it is difficult to discriminate between foreground and background). Fig. 5 shows one of its images and its corresponding ground-truth mask.

3.1.5. OVVV

The ObjectVideo Virtual Video (OVVV) dataset (Taylor et al., 2007) simulates several synchronized video streams from many camera configurations (e.g., static, PTZ, or omni-directional). It provides different types of ground-truth for each frame: target centroids, bounding boxes, and pixel-wise moving object masks. One image from this database and its corresponding ground-truth are shown in Fig. 6.

3.1.6. Wallflower

The Wallflower database (Toyama et al., 1999), is one of the first datasets focused to evaluate moving object detection strategies. Consequently, from its emergence, it has been extensively used by several authors (Cuevas and García, 2013; Maddalena and Petrosino, 2008).

It is composed of 7 sequences with a total of 16,158 images that address some potentially problematic situations in moving object detection: background subtraction, camouflage, bootstrapping, object subtracted from the background, and illumination changes.

Although this database includes sequences of great interest to evaluate moving object detection algorithms, its major drawback is that it only provides one binary mask per sequence to indicate what pixels belong to the moving object or to the background. One of these ground-truth masks and its corresponding original image are depicted in Fig. 7.

3.1.7. STAR

The STAR database (Li et al., 2004) is very similar to the Wallflower database. It is composed of 9 video sequences recorded in different indoor and outdoor environments: offices, campuses, shopping malls, etc. For each sequence, only a subset of 20 images, randomly selected, has been labeled at pixel-level.

Fig. 8 depicts an image from this dataset and its corresponding ground-truth mask.

3.2. Databases for tracking strategies

The creation of databases to evaluate the performance of moving object detection methods is a labor-intensive task. However, the elaboration of databases to test tracking algorithms is much easier, since they do not require pixel-level masks but bounding boxes around the moving objects. Consequently, a larger amount of datasets for the testing of tracking algorithms can be found.

Since some of these databases contain the typical challenges for motion detection strategies, they are frequently used to assess the quality of their results. However, since these databases do not contain pixel-level annotations, they only allow providing subjective evaluations of the detection algorithms. Next, we briefly describe those tracking databases that are typically used to assess the quality of moving object detection methods.

3.2.1. VPU-Lab

VPU-Lab dataset (García-Martín et al., 2012) is composed of 90 video sequences and their corresponding ground-truth data (bounding boxes for all the moving objects). It contains both indoor and outdoor sequences with different background complexities (lighting changes, dynamic background, etc.) and a great variability of people appearance and interactions.

3.2.2. i-LIDS

The Imagery Library for Intelligent Detection Systems (i-LIDS) (i LIDS Team, 2006) contains several sequences to test 4 situations related to moving object detection (abandoned objects, parked vehicles, doorway, and restricted areas). All the sequences are supplied with XML files describing temporal events and the size and spatial position of the moving objects.

3.2.3. PETS

The Performance Evaluation of Tracking and Surveillance (PETS) program (PETS, 2000-2006) collects videos for evaluating surveillance algorithms from 2000 to 2006. Most of these videos include bounding boxes as ground-truth data.

4. Database design criteria

To be complete, the proposed database must contain sequences to allow the evaluation of the main challenges in moving object detection. An overview of these challenges is given in this section.

4.1. Shadows and highlights

Shadows and highlights cast by moving objects are usually detected as part of such moving objects, drastically reducing the quality of the results of many detection algorithms (Xu et al., 2006). Consequently, several proposals to remove shadows and highlights from the detections have been proposed along the last years (Amato et al., 2014; Horprasert et al., 1999).

This problem not only appears in outdoor sequences where shadows are identified at first sight but also in most indoor sequences where moving objects continuously occlude and reflect light and, consequently, produce medium and soft shadows and highlights.

4.2. Dynamic background

Many sequences contain background elements that are not completely static but move periodically or irregularly (e.g., waving flags, trees and shrubs shaken by the wind, escalators, or water waves). However, these moving elements must be classified as part of the background (Sheikh and Shah, 2005).

4.3. Illumination changes

Moving object detection methods must be able to adapt their background models to gradual illumination changes (e.g., the light variations along the day) and to sudden illumination changes (e.g., turning on the lights in a room) (Bouwmans et al., 2014).

4.4. Persistent background changes

Another issue to consider is that related to sequences in which a persistent change occurs (Evangelio and Sikora, 2011): a moving object becomes part of the background (e.g., an abandoned object) or a background element is changed (e.g., a door is opened or someone moves a static object).

4.5. Camouflage

Camouflage situations occur when a moving object and the background behind it have similar color. In this situations, it is not easy to distinguish between background and foreground. Occasionally, when the camouflage is very intense, a moving object can be separated from the background only if its shape is previously known.

4.6. Occlusions

An occlusion is the result of an overlapping between two moving objects, or a situation in which part of the background is placed in front of a moving object. Some recent detection algorithms (Creusot, 2014; Cuevas et al., 2010; Gallego et al., 2012) use tracking methods to improve their results. For these methods, the presence of totally or partially occluded moving objects is a typical challenge (Zhang et al., 2013).

4.7. Displacements along the optical axis of the camera

Sometimes the sequences contain moving objects displacing along the optical axis of the camera. In these situations, large amounts of pixels belonging to such moving objects barely change their appearance throughout many frames. This results in the wrong classification of these moving objects as part of the background.

4.8. Stationary foreground objects

Another challenge to consider is that related to situations in which a moving object remains static a few seconds and then moves again. The decisions to make in these situations are very complicated. It can be considered that the stationary moving object continuous being a moving object. But it can also be considered that if it remains static too long, it should be classified as background. In this last case, it must also be decided at what time (after it stopped) it becomes part of the background. Many strategies for detecting stationary foreground objects have been proposed over the recent years (Baxter et al., 2015; Lin et al., 2015; Maddalena and Petrosino, 2013).

4.9. Bootstrapping

Traditionally, it is typical to initialize the detection algorithms with a set of images free from moving objects. However, in many current applications this initialization period is not available and the bootstrap strategies must be used to provide successful results (Hsiao and Leou, 2013).

4.10. Moving camera

The sequences to analyze could have been recorded with non-stabilized cameras (e.g., a camera endowed in a mobile phone) or with moving cameras (e.g., with pan or tilt motions). Therefore, motion detection algorithms that try to provide successful results in these situations have been also proposed by many authors (Kong et al., 2010; Sheikh et al., 2009; Ye et al., 2011).

5. LASIESTA

The LASIESTA database is composed of 48 sequences divided in two sets: indoor sequences (described in Section 5.1) and outdoor sequences (described in Section 5.2). Each set consists of different categories, each of which addresses a specific challenge. Each category has two different sequences to test the corresponding challenge, except the categories corresponding to the simulated camera motion (described in Section 5.3), which contains many more sequences.

The main characteristics of the proposed database are the following:

- It is complete: It contains sequences related to all the challenges described in Section 4. Moreover, all the images are supplied with ground-truth masks labeled at both pixel and object levels. The ground-truth data have been generated using the TSLAB (Cuevas et al., 2015), which is a novel tool that is able to provide very high-quality labels with the help of an intuitive graphical interface and some high-level analysis options (e.g. background subtraction methods and active contour-based algorithms).
- It is well structured: The challenges are well separated and, as far as possible, they are not repeated. Thus, every challenge can be addressed independently of other challenges.

- It is varied: The sequences have been recorded in multiple indoor and outdoor scenarios with very different content (e.g. halls, roads, corridors, cars, etc.) and climatology (rain, sun, clouds and snow).
- It is compact: The length of the sequences is just enough to assess the challenges. In this way, LASIESTA allows to carry out a fast evaluation of moving object detection algorithms, unlike many other databases (e.g. Goyette et al., 2012 or Li et al., 2004), which are composed of excessively long sequences. Furthermore, each sequence contains the required minimum number of moving objects. It must be noted that for the evaluation of most of the typical challenges in moving object detection (see Section 4) it is not necessary to use a high amount of moving objects, but it is generally sufficient with a single moving object. The sequences with multiple simultaneous moving objects are only necessary to analyze some specific cases, such as the partial or total occlusions between moving objects. On the other hand, the cost involved in labeling the sequences increases very significantly with the number of moving objects. For all these reasons, similarly to many other databases for assessing the quality of moving object detection methods (Fernandez-Sanchez et al., 2013a; 2013b; Li et al., 2004; Toyama et al., 1999; Yin et al., 2007), the sequences in LASIESTA contain a reduced amount of moving objects (up to 3 per sequence).

Another important contribution of the proposed database is that it contains some sequences where a moving object remains static permanently (e.g., an abandoned object) or along a short period of time (e.g., a person stops and, a while after, it continues moving). In these situations, in contrast to previous databases, the stopped moving objects are marked (at pixel level) with a specific label. Moreover, additional data are also provided to indicate the temporal interval over which these moving objects remain stopped. This allows each user of the database to decide at what time such objects become part of the background: from just the moment they stop moving, after remaining static for an arbitrary number of images, or never.

In addition, similarly to other datasets (e.g. ChangeDetection), we have defined a uncertainty zone around the contour of the labeled objects where, due to possible inaccuracies in the labeling process, it is difficult to determine if the pixels belong to the moving objects or to the background.

5.1. Indoor sequences

Table 3 shows the main characteristics of the indoor sequences in the LASIESTA database and Fig. 9 illustrates, for each one of these sequences, one image and its corresponding ground-truth mask.

The set of indoor sequences is formed by 7 categories focused in different particular challenges. The main features of these categories are the following:

- Simple sequences (I_SI): Sequences not containing camouflage, occlusions, illumination changes, persistent background changes, bootstrapping, or camera motion.
- Camouflage (I_CA): The sequences in this category contain moving objects with appearance similar to some background areas. Additionally, these moving objects remain static on such background regions along short periods of time.
- Occlusions (I_OC): It is formed by sequences with moving objects that, along their trajectories across the scene, appear partially or totally occluded.
- Illumination changes (I_IL): It contains sequences with gradual and sudden illumination changes.
- Modified background (I_MB): Sequences showing situations in which some background elements are subtracted or abandoned.
- Bootstrap (I_BS): It is formed by sequences containing moving objects from the first frame.
- Moving camera (I_MC): It contains sequences recorded with moving cameras (camera jitter and pan motion).

5.2. Outdoor sequences

Table 4 presents the main characteristics of the set of outdoor sequences in the LASIESTA database. Fig. 10 shows, for each one of these sequences, one image and its corresponding ground-truth mask.

Similarly to the set of indoor sequences, the outdoor sequences are organized in different categories focused in different particular challenges. The main features of the 5 categories in the set of outdoor videos are the following:

- Cloudy conditions (O_CL): It is formed by two sequences with cloudy weather, recorded in different scenarios. These sequences contain different types of moving objects (a car in the first one and two people in the second one) and dynamic backgrounds (trees and shrubs moved by the wind).
- Rainy conditions (O_RA): The sequences in this category have a very dynamic background due to the rain. Additionally, the trees and shrubs in their backgrounds are not completely static.
- Snowy conditions (O_SN): The two sequences in this category have been recorded in scenarios where it is snowing. Therefore, they contain very dynamic backgrounds.
- Sunny conditions (O_SU): Similarly to the previous categories, this one also contains dynamic background elements. However, unlike the rest of sequences, these sequences have been recorded in sunny scenarios to obtain hard shadows cast by the moving objects.
- Moving camera (O_MC): This category contains two sequences recorded with moving cameras (camera jitter and pan/tilt motion).

5.3. Sequences with simulated camera motion

The recent explosion of portable electronic devices with camera (e.g., cellular phones, vehicles, or robots) has resulted in an increasing interest in algorithms able to handle apparent displacements and motion blur due to sensor movements. Consequently, several strategies proposing methods to efficiently detect and track moving objects in this kind of sequences can be found (Ren et al., 2003; Rowe and Blake, 1996; Sheikh et al., 2009; Yi et al., 2013).

However, to our knowledge, there is not any adequate database for the evaluation of the results provided by these strategies. For this reason, we have decided to include two categories (one indoor and one outdoor) specifically focused to the evaluation of the performance of strategies for sequences recorded with moving cameras. These categories have been named respectively as I_SM and O_SM.

The assessment of a variety of camera motions (i.e. differing in both type of movement and intensity), would require recording and labeling a large amount of new sequences, which could be an unreachable labor. Therefore, in order to provide a wider set of moving camera examples and with a tightly controlled movement characteristics, we have simulated different types and intensities of motion from two single STATIC high-resolution sequences.

Tables 5 and 6 show, respectively, the main characteristics (intensities and types of motion) of the sets of indoor and outdoor sequences with simulated motion.

These sequences have been obtained by selecting, for each frame in the high-resolution sequences, the image inside a selection window (from now SW). Depending on the type of motion

Table 3
Indoor sequences in the LASIESTA database.

Identifier	Short description	Duration (s)	Number of images	Number of moving objects	Foreground pixels	Tags
I_SI_01	Three people cross a room walking perpendicularly to the optical axis of the camera.	12	300	3	4.50%	Smooth shadows
I_SI_02	One person crosses a corridor going towards the camera.	12	300	1	1.57%	Smooth shadows
I_CA_01	A person crosses a room and remains static for a few seconds in front of a door with similar color to his clothing.	14	350	1	5.64%	Stationary foreground object, smooth shadows, camouflage
I_CA_02	A person appears and stands in front of a wall with similar color to his T-shirt. A plant is constantly moving in the background.	21	525	1	7.67%	Stationary foreground object, camouflage, dynamic background
I_OC_01	A person crosses a room and passes behind a large column.	10	250	1	0.67%	Moderate shadows, total occlusion
I_OC_02	One person goes down some stairs and walks behind a railing.	10	250	1	3.46%	Smooth shadows, partial occlusion
I_IL_01	A person who is going through a room turns on the lights.	12	300	1	1.92%	Smooth shadows, illumination change
I_IL_02	One person walks towards a window and opens the blinds. Then the person walks out of the scene.	21	525	1	2.78%	Hard shadows, illumination change, permanent changes in the background
I_MB_01	One person enters a room with a bag on his shoulder, leaves the bag on the ground, and goes out of the scene.	18	450	2	6.65%	Smooth shadows, abandoned object
I_MB_02	One person enters a room, picks up a bag that is on the ground, and goes out.	14	350	1	4.26%	Moderate shadows, permanent changes in the background
I_BS_01	Two people shake their hands in a corridor and walk in opposite directions, entering two different rooms.	11	275	2	1.88%	Bootstrap, moderate shadows
I_BS_02	A person stands shortly in front of a sign before walking out of the scene.	11	275	1	0.73%	Bootstrap, moderate shadows
I_MC_01	One person climbs some steps in a hall with columns. The camera sweeps the scene.	12	300	1	3.36%	Camera motion, smooth shadows
I_MC_02	A person crosses a room. The camera is not completely static (soft jitter).	10	250	1	1.36%	Camera motion, moderate shadows

Table 4
Outdoor sequences in the LASIESTA database.

Identifier	Short description	Duration (s)	Number of images	Number of moving objects	Foreground pixels	Tags
O_CL_01	A car crosses a parking and passes behind a lamppost. The wind moves the vegetation and clouds are reflected in parked cars' windows.	9	225	1	3.22%	Dynamic background, smooth shadows, partial occlusion
O_CL_02	Two people go up a street, one immediately after another. The wind moves the vegetation.	17	425	2	10.9%	Dynamic background, smooth shadows
O_RA_01	In rainy conditions, a person walks in a parking. After a while, a parked car starts moving and leaves the scene.	56	1400	2	0.28%	Dynamic background, moderate shadows, rain, partial occlusion, permanent changes in the background.
O_RA_02	In rainy conditions, two people (one after another) go walking in front of a garden.	15	375	2	1.49%	Dynamic background, moderate shadows, rain
O_SN_01	It's snowing. A car crosses a parking and disappears under a gate.	20	500	1	1.17%	Dynamic background, snow, smooth shadows
O_SN_02	It is snowing. A car traveling very slowly appears in the bottom right corner of the scene.	34	850	1	0.31%	Dynamic background, snow, smooth shadows, partial occlusion
O_SU_01	A person crosses a street in a sunny area and a car moves in the background of the scene. The silhouette of the person is reflected in a puddle.	10	250	2	0.64%	Dynamic background, camouflage, hard shadows
O_SU_02	A person crosses a street in a sunny area. Later, another person crosses in opposite direction in a shaded area.	16	400	2	1.12%	Dynamic background, hard shadows
O_MC_01	A person crosses a parking and the camera sweeps the scene.	17	425	1	0.97%	Smooth shadows, camera motion
O_MC_02	A person crosses a parking. The camera is not completely static (soft jitter).	7	175	1	1.80%	Dynamic background, smooth shadows, camera motion

Fig. 9. Examples (original images and labeled masks) of the indoor sequences in the LASIESTA database.

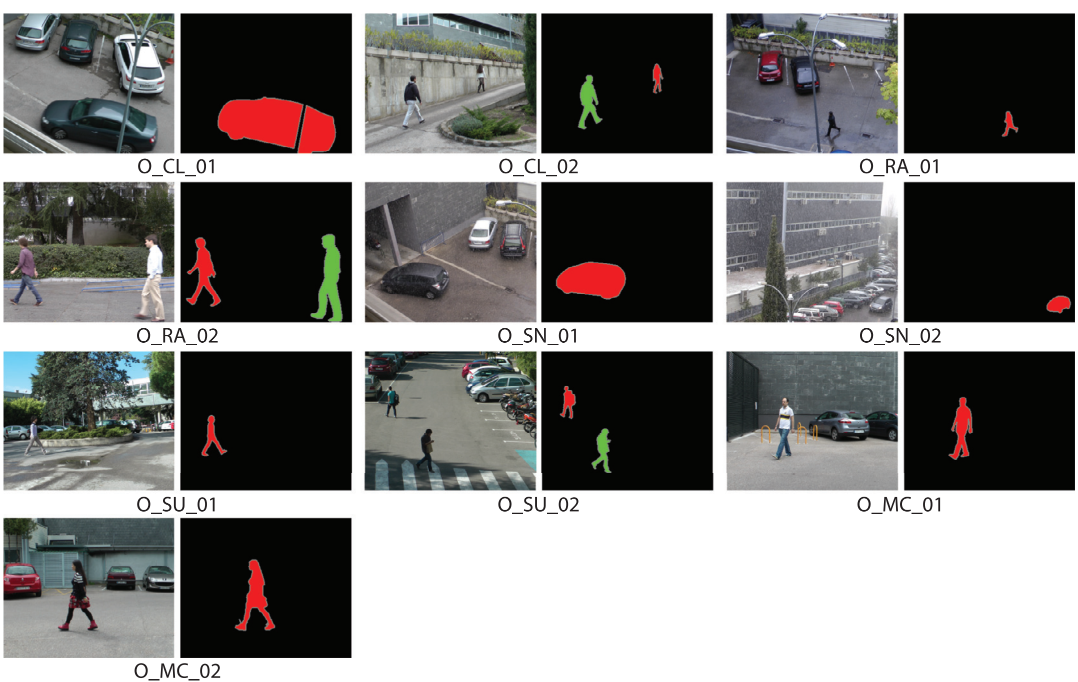

Fig. 10. Examples (original images and labeled masks) of the outdoor sequences in the LASIESTA database.

to be simulated, the SW is subjected, frame to frame, to a displacement, to a rotation, or to both. Two original images from the indoor and outdoor high-resolution sequences are shown, respectively, in Figs. 11 and 12. Additionally, in each one of these figures, three images corresponding to different intensities of simulated motion are also illustrated. The color rectangles on the high-resolution images show the SWs from which such trios of images have been obtained.

As it can be observed in the images of these two figures, the ground-truth images corresponding to the sequences with simu-

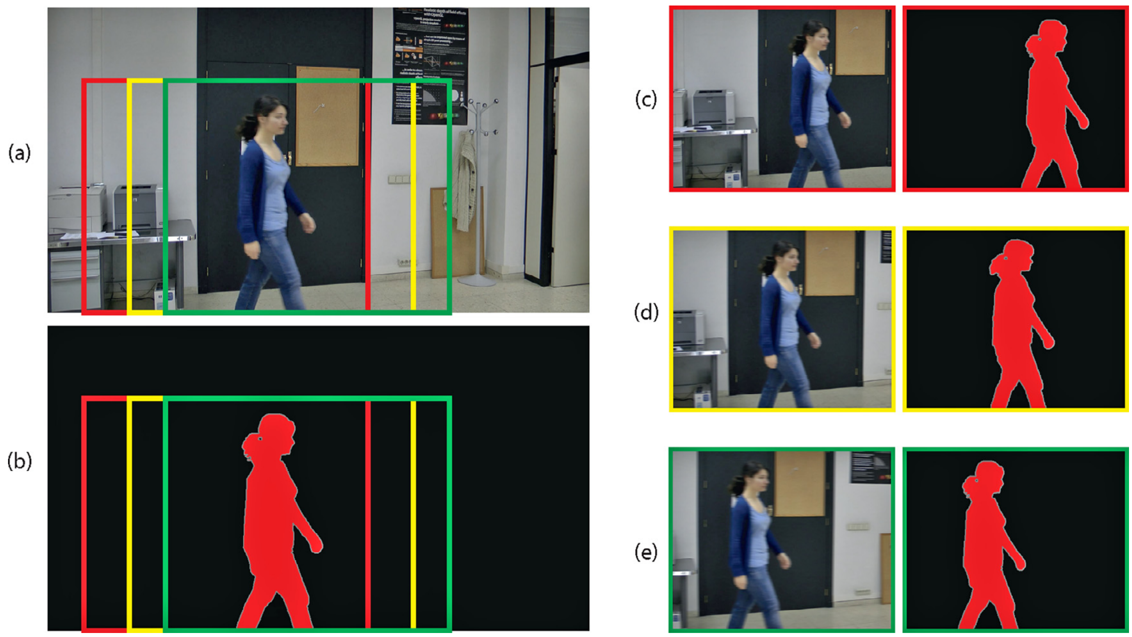

Fig. 11. One original image of the high-resolution sequence used as basis to obtain the indoor sequences with simulated motion (a), its corresponding ground-truth image (b), and three pairs of images (original and ground-truth) corresponding to sequences with different intensities of simulated motion: (c) low intensity, (d) medium intensity, and (e) high intensity. The color rectangles illustrate the areas of the high-resolution images that have been used to obtain the images with simulated motion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

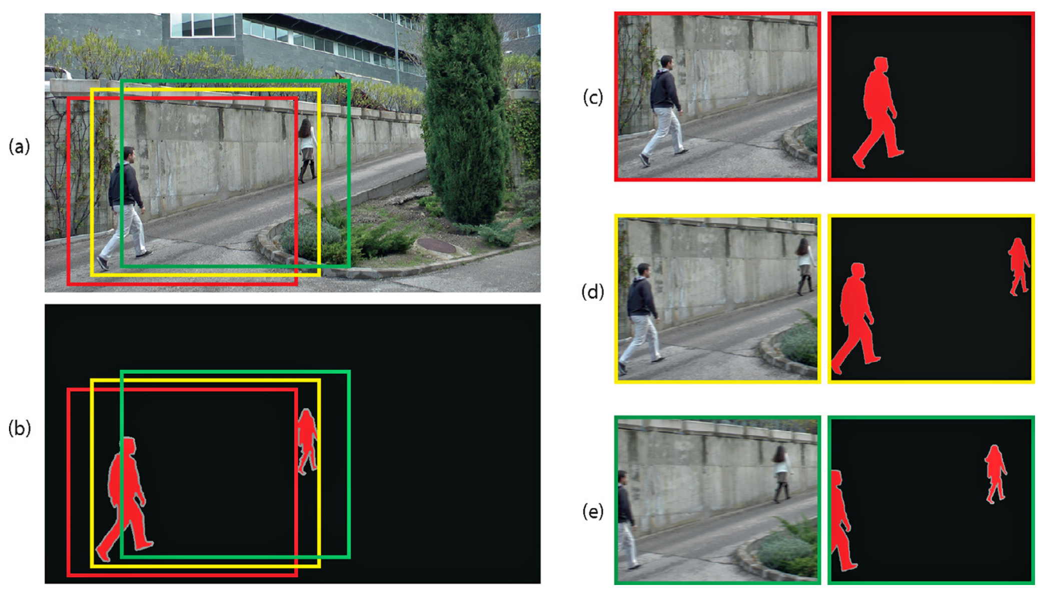

Fig. 12. One original image of the high-resolution sequence used as basis to obtain the outdoor sequences with simulated motion (a), its corresponding ground-truth image (b), and three pairs of images (original and ground-truth) corresponding to sequences with different intensities of simulated motion: (c) low intensity, (d) medium intensity, and (e) high intensity. The color rectangles illustrate the areas of the high-resolution images that have been used to obtain the images with simulated motion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lated motion are also obtained by selecting the data inside the SWs. In this way, we avoid the laborious task of creating new ground-truth data for each new sequence.

5.3.1. Simulation of translational motion

To simulate the translation movement, the SW has been uniformly displaced, frame to frame, according to a 2-dimensional motion vector $\mathbf{v} = (v_1, v_2)$, where v_1 and v_2 denote, respectively, the vertical and horizontal displacements in pixels.

In the case of the indoor sequences we have used: $\mathbf{v} = (0, 4)$ for the low intensity motion, $\mathbf{v} = (0, 8)$ for the medium intensity motion and $\mathbf{v} = (0, 12)$ for the high intensity motion.

To create the outdoor sequences we have used: $\mathbf{v} = (1, 4)$ for the low intensity motion, $\mathbf{v} = (2, 8)$ for the medium intensity motion and $\mathbf{v} = (3, 12)$ for the high intensity motion.

5.3.2. Simulation of camera jitter and rotational motion

To simulate the camera jitter and the rotational motion, at each new frame, the SW has been randomly rotated and randomly displaced (in both horizontal and vertical directions). The rotation angles have been (randomly) obtained from the Gaussian distribution $N(0, \sigma_r)$, whereas the horizontal and vertical displacements have been (randomly) obtained from the Gaussian distribution $N(0, \sigma_d)$.

Table 5
Indoor sequences with simulated motion.

Identifier	Motion	
	Type	Intensity
I_SM_01	Pan	Low
I_SM_02	Pan	Medium
I_SM_03	Pan	High
I_SM_04	Jitter/rotation	Low/low
I_SM_05	Jitter/rotation	Low/medium
I_SM_06	Jitter/rotation	Low/high
I_SM_07	Jitter/rotation	Medium/low
I_SM_08	Jitter/rotation	Medium/medium
I_SM_09	Jitter/rotation	Medium/high
I_SM_10	Jitter/rotation	High/low
I_SM_11	Jitter/rotation	High/medium
I_SM_12	Jitter/rotation	High/high

Table 6
Indoor sequences with simulated motion.

Identifier	Motion	
	Type	Intensity
O_SM_01	Pan/tilt	Low
O_SM_02	Pan/tilt	Medium
O_SM_03	Pan/tilt	High
O_SM_04	Jitter/rotation	Low/low
O_SM_05	Jitter/rotation	Low/medium
O_SM_06	Jitter/rotation	Low/high
O_SM_07	Jitter/rotation	Medium/low
O_SM_08	Jitter/rotation	Medium/medium
O_SM_09	Jitter/rotation	Medium/high
O_SM_10	Jitter/rotation	High/low
O_SM_11	Jitter/rotation	High/medium
O_SM_12	Jitter/rotation	High/high

The standard deviations of these distributions have been set according to the intensity of the simulated motion:

- Low intensity: $\sigma_r = 0.1$ degrees and $\sigma_d = 1$ pixel.
- Medium intensity: $\sigma_r = 0.5$ degrees and $\sigma_d = 4$ pixels.
- High intensity: $\sigma_r = 0.9$ degrees and $\sigma_d = 7$ pixels.

5.3.3. Blur simulation

To approximate as closely as possible to the real moving camera sequences features, the simulated sequences include simulated motion blur (corresponding to the simulated motion). This blurring has been performed by using low-pass filters oriented along the direction of the simulated motion and with a length equal to the amount of motion between consecutive frames. The trios of images in Figs. 11 and 12 allow observing that the amount of added blur is proportional to the intensity of the simulated motion. That is, the images corresponding to sequences with low intensity motion have less blur than the images corresponding to the sequences with the high intensity motion.

It must be noted that the blurring has been applied exclusively to the original images but not to the ground-truth images: the areas of uncertainty have not been increased depending on the amount of simulated motion. This is because the areas of uncertainty have been created exclusively to deal with the inaccuracies in the labeling process and, obviously, these inaccuracies will remain the same regardless of the performed motion simulation.

6. Results

The great effort devoted along the last decades to address moving object detection has resulted in hundreds of proposals (Bouwmans, 2014). These proposals have evolved from methods based on simple frame differencing to methods based on statistical analysis, which are able to automatically capture the representative statistics of the scene and to adapt the detection process to avoid misdetections due to uninteresting pixel variations. To test the usefulness of the proposed database, we have selected the following seven outstanding methods:

- Wren et al. (1997): This strategy, commonly known as Running Gaussian Average (RGA), tries fitting a Gaussian probabilistic density function on the most recent values of each pixel. Because of its high usability (it depends on few parameters), its low computational cost and its successful detections in many common scenarios, it is one of the most popular moving object detection strategies and it has been used in many recent video analysis strategies (Li et al., 2009; Lingam and Kumar, 2014; Mirmahboub et al., 2013). Moreover, a recent survey (Sobral and Vacavant, 2014) has proven that, for some datasets, the RGA is among the top five detection strategies.
- Stauffer and Grimson (2000): To deal with scenarios with complex and dynamic backgrounds, this strategy uses Gaussian Mixture Models (GMMs) to try to represent the variations of the pixels along the time. Because of its high-quality results in a huge amount of situations, it has been taken as starting point by hundreds of authors (Bouwmans et al., 2008) and it has thousands of citations.
- Zivkovic and van der Heijden (2006): It is an improved version of the GMM algorithm in Stauffer and Grimson (2000). This strategy, from a Bayesian perspective, applies a criterion selection to automatically determine the adequate number of components for each pixel according to the characteristics of the scene. This method provides adequate results in multiple scenarios and, additionally, is fast and easy to use. Consequently, it is one of the most used and has been recently included in the OpenCV (Open Source Computer Vision) library (Intel Corporation, 2015), which is probably the most popular C++ library of programming functions aimed to computer vision.
- Maddalena and Petrosino (2008): This method is based on Artificial Neural Networks (ANN) and it is able to handle scenes containing moving backgrounds, gradual illumination variations, camouflage and bootstrapping without prior knowledge about the involved patterns. Additionally, it allows including into the background model the shadows cast by moving objects.
- Maddalena and Petrosino (2012): It is an improved version of the ANN-based algorithm in Maddalena and Petrosino (2008) that, by the introduction of spatial coherence into the background update procedure, is able to provide further robustness against false detections.
- Cuevas and García (2013): It is a nonparametric Kernel Density Estimation (KDE) method that is able to improve the quality of the results provided by parametric strategies (e.g. RGA or GMM) in environments where pixel variations cannot be described parametrically. Thanks to a selectively updated spatio-temporal background model and a shadow removal method, the strategy in Cuevas and García (2013) has shown to be able to significantly improve both the usability and the quality of previous nonparametric approaches (Cuevas and Garcia, 2013; Elgammal et al., 2000; Sheikh and Shah, 2005).
- Haines and Xiang (2014): This strategy is based on the use of Dirichlet process Gaussian mixture models to estimate per-pixel background distributions. This recent strategy has shown to be able of improving the results of many previous methods in a very large amount of video datasets. Moreover, its usability is very high (barely requires adjusting its parameters) and it is able to run at real-time.

The quality of the detections provided by these seven strategies has been measured using the harmonic mean of the recall ($r = \frac{cd}{cd+md}$) and precision ($p = \frac{cd}{cd+fd}$) or *F*-score ($F = 2\frac{r \cdot p}{r+p}$), where

Table 7
F-score values obtained in the LASIESTA database. Algorithm ranks are given by the numbers in brackets. The last column contains the average for all the tests.

Method	I_SI_01	I_SI_02	I_CA_01	I_CA_02	I_OC_01	I_OC_02	I_IL_01	I_IL_02	I_MB_01
Wren et al. (1997)	0.8348 (6)	0.8016 (6)	0.8012 (7)	0.7123 (6)	0.9512 (4)	0.8235 (7)	0.5191 (4)	0.4568 (3)	0.8084 (6)
Stauffer and Grimson (2000)	0.8409 (5)	0.8247 (4)	0.8811 (4)	0.7733 (3)	0.9500 (5)	0.8271 (6)	0.3498 (5)	0.2392 (6)	0.8342 (5)
Zivkovic and van der Heijden (2006)	0.9118 (3)	0.8990 (2)	0.9094 (3)	0.7546 (4)	**0.9880 (1)**	0.9135 (3)	0.1648 (7)	0.3135 (5)	0.9321 (3)
Maddalena and Petrosino (2008)	0.8928 (4)	0.8465 (3)	**0.9532 (1)**	0.7394 (5)	0.9803 (2)	0.8466 (5)	0.8533 (2)	0.3750 (4)	0.8473 (4)
Maddalena and Petrosino (2012)	0.9559 (2)	**0.9409 (1)**	0.8416 (6)	**0.8731 (1)**	0.9573 (3)	0.9508 (2)	0.1898 (6)	0.2312 (7)	0.9728 (2)
Cuevas and García (2013)	0.8143 (7)	0.7576 (7)	0.8424 (5)	0.6296 (7)	0.8274 (7)	0.8781 (4)	0.7966 (3)	0.7864 (2)	0.7779 (7)
Haines and Xiang (2014)	**0.9622 (1)**	0.8130 (5)	0.9220 (2)	0.8656 (2)	0.8920 (6)	**0.9526 (1)**	**0.8861 (1)**	**0.8122 (1)**	**0.9816 (1)**

Method	I_MB_02	I_BS_01	I_BS_02	O_CL_01	O_CL_02	O_RA_01	O_RA_02	O_SN_01	O_SN_02	O_SU_01	O_SU_02	Average
Wren et al. (1997)	0.6722 (7)	0.4622 (4)	0.4893 (4)	0.8724 (6)	0.8474 (4)	**0.8868 (1)**	0.8248 (7)	0.7418 (3)	0.4515 (4)	0.6808 (4)	0.8304 (5)	0.7234 (5)
Stauffer and Grimson (2000)	0.6940 (4)	0.3555 (7)	0.3694 (7)	0.8922 (5)	0.8457 (6)	0.7435 (7)	0.8275 (6)	0.7369 (4)	0.4724 (2)	0.6177 (6)	0.8304 (6)	0.6953 (7)
Zivkovic and van der Heijden (2006)	0.8015 (2)	0.5472 (2)	0.5195 (3)	0.9303 (2)	0.8226 (7)	0.8586 (2)	0.8980 (3)	0.5206 (6)	0.2402 (5)	0.5426 (7)	0.8775 (3)	0.7173 (6)
Maddalena and Petrosino (2008)	0.6761 (6)	0.4023 (5)	0.4465 (5)	0.8985 (4)	0.8547 (5)	0.8252 (4)	0.8588 (5)	0.6977 (5)	0.4595 (3)	0.7467 (3)	0.8562 (4)	0.7528 (3)
Maddalena and Petrosino (2012)	**0.8517 (1)**	0.4015 (6)	0.4021 (6)	**0.9657 (1)**	**0.9760 (1)**	0.8353 (3)	**0.9591 (1)**	**0.9093 (1)**	**0.7116 (1)**	**0.8742 (1)**	0.8843 (2)	**0.7842 (1)**
Cuevas and García (2013)	0.6797 (5)	0.5065 (3)	0.6607 (2)	0.9280 (3)	0.8995 (3)	0.7462 (6)	0.8699 (4)	0.8214 (2)	0.0895 (6)	0.6527 (5)	0.8074 (7)	0.7386 (4)
Haines and Xiang (2014)	0.7064 (3)	**0.6285 (1)**	**0.7333 (1)**	0.6946 (7)	0.9588 (2)	0.8225 (5)	0.9590 (2)	0.3054 (7)	0.0426 (7)	0.8115 (2)	**0.9021 (1)**	0.7826 (2)

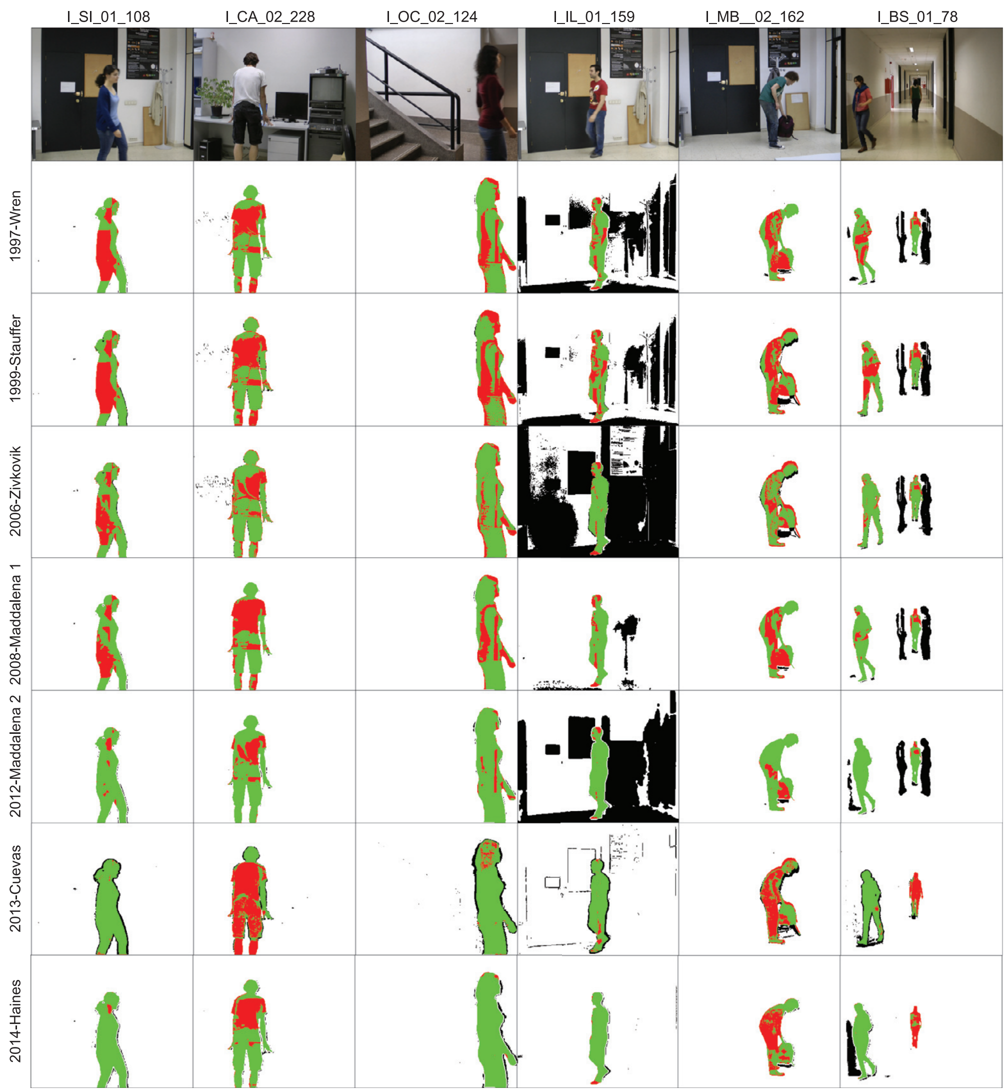

Fig. 13. Some representative results obtained in indoor sequences. Color notation: correct detections in green, false detections in black and misdetections in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cd is the number of correct detections, *fd* is the amount of false detections and *md* is the amount of misdetections.

The evaluation of the sequences has been carried out without training period. Additionally, all the algorithms have been run with a single set of parameters throughout the sequences. In the case of the algorithms that rely on few parameters (Wren et al., 1997; Stauffer and Grimson, 2000; Zivkovic and van der Heijden, 2006; Cuevas and García, 2013), we have set such parameters with the values that give the best overall results on all the test sequences. For the algorithms depending on much more parameters

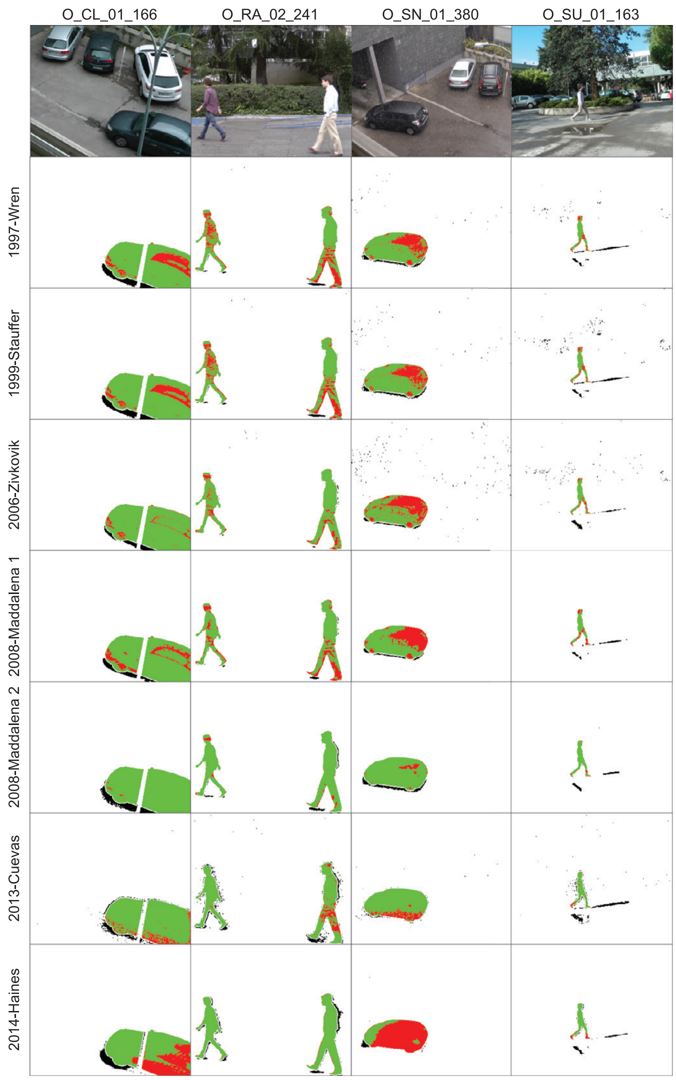

Fig. 14. Some representative results obtained in outdoor sequences. Color notation: correct detections in green, false detections in black and misdetections in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Maddalena and Petrosino, 2008; Maddalena and Petrosino, 2012; Haines and Xiang, 2014), the default configurations recommended by their authors have been used.

Regarding the sequences with stationary foreground objects (i.e., moving objects that remain static throughout several frames), none of the evaluated strategies includes mechanisms to detect these stationary objects. Consequently, to avoid making a decision about at what time such objects should become part of the background, we have decided to ignore the results corresponding to the pixels labeled as part of stationary foreground objects. That is, the pixels with this label have been treated as uncertainty zones.

Finally, it must be noted that all the evaluated strategies are focused on detecting moving objects recorded with fixed cameras: they are able of dealing with slight camera motion (e.g. camera vibrations), but in no case they were conceived to deal with the amount of motion in the motion sequences of LASIESTA. For this reason, in the current analysis we have omitted the sequences with motion (both natural or simulated).

Table 7 summarizes the obtained *F*-score values and some representative results are illustrated in Figs. 13 and 14 for, respectively, indoor and outdoor sequences (one image per category).

These results show that none of the analyzed strategies is clearly better than the rest for all the categories in LASIESTA. On the contrary, since many of them have been designed to solve specific problems, the quality of each strategy highly depends on the evaluated challenge. On the one hand, the traditional parametric methods are able to achieve the best results in some sequences with simple background and few shadows (e.g. I_OC_01 or O_RA_01). On the other hand, multimodal methods are those obtaining the best quality detections in scenarios with dynamic background. Additionally, it can be observed that the strategies in 2013-Cuevas and 2014-Haines are those that adapt faster the changes in the scene. Consequently, they obtain the best results in sequences with illumination changes and in bootstrapping sequences. However, this fast adaptation also leads to many false detections when the objects move very slow or when they start stopping.

In other databases (e.g. CITIC (Fernandez-Sanchez et al., 2013a) or STAR (Li et al., 2004)) the main challenges in moving object detection are not separated adequately, which makes it difficult to determine the quality of the strategies for each specific challenge. On the contrary, the results obtained for LASIESTA show that it separates adequately the challenges, which makes it possible a much more detailed assessment of the strengths and weaknesses of each strategy.

This important advantage of LASIESTA, together with other virtues mentioned in the previous sections of this article (e.g. it is the only existing data base composed by natural content that is 100% annotated at both object and pixel levels and it is also the unique that contains information about stationary foreground objects), make it clear that the proposed database is the most suitable for the evaluation of future moving object detection approaches.

Moreover, as described in Section 5, the length of the sequences in LASIESTA is just enough to assess the challenges, which facilitates the performance of experiments. On the contrary, many other databases (e.g. ChangeDetection (Goyette et al., 2012) or SABS (Brutzer et al., 2011)) are composed of sequences with an excessively high length.

7. Conclusions

A new, complete, compact, and well structured database for moving object detection strategies has been proposed.

This database contains two main sets of real sequences, indoor and outdoor, composed of a large variety of sequences recorded in many scenarios. Each main set consists of different categories addressing different challenges (illumination changes, dynamic backgrounds, camouflage, etc.). Additionally, the database includes some categories specifically designed to assess the quality of the results of strategies designed to work with moving cameras.

In contrast to other databases with real sequences, all the sequences in the proposed one are fully annotated. Additionally, the annotations have been carried out at both pixel and object levels, which makes LASIESTA suitable for assessing the quality of moving object detection methods integrating tracking strategies. It also includes information about moving objects remaining temporarily static, allowing each user to decide at what time these objects become part of the background.

The proposed database has been used to evaluate the performance of some classical and current moving object detection methods. The overall results obtained with all these methods have shown that, thanks to the great variety of well separated challenges it contains, LASIESTA makes it possible a very detailed assessment of the strengths and weaknesses of the evaluated strategies. Therefore, the LASIESTA database has shown to be more suitable than other databases for the evaluation of future moving object detection approaches.

Acknowledgments

This work was supported in part by the Ministerio de Economía y Competitividad of the Spanish Government under project and TEC2013-48453 (MR-UHDTV).

References

Amato, A., Huerta, I., Mozerov, M.G., Roca, F.X., Gonzàlez, J., 2014. Moving cast shadows detection methods for video surveillance applications. In: Wide Area Surveillance. Springer, pp. 23–47.
Baxter, R.H., Robertson, N.M., Lane, D.M., 2015. Human behaviour recognition in data-scarce domains. Pattern Recognit. 48 (8), 2377–2393.
Blunsden, S., Fisher, R., 2010. The behave video dataset: ground truthed video for multi-person behavior classification. Ann. Br. Mach. Vis. Assoc. 2010 (4), 1–12.
Bourdis, N., Marraud, D., Sahbi, H., 2011. Constrained optical flow for aerial image change detection. In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 4176–4179.
Bouwmans, T., 2014. Traditional and recent approaches in background modeling for foreground detection: an overview. Comput. Sci. Rev. 11, 31–66.
Bouwmans, T., El Baf, F., Vachon, B., 2008. Background modeling using mixture of Gaussians for foreground detection-a survey. Recent Pat. Comput. Sci. 1 (3), 219–237.
Bouwmans, T., Gonzàlez, J., Shan, C., Piccardi, M., Davis, L., 2014. Special issue on background modeling for foreground detection in real-world dynamic scenes. Mach. Vis. Appl. 1–3.
Brutzer, S., Hoferlin, B., Heidemann, G., 2011. Evaluation of background subtraction techniques for video surveillance. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1937–1944.
Chang, H.J., Jeong, H., Choi, J.Y., 2012. Active attentional sampling for speed-up of background subtraction. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2088–2095.
Chiranjeevi, P., Sengupta, S., 2012. Robust detection of moving objects in video sequences through rough set theory framework. Image Vis. Comput. 30 (11), 829–842.
Creusot, C., 2014. Local segmentation for pedestrian tracking in dense crowds. In: MultiMedia Modeling. Springer, pp. 266–277.
Cuevas, C., Del Blanco, C.R., García, N., Jaureguizar, F., 2010. Segmentation-tracking feedback approach for high-performance video surveillance applications. In: IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI). IEEE, pp. 41–44.
Cuevas, C., Garcia, N., 2013. Efficient moving object detection for lightweight applications on smart cameras. IEEE Trans. Circuits Syst. Video Technol. 23 (1), 1–14.
Cuevas, C., García, N., 2013. Improved background modeling for real-time spatio-temporal non-parametric moving object detection strategies. Image Vis. Comput. 31 (9), 616–630.
Cuevas, C., Yáñez, E.M., García, N., 2015. Tool for semiautomatic labeling of moving objects in video sequences: TSLAB. Sensors 15 (7), 15159–15178.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. IEEE, pp. 248–255.
Denina, G., Bhanu, B., Nguyen, H.T., Ding, C., Kamal, A., Ravishankar, C., Roy-Chowdhury, A., Ivers, A., Varda, B., 2011. Videoweb dataset for multi-camera activities and non-verbal communication. In: Distributed Video Sensor Networks. Springer, pp. 335–347.
Desurmont, X., Wijnhoven, R., Jaspers, E., Caignart, O., Barais, M., Favoreel, W., Delaigle, J.-F., 2005. Performance evaluation of real-time video content analysis systems in the Candela project. In: Electronic Imaging 2005. International Society for Optics and Photonics, pp. 200–211.
Dollár, P., Wojek, C., Schiele, B., Perona, P., 2009. Pedestrian detection: a benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 304–311.
Elgammal, A., Harwood, D., Davis, L., 2000. Non-parametric model for background subtraction. In: Computer VisionECCV 2000. Springer, pp. 751–767.
Evangelio, R.H., Sikora, T., 2011. Static object detection based on a dual background model and a finite-state machine. EURASIP J. Image Video Process. 2011, 1–11.
Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The Pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88 (2), 303–338.
Fernandez-Sanchez, E.J., Diaz, J., Ros, E., 2013a. Background subtraction based on color and depth using active sensors. Sensors 13 (7), 8895–8915.
Fernandez-Sanchez, E.J., Rubio, L., Diaz, J., Ros, E., 2013b. Background subtraction model based on color and depth cues. Mach. Vis. Appl. 25, 1211–1225.
Fisher, R., Santos-Victor, J., Crowley, J., 2005. Caviar: Context Aware Vision Using Image-Based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR.
Gallego, J., Pardàs, M., Haro, G., 2012. Enhanced foreground segmentation and tracking combining Bayesian background, shadow and foreground modeling. Pattern Recognit. Lett. 33 (12), 1558–1568.
García-Martín, Á., Martínez, J.M., Bescós, J., 2012. A corpus for benchmarking of people detection algorithms. Pattern Recognit. Lett. 33 (2), 152–156.
Goldmann, L., De Simone, F., Ebrahimi, T., 2010. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics. 75260S–75260S.
Goyette, N., Jodoin, P.-M., Porikli, F., Konrad, J., Ishwar, P., 2012. Changedetection.net: a new change detection benchmark dataset. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 1–8.
Grgic, M., Delac, K., Grgic, S., 2011. Scface–surveillance cameras face database. Multimedia ToolsAppl. 51 (3), 863–879.
Haines, T.S., Xiang, T., 2014. Background subtraction with Dirichletprocess mixture models. Pattern Anal. Mach. Intell. IEEE Trans. 36 (4), 670–683.
Horprasert, T., Harwood, D., Davis, L.S., 1999. A statistical approach for real-time robust background subtraction and shadow detection. In: IEEE International Conference on Computer Vision, 99, pp. 1–19.
Hsiao, H.-H., Leou, J.-J., 2013. Background initialization and foreground segmentation for bootstrapping video sequences. EURASIP J. Image Video Process. 2013 (1), 1–12.
Hull, J.J., 1994. A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. 16 (5), 550–554.
Intel Corporation, 2015. Opencv: Open Source Computer Vision. http://opencv.org/.
Kanade, T., Cohn, J.F., Tian, Y., 2000. Comprehensive database for facial expression analysis. In: IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, pp. 46–53.
Kim, J., Rivera, A.R., Song, G., Ryu, B., Chae, O., 2013. Edge-segment-based background modeling: non-parametric online background update. In: International Conference on Advanced Video and Signal Based Surveillance. IEEE, pp. 214–219.
Kolekar, M.H., Palaniappan, K., Sengupta, S., 2008. Semantic event detection and classification in cricket video sequence. In: IEEE Indian Conference on Computer Vision, Graphics & Image Processing. IEEE, pp. 382–389.
Kong, H., Audibert, J.-Y., Ponce, J., 2010. Detecting abandoned objects with a moving camera. IEEE Trans. Image Process. 19 (8), 2201–2210.
Kryjak, T., Komorkiewicz, M., Gorgon, M., 2013. Hardware implementation of the pbas foreground detection method in fpga. In: International Conference on Mixed Design of Integrated Circuits and Systems. IEEE, pp. 479–484.
Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. Hmdb: a large video database for human motion recognition. In: IEEE International Conference on Computer Vision. IEEE, pp. 2556–2563.
Lai, K., Bo, L., Ren, X., Fox, D., 2011. A large-scale hierarchical multi-view rgb-d object dataset. In: IEEE International Conference on Robotics and Automation. IEEE, pp. 1817–1824.
Li, L., Huang, W., Gu, I.-H., Tian, Q., 2004. Statistical modeling of complex backgrounds for foreground object detection. IEEE Trans. Image Process. 13 (11), 1459–1472.
Li, Q., Mao, Y., Wang, Z., Xiang, W., 2009. Robust real-time detection of abandoned and removed objects. In: Image and Graphics, 2009. ICIG'09. Fifth International Conference on. IEEE, pp. 156–161.
i LIDS Team, 2006. Imagery library for intelligent detection systems (i-lids); a standard for testing video based detection systems. In: IEEE Carnahan Conference on Security Technology. IEEE, pp. 75–80.
Lin, K., Chen, S.-C., Chen, C.-S., Lin, D.-T. D., Hung, Y.-P., 2015. Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance. Inf. Forensics Secur. IEEE Trans. 10 (7), 1359–1370.
Lingam, R.A., Kumar, K.S., 2014. Statistically tuned Gaussian background subtraction technique for uav videos. Sadhana 39 (4), 785–808.
Liu, C.-L., Yin, F., Wang, D.-H., Wang, Q.-F., 2011. Casia online and offline chinese handwriting databases. In: IEEE International Conference on Document Analysis and Recognition. IEEE, pp. 37–41.
Maddalena, L., Petrosino, A., 2008. A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans. Image Process. 17 (7), 1168–1177.
Maddalena, L., Petrosino, A., 2012. The sobs algorithm: what are the limits? In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. IEEE, pp. 21–26.
Maddalena, L., Petrosino, A., 2013. Stopped object detection by learning foreground model in videos. Neural Netw. Learn. Syst. IEEE Trans. 24 (5), 723–735.
Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: IEEE International Conference on Computer Vision, 2001, 2. IEEE, pp. 416–423.
Mirmahboub, B., Samavi, S., Karimi, N., Shirani, S., 2013. Automatic monocular system for human fall detection based on variations in silhouette area. Biomed. Eng. IEEE Trans. 60 (2), 427–436.
Müller, H., Michoux, N., Bandon, D., Geissbuhler, A., 2004. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int. J. Med. Inform. 73 (1), 1–23.
Nghiem, A.T., Bremond, F., Thonnat, M., Valentin, V., 2007. Etiseo, performance evaluation for video surveillance systems. In: IEEE Conference on Advanced Video and Signal Based Surveillance. IEEE, pp. 476–481.
Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al., 2011. A large-scale benchmark dataset for event recognition in surveillance video. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3153–3160.
Over, P., Awad, G., Fiscus, J., Antonishek, B., Michel, M., Smeaton, A.F., Kraaij, W., Quénot, G., et al., 2011. Trecvid 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics. TRECVID 2011-TREC Video Retrieval Evaluation Online.
PETS: The Performance Evaluation of Tracking and Surveillance. http://www.cvg.rdg.ac.uk/.

Qian, X., Hou, X., Tang, Y., Wang, H., Li, Z., 2012. Hidden conditional random field-based soccer video events detection. IET Image Process. 6 (9), 1338–1347.
Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B., 2005. Image change detection algorithms: a systematic survey. IEEE Trans. Image Process. 14 (3), 294–307.
Ragheb, H., Velastin, S., Remagnino, P., Ellis, T., 2008. Vihasi: virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods. In: ACM/IEEE International Conference on Distributed Smart Cameras. IEEE, pp. 1–10.
Ren, Y., Chua, C.-S., Ho, Y.-K., 2003. Statistical background modeling for non-stationary camera. Pattern Recognit. Lett. 24 (1), 183–196.
Rowe, S., Blake, A., 1996. Statistical mosaics for tracking. Image Vis. Comput. 14 (8), 549–564.
Saptharishi, M., Lipchin, A., Lisin, D., 2012. Discriminative focus of attention for real-time object detection in video. In: Workshop on Signal Processing Systems. IEEE, pp. 85–90.
Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: a local svm approach. In: IEEE International Conference on Pattern Recognition, 3. IEEE, pp. 32–36.
Seidel, F., Hage, C., Kleinsteuber, M., 2013. pROST: a smoothed lp-norm robust online subspace tracking method for background subtraction in video. Mach. Vis. Appl. 25 (5), 1227–1240.
Shapiro, G., 2013. Consumer electronics association's five technology trends to watch: exploring new tech that will impact our lives. IEEE Consum. Electron. Mag. 2 (1), 32–35.
Sheikh, Y., Javed, O., Kanade, T., 2009. Background subtraction for freely moving cameras. In: IEEE International Conference on Computer Vision. IEEE, pp. 1219–1225.
Sheikh, Y., Shah, M., 2005. Bayesian modeling of dynamic scenes for object detection. IEEE Trans. Pattern Anal. Mach.Intell. 27 (11), 1778–1792.
Shimada, A., Nagahara, H., Taniguchi, R.-i., 2013. Background modeling based on bidirectional analysis. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1979–1986.
Sigal, L., Black, M.J., 2006. Humaneva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion, 120. Brown Univertsity TR.
Singh, S., Velastin, S.A., Ragheb, H., 2010. Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods. In: IEEE International Conference on Advanced Video and Signal Based Surveillance. IEEE, pp. 48–55.
Sobral, A., Vacavant, A., 2014. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Comput. Vis. Image Understanding 122, 4–21.
Spinello, L., Arras, K.O., 2011. People detection in rgb-d data. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, pp. 3838–3843.
Spinello, L., Arras, K.O., Triebel, R., Siegwart, R., 2010. A layered approach to people detection in 3d range data. In: AAAI Conference on Artificial Intelligence. AAAI, pp. 1625–1630.
Spitzer, V.M., Whitlock, D.G., 1998. The visible human dataset: the anatomical platform for human simulation. Anat. Rec. 253 (2), 49–57.
Stauffer, C., Grimson, W.E.L., 2000. Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. 22 (8), 747–757.
Taylor, G.R., Chosak, A.J., Brewer, P.C., 2007. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.
Toyama, K., Krumm, J., Brumitt, B., Meyers, B., 1999. Wallflower: principles and practice of background maintenance. In: IEEE International Conference on Computer Vision, 1. IEEE, pp. 255–261.
Urvoy, M., Barkowsky, M., Cousseau, R., Koudota, Y., Ricorde, V., Le Callet, P., Gutiérrez, J., García, N., 2012. NAMA3DS1-COSPAD1: subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In: International Workshop on Quality of Multimedia Experience. IEEE, pp. 109–114.
Wali, A., Alimi, A.M., 2010. Hybrid approach for event detection from video surveillance sequences. In: IEEE International Symposium on I/V Communications and Mobile Network. IEEE, pp. 1–6.
Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P., 1997. Pfinder: real-time tracking of the human body. Pattern Anal. Mach. Intell. IEEE Trans. 19 (7), 780–785.
Xu, L., Qi, F., Jiang, R., 2006. Shadow removal from a single image. In: IEEE International Conference on Intelligent Systems Design and Applications, 2. IEEE, pp. 1049–1054.
Ye, Y., Ci, S., Liu, Y., Tang, H., 2011. Dynamic video object detection with single ptu camera. In: IEEE Visual Communications and Image Processing. IEEE, pp. 1–4.
Yi, K.M., Yun, K., Kim, S.W., Chang, H.J., Choi, J.Y., 2013. Detection of moving objects with non-stationary cameras in 5.8 ms: bringing motion detection to your mobile device. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 27–34.
Yin, P., Criminisi, A., Winn, J., Essa, I., 2007. Tree-based classifiers for bilayer video segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.
Zhang, T., Ghanem, B., Liu, S., Ahuja, N., 2013. Robust visual tracking via structured multi-task sparse learning. Int. J. Comput. Vis. 101 (2), 367–383.
Zivkovic, Z., van der Heijden, F., 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognit. Lett. 27 (7), 773–780.