TEC2014-53176-R HAVideo (2015-2017)

High Availability Video Analysis for People Behaviour Understanding

D4.1 v3

Evaluation methodology and datasets

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

AUTHORS LIST

Álvaro García Martín (agm)	alvaro.garcia@uam.es

HISTORY

Version	Date	Editor	Description
0.9	07/12/2015	Álvaro García Martín	Final Working Draft
1.0	15/12/2015	José M. Martínez	Editorial checking
1.9	17/12/2016	Álvaro García Martín	Final Working Draft version 2
2.0	19/12/2016	José M. Martínez	Editorial checking
2.1	21/06/2017	Álvaro García Martín	Parking Lot dataset inclusion
3.0	25/06/2017	José M. Martínez	Editorial checking

CONTENTS:

1. Introduction

1.1. Motivation

The work package 4 (WP4) aims at evaluating and integrating the algorithms developed within WP1, WP2 and WP3, in order to conform the global analysis chain to provide solutions to long-term video analysis for people behaviour understanding. In particular, this deliverable describes the work related with the task T.4.1: Evaluation framework.

In order to identify which approaches operate better in certain situations or applications, performance evaluation has been proposed in the literature as a way to determine their strengths and weaknesses. The widely used empirical approach consists on the performance evaluation through the analysis of the obtained results using previously annotated ground truth. For such analysis, two main aspects 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 to measure the precision of algorithms (which allow to quantify their performance).

For these reason, this first task T.4.1 goal is the definition of the evaluation methodologies and the establishment of evaluation frameworks for the approaches developed in the project. It includes the selection of appropriate datasets (sequences and associated ground-truth) and their generation if required.

1.2. Document structure

This document contains the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: Overview of the evaluation proposed in the HAVideo project
- Chapter 3: Describes the available evaluation material for the main stages of video surveillance systems that are also studied in the HAVideo project
- Chapter 4: Defines the evaluation methodologies used in the HAVideo project
- Chapter 5: Conclusions

2. Overview of the evaluation proposed in the HAVideo project

2.1. Selected analysis stages

The people behaviour understanding in this project has been already designed as a sequential combination of object segmentation, people detection, object tracking and behaviour recognition. For the HAVideo project, we consider the already commented stages that compose a video surveillance system. They are:

- Object segmentation: Video object segmentation [1] isolates the objects of interest and feeds the other system stages. In fixed camera scenarios, it is often approached by background subtraction (BS) [2] to segregate the foreground from the background of the scene.

- People detection: The people detection stage [3] spatially locates people in frames by adjusting a person model to the candidate regions in the scene. Such model is learned offline based on characteristic parameters (motion, silhouette, size, etc). This stage can be combined with object segmentation to improve the detection accuracy or used in moving camera scenarios where BS fails.

- Object tracking: Video object tracking [4] computes the spatial location of objects over time by associating object (or people) locations in consecutive frames. This stage considers models of objects, created at initialization time, which are updated over time to cope with object variations.

- Behaviour recognition: The goal of behaviour recognition is to detect (spatial and temporal) the activities of interest in the scene by extracting features from the segmented and tracked objects. Two main strategies exist based on semantic and probabilistic analysis [5]. The former defines rules to infer the activity occurrence whereas the latter considers learning models to account for the intrinsic variability of extracted features. This stage relies on the information generated by the whole system and therefore, successful behaviour recognition has a direct dependence with the accuracy of the system stages.

3. Evaluation material

During the first part of the project there have been a focus on evaluating different approaches for segmentation, people detection and tracking. For these three analysis stages mentioned in section 2.1, we describe the available evaluation material based on visual information to be used within the HAVideo project.

3.1. Object segmentation

For video object segmentation, some datasets from the state of the art has been used by the VPULab focused on the main problems that affect video-object segmentation. Moreover, an analysis of publicly available datasets is also provided in the appendix.

3.1.1. CUHK Square dataset

The CUHK dataset [6] dataset includes one video sequence of 60 minutes (90000 frames) with resolution 720x576, resized to 360x288 in the paper for faster operation. It is originally designed for adapting generic pedestrian detectors but it is useful for static and non-static object segmentation analysis. It is recorded by a stationary camera in an outdoors scenario, see Figure 1.

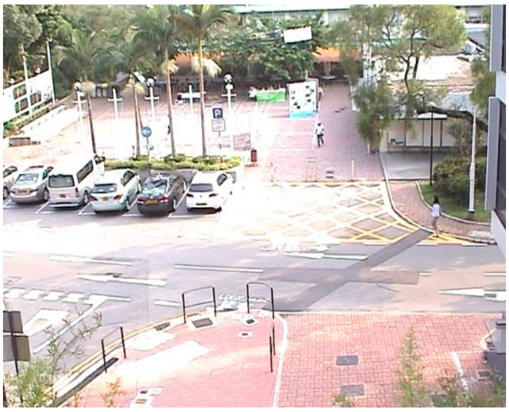


Figure 1. Sample frame for the CUHK dataset.

3.1.2. IDIAP dataset

The IDIAP dataset [7] includes one video sequence of 44.13 minutes (66324 frames) with resolution of 360x288. The video contains multiple instances of rare or unusual events such as vehicle stopping after the stop line, people crossing the road away from the zebra crossing, jay walking, and car entering pedestrian area. It is originally used for the detection of abnormal events such as vehicle stopping after the stop line, people crossing the road away from the zebra crossing, jay walking, and car entering pedestrian area. Therefore, it is also useful for static and non-static object segmentation analysis, see Figure 2.

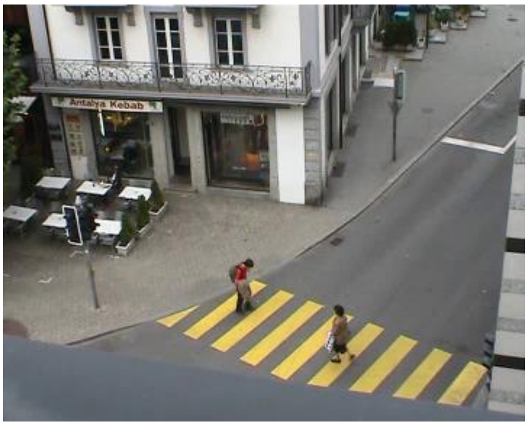


Figure 2. Sample frame for the IDIAP dataset.

3.1.3. Virat dataset

The Virat dataset [8] is a large video-surveillance dataset with 11 environments and approximately 8.5 hours of recorded HD videos. These sequences contain 12 event types annotated in ground and Aerial Videos. In particular, for the object detection segmentation task, we have merged all the short clips from VIRAT which are continuous in time and useful for Stationary Object Detection (SOD), conforming 4 sequences, see one example in Figure 3.

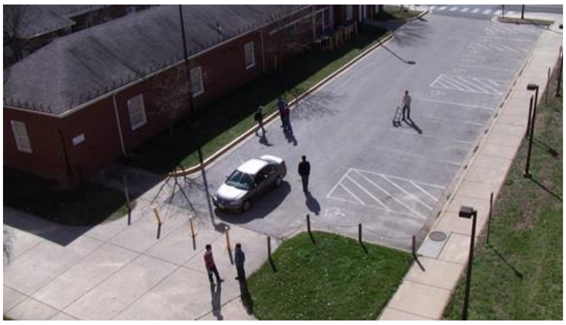


Figure 3. Sample frame for the Virat dataset.

3.1.4. AVSS2007

The AVSS2007 is a dataset for event detection in CCTV footage and is a sub-set of the i-Lids dataset. The events of interest appearing in the dataset are abandoned baggage and parked vehicle. For both tasks, there are 7 sequences (25 fps) with a total amount of 35000 frames at 25 fps. The original resolution of 720 x 576 pixels has been reduced to 360x288 to achieve faster operation. Furthermore, there are two additional long sequences for each task with around 17 minutes and 22 minutes, respectively, see one example in *Figure 4*.

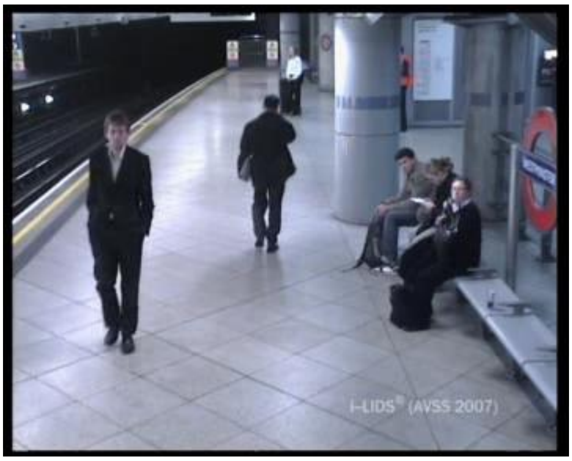


Figure 4. *Sample frame for the AVSS2007 dataset.*

3.1.5. PETS2006

The PETS2006 [9] dataset contains multi-camera video sequences for left-luggage detection recorded with 25 fps and with resolution 768x576 pixels. The same scenario is used to record seven video sequences of increasing complexity with four cameras each. We have used the camera 3 from sequence one as in related work and reduced the original resolution of 768 x 576 pixels to 384x288 for faster operation, see one example in Figure 5.

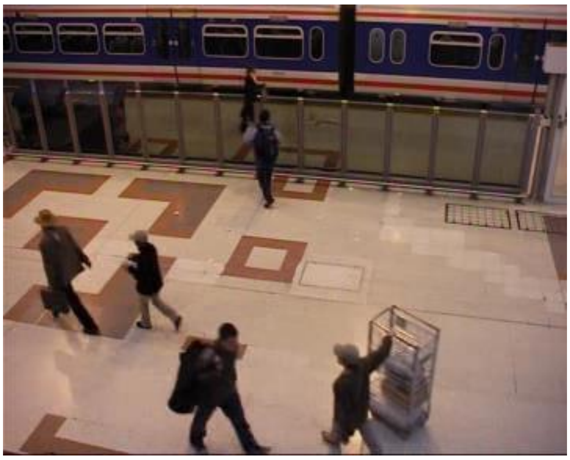


Figure 5. Sample frame for the PETS2006 dataset.

3.1.6. SBMnet dataset

The SBMnet (Scene Background Modeling) dataset (http://scenebackgroundmodeling.net/) provides a realistic and diverse set of videos. They have been selected to cover a wide range of detection challenges and are representative of typical indoor and outdoor visual data captured today in surveillance, smart environment, and video database scenarios. These videos come from our personal collection as well as from public datasets, namely CDnet, BMC2012, VSSN, the

SABS, LASIESTA, LIMU, CMU, ICRA, IPPR, CIRL, ATON, UCF, MIT, Fish4Knowledge and PETS. SBMnet was developed as part of the ICPR 2016 Scene Background Modeling Contest challenge (SBMC2016 tab). This dataset consists of 79 camera-captured videos spanning 8 categories selected to include diverse change and motion detection challenges:

1. Basic category represents a mixture of mild challenges typical of the shadows, Dynamic Background, Camera Jitter and Intermittent Object Motion categories. Some videos have subtle background motion, others have isolated shadows, some have an abandoned object and others have pedestrians that stop for a short while and then move away. These videos are fairly easy, but not trivial, to process, and are provided mainly as reference.

2. Intermittent Motion category includes videos with scenarios known for causing “ghosting” artefacts in the detected motion, i.e., objects move, then stop for a short while, after which they start moving again. Some videos include still objects that suddenly start moving, e.g., a parked vehicle driving away, and also abandoned objects. This category is intended for testing how various algorithms adapt to background changes.

3. Clutter category of videos containing a large number of foreground moving objects occluding a large portion of the background.

4. Jitter category contains indoor and outdoor videos captured by unstable (e.g., vibrating) cameras. The jitter magnitude varies from one video to another.

5. Illumination Changes: indoor videos containing strong and mild illumination changes due to a light switch, curtains opening or automatic camera brightness change.

6. Background Motion category includes scenes with strong (parasitic) background motion: boats on shimmering water, cars passing next to a fountain, or pedestrians, cars and trucks passing in front of a tree shaken by the wind.

7. Very Long: videos containing more than 3,500 frames.

8. Very Short: videos containing a limited number of frames (less than 20) with a very low framerate.

The videos have been obtained with different cameras ranging from low-resolution IP cameras through mid-resolution camcorders. As a consequence, spatial resolutions of the videos vary from 240x240 to 800x600. Also, due to diverse lighting conditions present and compression parameters used, the level of noise and compression artefacts varies from one video to another. The length of the videos also varies from 6 to 9,370 frames and the videos shot by low-end IP cameras suffer from noticeable radial distortion. Different cameras may have different hue bias (due to different white balancing algorithms employed) and some cameras apply automatic exposure adjustment resulting in global brightness fluctuations in time. The frame rate also varies from one video to another, often due to a limited bandwidth.

3.1.7. A Background Estimation dataset – Beds

The dataset [30] is focused on 4 challenges or categories conformed by 10 video sequences each and the associated ground-truth background image. Ground-truth images have been generated manually by selecting a unique frame when the background is entirely visible in a temporal instant or by mixing different areas of different temporal instants.

The generated dataset contains 4 categories or challenges key to evaluate the Background Estimation with 10 sequences containing the challenges: Baseline, Clutter, Low framerate and Static objects. Baseline, containing simple sequences with low object density and no stationary objects, to evaluate the BE task in simple scenarios. Clutter, containing sequences with high foreground motion and continuous background occlusions, situations where BE complexity is increased. Low framerate, containing simple sequences recorded with low framerate to evaluate the impact of motion velocity in the BE task. Static objects, containing stationary objects in the scene for more than and less than 50% and 100% of the video sequences duration, respectively. Figure 1 shows images of the background scenarios of all sequences from each category. This dataset and its annotated ground truth are public and available (http://www-vpu.eps.uam.es/DS/BEds/).

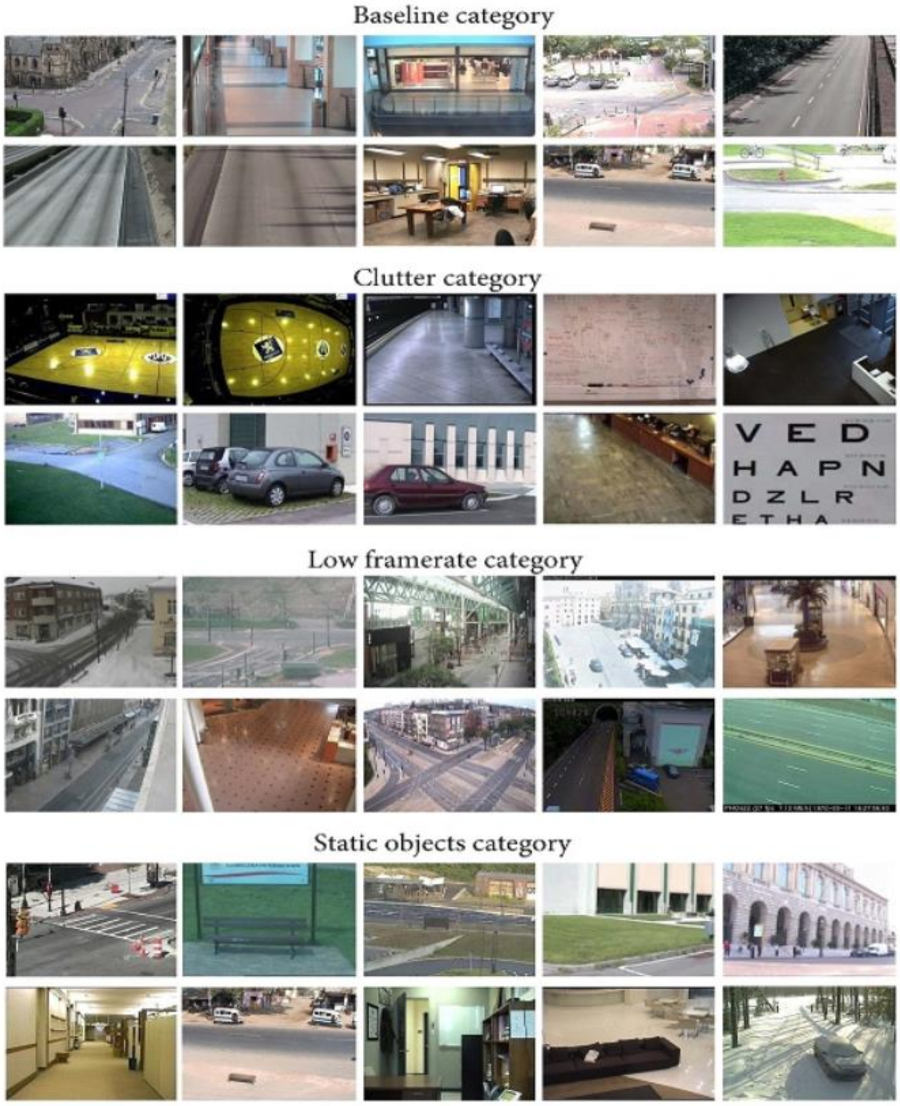


Figure 6. Images of the background scenarios of all sequences from each category.

3.2. People detection

For people detection, some datasets from the state of the art has been used by the VPULab focused on the main problems that affect people detection in surveillance videos. In addition, one dataset has been created by the VPULab (Wheelchair Users dataset, WUds) including multiple camera video sequences in a real in-door senior residence environment containing wheelchairs users and standing people. Moreover, an analysis of publicly available datasets is also provided in the appendix.

3.2.1. People detection benchmark repository - PDbm

The PDbm dataset [11] provides a realistic, camera-captured, diverse set of videos. The chosen sequences has been extracted from the Change detection dataset 2012 [10]. The video sequences have been chosen in order to cover typical people detection challenges. The dataset includes traditional indoor and outdoor scenarios in computer vision applications: video surveillance, smart cities, etc. The Change detection dataset 2012 includes the following challenges: dynamic background, camera jitter, intermittent object motion, shadows and thermal signatures.

The proposed People detection challenge includes 16 selected sequences from the whole original dataset (31 sequences). We have selected all the sequences including people (currently excluding thermal cameras because detection algorithms rarely consider thermal images). Each sequence is accompanied by a newly developed accurate people detection ground-truth (see Figure 7).

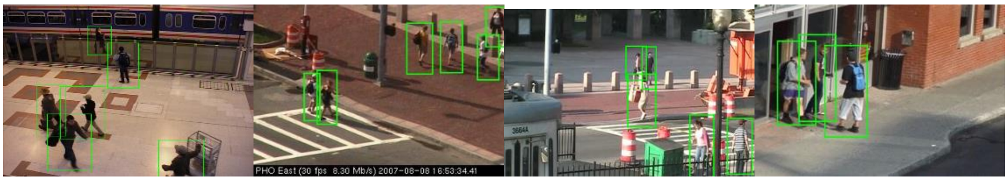


Figure 7. Sample frames for the PDbm dataset.

3.2.2. PETS

PETS [12] is the most extended database nowadays. A new database is released each year since 2000, along with a different challenge proposed. With the algorithms provided researchers can test or develop new algorithms. The best ones are presented in the conference held each year.

Since the amount of data is extensive and cover real situations, these databases are by far the most used and are almost considered a de facto standard. Despite this, it is important to say that the PETS databases are not ideal. One of its disadvantages is the fact that since PETS became a surveillance project, the challenges proposed are focused on high level applications of that field, leaving aside the tracking approach. Therefore, some important issues (such as illumination or target scale changes) are not considered. In particular for people detection, the most used one is the 2009/2010 version (see Figure 8).

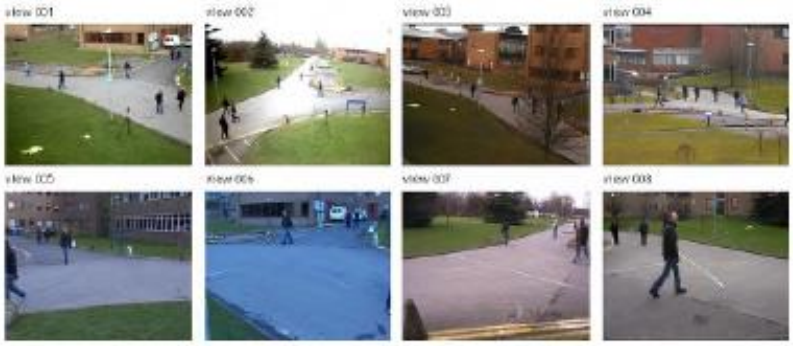


Figure 8. Sample frames for the PETS 2009/2010 dataset.

3.2.3. Smile Lab wheelchair dataset

This dataset was created by the Smile Lab (http://smile.ee.ncku.edu.tw/) at the Department of Electrical Engineering, National Cheng Kung University, Taiwan. The dataset is divided into two main image sets, the train sequences and the test sequences. Each of the frames has a resolution of 720x480 pixels. The training sequences are composed of 8 folders and a total of 3674 images, each one of them contains a set of images of wheelchairs with a defined orientation relative to the camera. The different orientations and models are shown and defined in [13].

The test sequences are composed of 4 folders, each one of them containing a sequence with a wheelchair and some people standing walking around. Unlike the training set, each of these folders contains a continuous recording, allowing to use tracking techniques to improve detection, as shown in [13]. The test set contains a total of 1314 frames divided in 4 folders. Table I shows the properties of each sequence: number of frames, number of wheelchair users and number of standing people.

The ground truth of this dataset was not available, so we created it annotating manually each of the frames from both sets. This ground truth is available for downloading as additional content in our dataset webpage (http://www-vpu.eps.uam.es/DS/WUds/), see following section.

3.2.4. Wheelchair Users dataset - WUds

This dataset was recorded by the Video Processing and Understanding Lab due to the lack of public wheelchair datasets. We used it to test the trained wheelchair users detector, as it contains sequences with a higher number of wheelchairs (up to four) and some more complex situations and scenarios (illumination changes, occlusions, etc). The sequences were recorded in a real environment of a senior residence, in order to work with an environment as realistic as possible (due to privacy issues, real recording with actual residents was not possible). Each of the frames has a resolution of 768x432 pixels and the sequences are recorded at 25 fps. Compared to the other dataset, this one contains a new environment with a larger number of sequences, a greater number of frames per sequence, and more wheelchair models (three different).

The dataset consists of 11 sequences (S1 to S11), each of them recorded from two points of views (V1 and V2), resulting in a total of 22 sequences. All sequences were recorded in the same room, using two GoPro cameras, HERO3 White edition. The fisheye effect was corrected using the GoPro Studio software tool. Each camera views are shown in Figure 9 and a room top view map is shown in Figure 10. This dataset and its annotated ground truth are public and available (http://www-vpu.eps.uam.es/DS/WUds/). The ground truth of this dataset was manually annotated for each frame of each sequence. The annotated ground truth considers the wheelchair users and the standing people present in every frame, even if they are highly occluded.

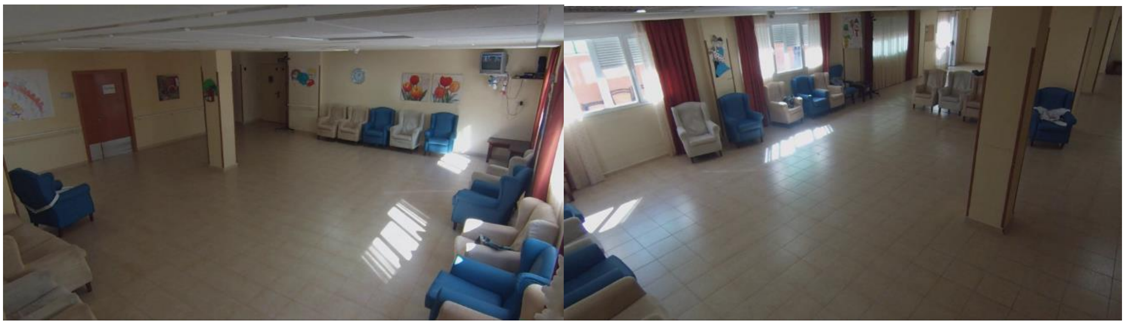


Figure 9. Camera views of the Wheelchair Users dataset. Left: viewpoint 1. Right: viewpoint 2.

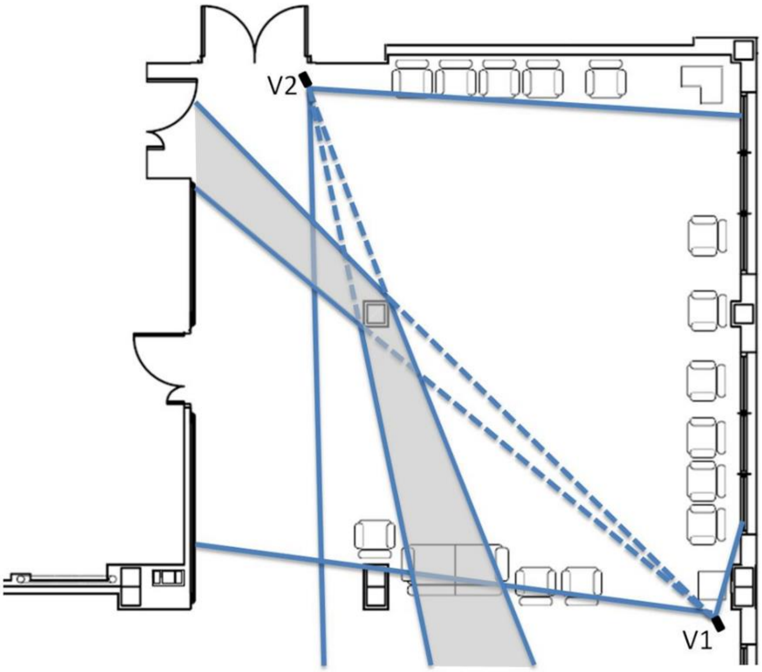


Figure 10. Top view map of the Wheelchair Users dataset. V1 and V2 represent camera 1 and camera 2 locations and fields of view.

3.2.5. Parking Lot Dataset – PLds

This dataset was recorded due to the lack of public parking vehicle datasets. The dataset consists of two main image sets: a training set and a test set, containing a total of 8616 frames. The sequences were recorded in a real environment of the Pittsburgh International Airport parking, in order to work with an environment as realistic as possible. Each of the frames has a resolution of 1280x960 pixels, recorded using Panasonic WV-SW155 cameras.

The dataset consists of two main image sets, a training set used to generate the detector models, and a test set used for the experimental evaluation. The training set consist of a longer set of images, and the test set consist of long and short versions of the images, with 1000 and 100 frames, respectively. The short versions (multicamera sets) are contained in the long version and have the frames synchronized between the two cameras, to be able to evaluate experiments combining the information of both cameras.

In addition to generating the images, the vehicles of all images have been manually annotated. The training images have been annotated for its use in the generation of the parked vehicle model, and the test images for the evaluation of the system. In the case of the synchronized multicamera set, the vehicle occupancy matrix has been manually generated.

This dataset and its annotated ground truth are publicly available (http://www-vpu.eps.uam.es/DS/PLds/). Each camera views are shown in Figure 11.

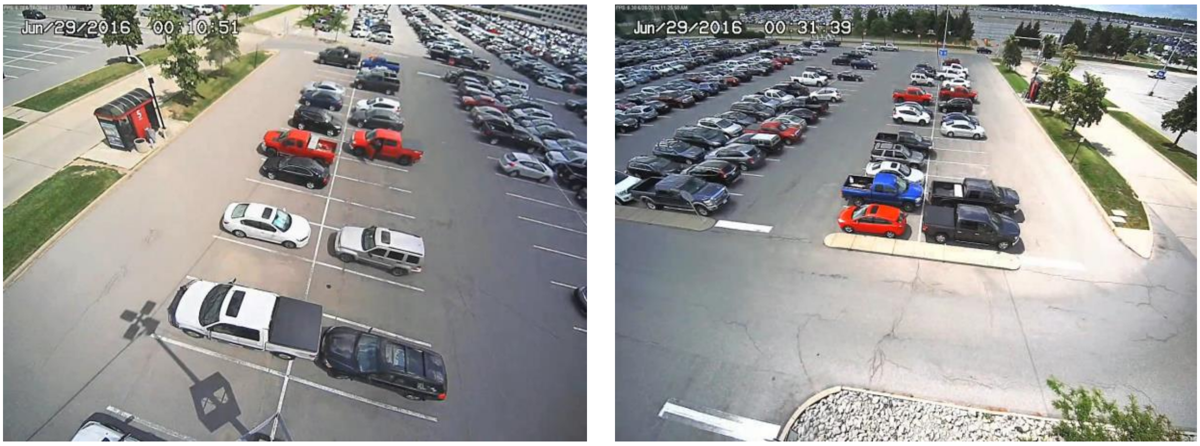


Figure 11. Examples of dataset frames.

3.3. Object tracking

For object tracking, some datasets from the state of the art has been used by the VPULab focused on the main problems that affect video-object tracking. Moreover, an analysis of publicly available datasets is also provided in the appendix.

3.3.1. VOT challenges datasets

The VOT challenges provide the visual tracking community with a precisely defined and repeatable way of comparing trackers as well as a common platform for discussing the evaluation and advancements made in the field of visual tracking. In particular this last two years there have been two different evaluations datasets: VOT2015/2016 [20]/23[28] and VOT-TIR2015/2016 [21]/[29].

3.3.1.1. VOT2015 Dataset

The Visual Object Tracking challenge 2015, VOT2015, aims at comparing short-term single-object visual trackers that do not apply pre-learned models of object appearance. The dataset comprises 60 short sequences showing various objects in challenging backgrounds. The sequences were chosen from a large pool of sequences including the ALOV dataset, OTB2 dataset, non-tracking datasets, Computer Vision Online, Professor Bob Fisher’s Image Database, Videezy, Center for Research in Computer Vision, University of Central Florida, USA, NYU Center for Genomics and Systems Biology, Data Wrangling, Open Access Directory and Learning and Recognition in Vision Group, INRIA, France. The VOT sequence selection protocol was applied to obtain a representative set of challenging sequences. Figure 12 gives an overview of the VOT2015 dataset.

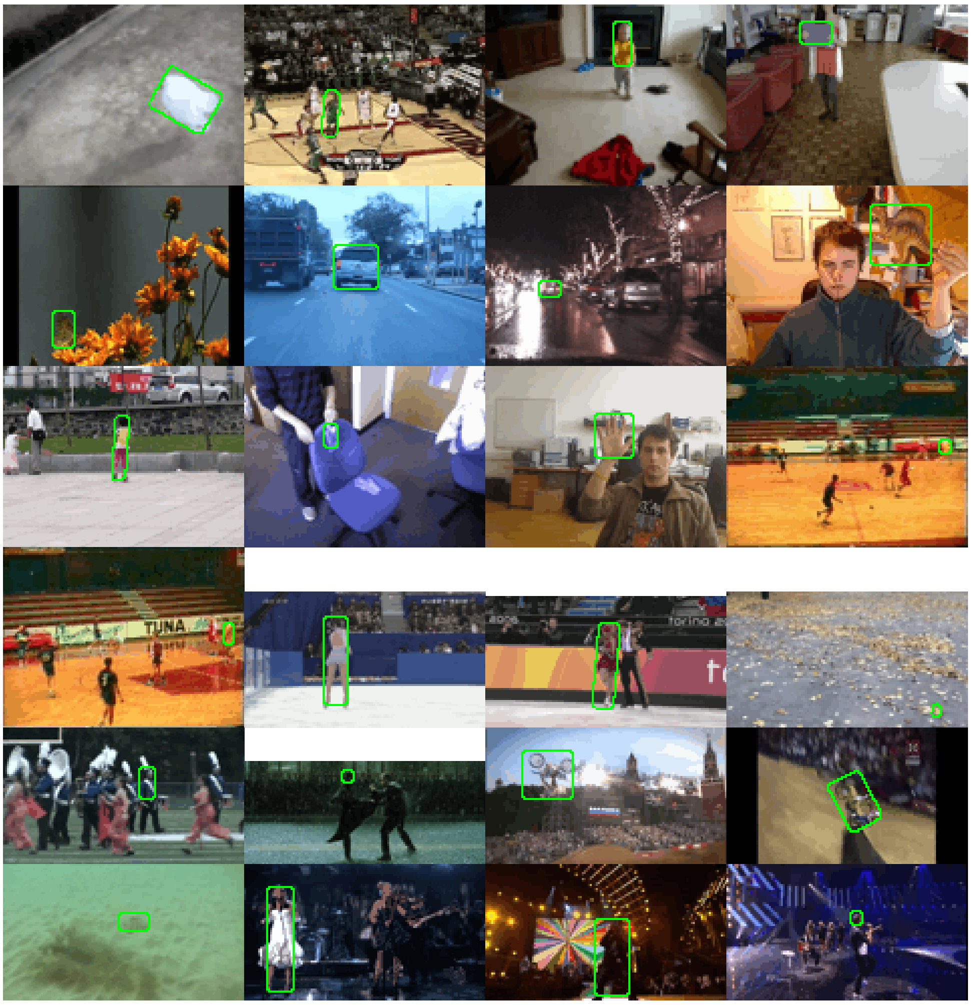


Figure 12. Overview of the VOT2015 dataset.

3.3.1.1. VOT2016 Dataset

The VOT2016 [28] dataset contains all 60 sequences from VOT2015 [20], where each sequence is per-frame annotated by the following visual attributes: (i) occlusion, (ii) illumination change, (iii) motion change, (iv) size change, (v) camera motion. In case a particular frame did not correspond to any of the five attributes, we denoted it as (vi) unassigned. However, the bounding boxes annotations have been refined.

3.3.1.2. VOT-TIR2015 Dataset

The Thermal Infrared Visual Object Tracking challenge 2015, VOT-TIR2015, aims at comparing short-term single-object visual trackers that work on thermal infrared (TIR) sequences and do not apply pre-learned models of object appearance. The VOT-TIR2015 dataset consists of 20 sequences of which eight has been recorded specifically for this dataset. The other twelve sequences have been collected from different sources including Termisk Systemteknik

AB, the Department of Electrical Engineering at Linköping University, the School of Mechanical Engineering at University of Birmingham, ETH Zürich, Fraunhofer IOSB, Aalborg University, and finally the EU FP7 project P5.

The raw signal values from a thermal infrared sensor is typically stored in 16-bit format. Since not all trackers can handle 16-bit data and for the purpose of visualization, all sequences in the dataset have been truncated to 8-bit. In practice, this is a common procedure since not all sensors give access to the 16-bit values. Therefore, the sequences are not radiometric (the corresponding temperature value is unknown) and the dynamic may adaptively change during the course of a sequence. Figure 13 gives an overview of the VOT- TIR20152015 dataset.

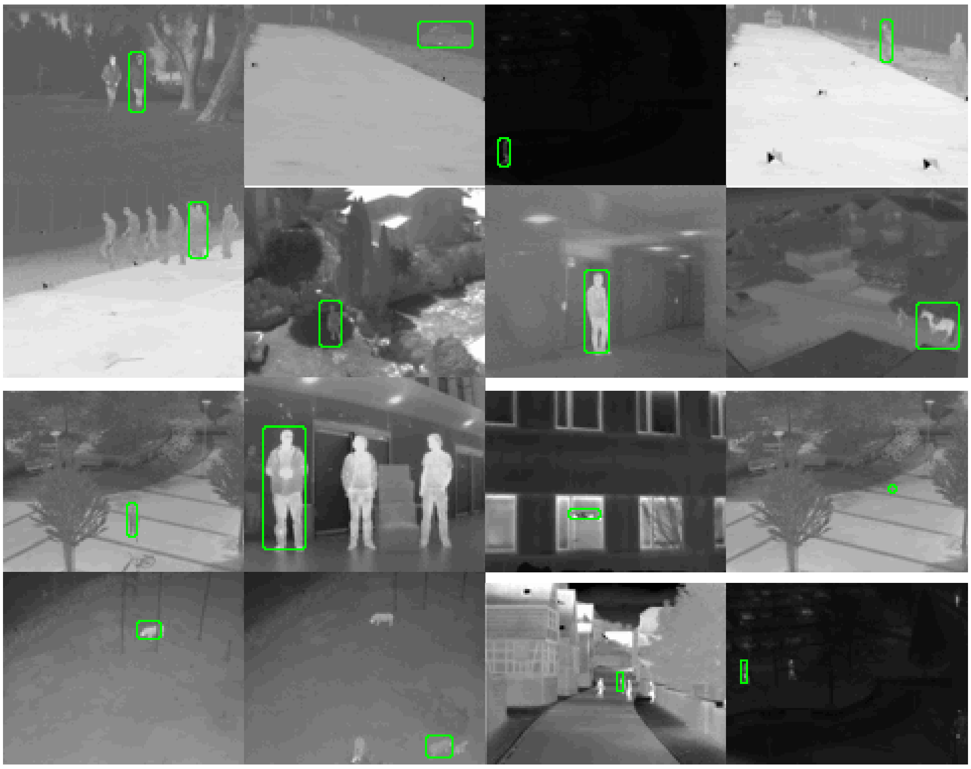


Figure 13. Overview of the VOT-TIR20152015 dataset.

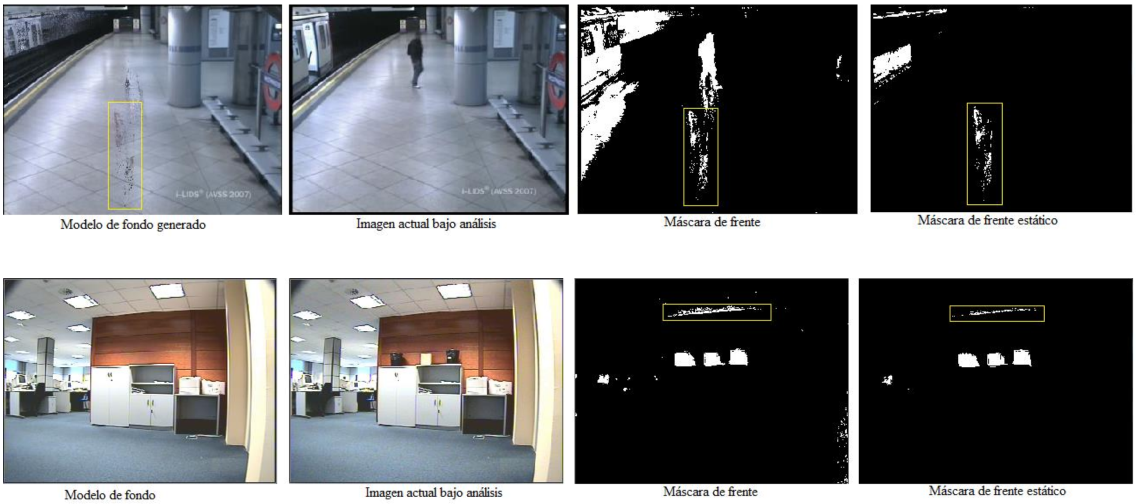
3.3.1.1. VOT-TIR2016 Dataset

Compared to VOT-TIR2015 [20], the sequences *Crossing*, *Horse*, and *Rhino behind tree* have been removed. The newly added sequences are *Bird*, *Boat1*, *Boat2*, *Car2*, *Dog*, *Excavator*, *Ragged*, and *Trees2*.

3.4. Behaviour recognition analysis tools

3.4.1. Abandoned-stolen object detection dataset

In [31] a configurable abandoned-stolen object detection system in security-video is proposed that integrates the most relevant techniques in each one of its stages. For the evaluation framework, this project proposes the evaluation of seven video sequences classified in two complexity levels: low and medium complexity. Figure 14 shows one example of each category.



Modelo de fondo generado | Imagen actual bajo análisis | Máscara de frente | Máscara de frente estático

Modelo de fondo | Imagen actual bajo análisis | Máscara de frente | Máscara de frente estático

Figure 14. Example of video sequences for object detection algorithms.

4. Defines the evaluation methodologies used in the HAVideo project

During the first part of the project there have been a focus on evaluating different approaches for segmentation, people detection and tracking. For these three analysis stages mentioned in section 2.1, we describe the evaluation methodologies to be used within the HAVideo project.

4.1. Object segmentation

The metrics most commonly used in literature are those based on ground-truth pixel level evaluation:

- True positive (TP): The number of pixels correctly classified as foreground (pixel value 1).
- True negative (TN): The number of pixels correctly classified as background (pixel value 0).

- False positive (FP): The number of pixels incorrectly classified as foreground.
- False negative (FN): The number of pixels incorrectly classified as background.

To evaluate the results, we use standard Precision (P), Recall (R) and F-score (F) measures:

- Precision: It is defined as the total number of pixels correctly classified as foreground/ background vs the total number of pixels correctly or incorrectly classified as foreground/ background.

$$P = \frac{TP}{TP + FP}$$

- Recall: It is defined as the total number of pixels correctly classified as foreground/ background vs the total real (ground truth) number of pixels of foreground/ background.

$$R = \frac{TP}{TP + FN}$$

- F-score: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

$$F = \frac{2 \times P \times R}{P + R}$$

4.1.1. Performance measures for background estimation task: SBMnet challenge metrics

The aim of the contest (http://scenebackgroundmodeling.net/) is to advance the development of algorithms and methods for scene background modeling through objective evaluation on a common dataset. In addition to providing a fine-grained videos dataset, the challenge also provides tools to compute performance metrics and thus identify algorithms that are robust across various challenges.

Methods that perform very well under one challenge (e.g., background motion) may not perform well in the presence of another challenge (e.g., strong shadows or night videos). In order to gauge performance and rank methods, they rely on the following widely used metrics:

AGE: (Average Gray-level Error). Average of the gray-level absolute difference between GT and the computed background (CB) image.

pEPs: (Percentage of Error Pixels). Percentage of EPs (number of pixels in CB whose value differs from the value of the corresponding pixel in GT by more than a threshold) with respect to the total number of pixels in the image.

pCEPS: (Percentage of Clustered Error Pixels). Percentage of CEPs (number of pixels whose 4-connected neighbors are also error pixels) with respect to the total number of pixels in the image.

MSSSIM: (MultiScale Structural Similarity Index). Estimate of the perceived visual distortion.

PSNR: (Peak-Signal-to-Noise-Ratio) Amounts to 10log_10((L-1)^2/MSE) where L is the maximum number of grey levels and MSE is the Mean Squared Error between GT and CB images.

CQM: (Color image Quality Measure). Based on a reversible transformation of the YUV color space and on the PSNR computed in the single YUV bands. It assumes values in db and the higher the CQM value, the better is the background estimate.

4.2. People detection

In order to evaluate different people detection approaches, we need to quantify the different performance results.

Some people detection literature studies compare methods based on false positive per image measure, which is generally used to evaluate the selected classifier. In order to evaluate a video surveillance system, it is more interesting to compare the overall performance. Global sequence performance is usually measured in terms of Precision-Recall (PR) curves [14], [18], [15]. These curves compare the similarities between the output and ground truth bounding boxes. In addition, in order to evaluate not only the yes/no detection decision but also the precise people locations and extents, we take into account the three evaluation criteria defined in [16], that allow to compare hypotheses at different scales: relative distance (dr), cover and overlap. A detection is considered true if dr≤0.5 (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%.

The integrated Average Precision (AP) is generally used to summarize the algorithm performance in a single value, represented geometrically as the area under the PR curve (AUC-PR). In order to approximate the area correctly, we use the approximation described by [17].

Figure 15 shows one example of PR curves over our Wheelchair Users dataset (WUds) presented in previous section.

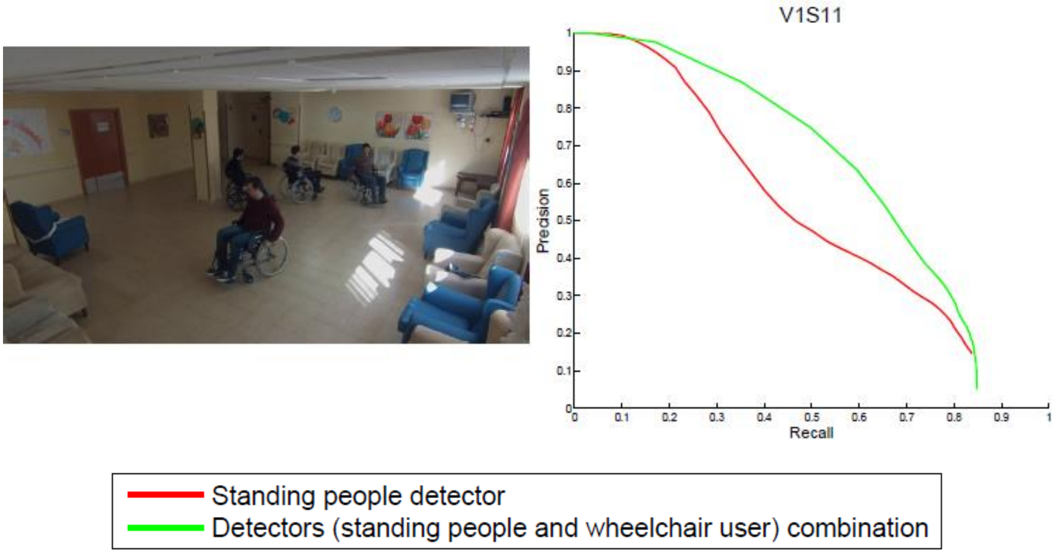


Figure 15. Precision vs Recall detection curves for the Wheelchair Users dataset sequence V1S11. AUC of 51.4% standing people detectors versus AUC of 62.7% using detectors combination.

4.3. Object tracking

4.3.1. Performance measures VOT 2015 and VOT-TIR2015

As in VOT2014 [19], the following two weakly correlated performance measures are used due to their high level of interpretability: (i) accuracy and (ii) robustness.

The accuracy measures how well the bounding box predicted by the tracker overlaps with the ground truth bounding box. On the other hand, the robustness measures how many times the tracker loses the target (fails) during tracking. A failure is indicated when the overlap measure becomes zero. To reduce the bias in robustness measure, the tracker is re-initialized five frames after the failure and ten frames after initialization are ignored in computation to further reduce the bias in accuracy measure. Stochastic trackers are run 15 times on each sequence to obtain better statistics on performance measures. The per-frame accuracy is obtained as an average over these runs. Averaging per-frame accuracies gives per-sequence accuracy, while per-sequence robustness is computed by averaging failure rates over different runs. The trackers are ranked with respect to each measure. Figure 16 shows examples of VOT 2015 final ranked in terms of Accuracy and Robustness (AR) results extracted from [20].

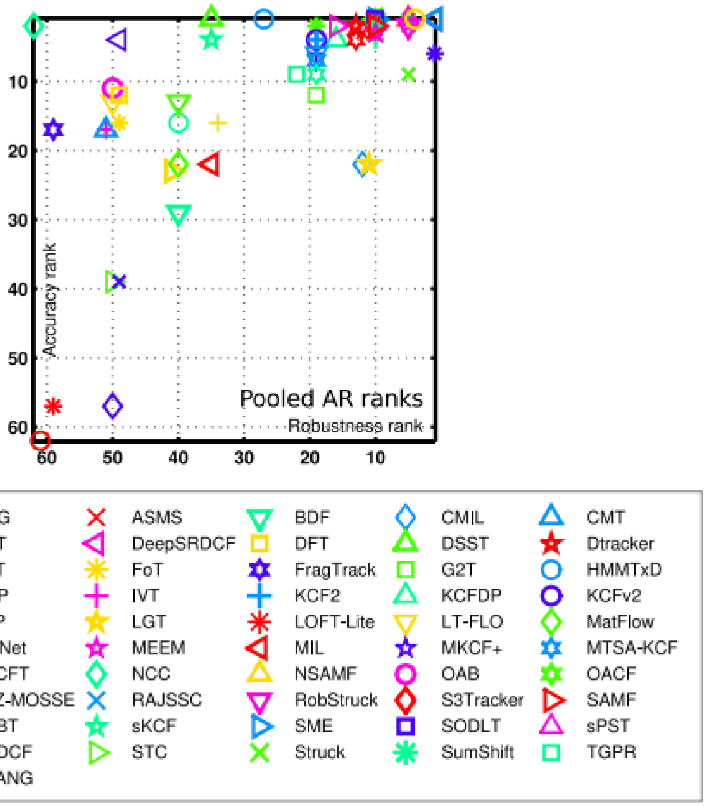


Figure 16. VOT 2015 final ranked in terms of Accuracy and Robustness (AR) results extracted from [20].

4.3.2. Performance measures VOT 2016 and VOT-TIR2016

VOT 2016 [28] and VOT-TIR2016 [29] adopt the performance measures VOT 2015 [20] and VOT-TIR2015 [21]. However, in addition to the standard reset-based VOT experiment, the VOT2016 and VOT-TIR2016 also carried out a no-reset experiment analysis.

4.4. Behaviour recognition analysis tools

4.4.1. Performance measures for abandoned-stolen object detection

In [31] a configurable abandoned-stolen object detection system in security-video is proposed that integrates the most relevant techniques in each one of its stages. For the evaluation framework, this project proposes the use of the most common metrics for abandoned-stolen object detection: Precision, Recall and FScore at object level, see object segmentation section for a more detailed description of these metrics.

5. Conclusions

In this document, we have presented the material to be used for performance evaluation within the HAVideo project. During the two first years of the project there have been a focus on evaluating different approaches for segmentation, people detection, tracking and behaviour recognition. Then, we have described the datasets used in section 3 and the methodologies for the evaluation of each stage in section 4.

In addition to the selection of appropriate datasets (sequences and associated ground-truth) and evaluation frameworks from the state of the art for segmentation, people detection, tracking and behaviour recognition. Due to the lack of public wheelchair and wheelchair users datasets, we have presented a new dataset, named Wheelchair Users dataset (WUds dataset, http://www-vpu.eps.uam.es/DS/WUds/) with sequences recorded in a real environment of a senior residence. We have also presented one dataset and evaluation framework for background estimation (Beds dataset, http://www-vpu.eps.uam.es/DS/BEds/).

6. References

[1] Weinland et al. “A survey of vision-based methods for action representation, segmentation and recognition”, Comput. Vis. Image Understand., 115(2):224-241, 2011.

[2] T. Bouwmans, "Traditional and recent approaches in background modelling for foreground detection: An overview", Comput. Science Review, 11:31-66, May 2014.

[3] P. Dollar et al, "Pedestrian Detection: An Evaluation of the State of the Art", IEEE Trans. Pattern Anal. Mach. Intell., 34(4):743-761, May 2012.

[4] A. Smeulder et al., “Visual Tracking: an Experimental Survey”, IEEE Trans. Pattern Anal. Mach. Intell., 36(7):1442-1468 July 2014.

[5] P. Borges et al., "Video-Based Human Behavior Understanding: A Survey," IEEE Trans. Circ. Syst. Video Technol., 23(11):1993-2008, Nov. 2013.

[6] M. Wang, W. Li and X. Wang, "Transferring a Generic Pedestrian Detector Towards Specific Scenes" in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[7] J. Varadarajan and J.M. Odobez, “Topic Models for Scene Analysis and Abnormality Detection”, in Proceedings of International Conference on Computer Vision Worshop (ICCVW), 2009.

[8] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen et al., “A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video”, in Proceedings of IEEE Conference on Comptuer Vision and Pattern Recognition (CVPR), 2011.

[9] J. Ferryman and D. Thirde, “An Overview of the PETS2006 Dataset”, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Worshop (CVPRW), 2006.

[10] N. Goyette, P. Jodoin, F Porikli, J. Konrad, P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in Proc. Of CVPRW, 2012, pp. 1-8.

[11] A. García-Martín, J.M. Martínez, J. Bescós,"PDbm: People detection benchmark repository", Electronic Letters Volume 51, Issue 7, p. 559 – 560.

[12] PETS, "PETS, IEEE int. workshop perform. eval. track. surveill.," Last accessed, 24 March 2016.

[13] C.-R. Huang, P.-C. Chung, K.-W. Lin, and S.-C. Tseng, "Wheelchair detection using cascaded decision tree," Information Technology in Biomedicine, vol. 14(2), pp. 292–300, 2010.

[14] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[15] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in Computer Vision and Pattern Recognition, 2009, pp. 794–801.

[16] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in Computer Vision and Pattern Recognition, 2005, pp. 878–885.

[17] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in International Conference on Machine Learning, 2006, pp. 233–240.

[18] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," International Journal of Computer Vision, vol. 77(1-3), pp. 259–289, 2008.

[19] M. Kristan et al. The visual object tracking vot2014 challenge results. In ECCV2014 Workshops, Workshop on visual object tracking challenge, 2014.

[20] Matej Kristan et al., "The Visual Object Tracking VOT2015 challenge results", Proc. of 3rd Visual Object Tracking Challenge Workshop at International Conference on Computer Vision, Santiago, Chile, December 2015, pp.564-586.

[21] Michael Felsberg et al., "The Visual Object Tracking VOT-TIR2015 challenge results", Proc. of 3rd Visual Object Tracking Challenge Workshop at International Conference on Computer Vision, Santiago, Chile, December 2015, pp.639-651.

[22] A. García, J. M. Martinez, J. Bescós: "A corpus for benchmarking of people detection algorithms", Pattern Recognition Letters, 33 (2): pp. 152-156, January 2012, ISSN 0167-8655

[23] Andriluka, M.; Roth, S.; Schiele, B.: "People-tracking-by-detection and people-detection-by-tracking", Proc. of Computer Vision and Pattern Recognition (CVPR), Anchorage, (Alaska, USA), pp. 1-8, 2008.

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

[25] "SPEVI, surveillance performance evaluation initiative," http://www.eecs.qmul.ac.uk/ andrea/spevi.html, Last accessed, March 2016.

[26] "CAVIAR context aware vision using image-based active recognition," http://homepages.inf.ed.ac.uk/rbf/CAVIAR/, Last accessed, March 2016.

[27] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, CDnet 2014: An Expanded Change Detection Benchmark Dataset, in Proc. IEEE Workshop on Change Detection (CDW-2014) at CVPR-2014, pp. 387-394. 2014

[28] Matej Kristan et al., “The Visual Object Tracking VOT2016 challenge results”, Proc. of 4th Visual Object Tracking Challenge Workshop at European Conference on Computer Vision, Amsterdam, The Netherlands, November 2016, pp. 777-823.

[29] Michael Felsberg et al., “The Thermal Infrared Visual Object Tracking VOT-TIR2016 Challenge Results”, Proc. of 4th Visual Object Tracking Workshop at European Conference on Computer Vision, Amsterdam, The Netherlands, November 2016, pp. 824-849.

[30] Reconstrucción de fondo de escena a partir de secuencias de vídeo (Background reconstruction from video sequences), Carolina Fernández-Pedraza (advisor: Diego Ortego), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2016.

[31] Integración y evaluación de sistemas de robo-abandono de objetos en video-seguridad, Jorge Gómez Vicente, Proyecto Fin de Carrera, Ing. Telecomunicación, Univ. Autónoma de Madrid en curso de realización. (tutor: Juan C. SanMiguel), Julio 2016.

Appendix

7. Additional datasets for evaluation

In this appendix, we list additional datasets for the evaluation of the selected stages in the HAVideo project.

7.1. Object segmentation

7.1.1. VSSN2006

The VSSN Workshop 2006 [24] included a motion segmentation for surveillance competition. The artificial data input sequences and corresponding ground-truth data were provided in order to have a common framework for a fair comparison of the algorithms. Each test video will consist of a video consisting of some (maybe dynamic) background and one or several foreground objects and a foreground mask video (ground truth video) specifying each pixel belonging to a foreground object (pixel values above 128; same pixel values belong to the same object, while different values belong to different objects). The dataset includes 10 sequences with ground truth and 4 sequences without ground truth.

7.1.2. CDNET dataset 2012/2014

The CDNET (ChangeDetection.NET) dataset 2014 [27] enhances the CDNET 2012 [10] dataset by incorporating 5 new categories. CDNET 2012/2014 aims to initiate a rigorous and comprehensive academic benchmarking effort for testing and ranking existing and new algorithms for change and motion detection much. It is representative of indoor and outdoor visual data captured today in surveillance and smart environment scenarios. This dataset contains 11 video categories with 4 to 6 videos sequences in each category (see Figure 17).

In overall, the dataset provides a diverse and representative set of videos. These videos have been selected to cover a wide range of foreground segregation challenges and are claimed to be representative of typical indoor and outdoor visual signals common in applications such as surveillance and smart environments. It is composed of 53 sequences, represented by colour video or thermal JPEG frames of multiple sizes with segmentation ground-truth data available.

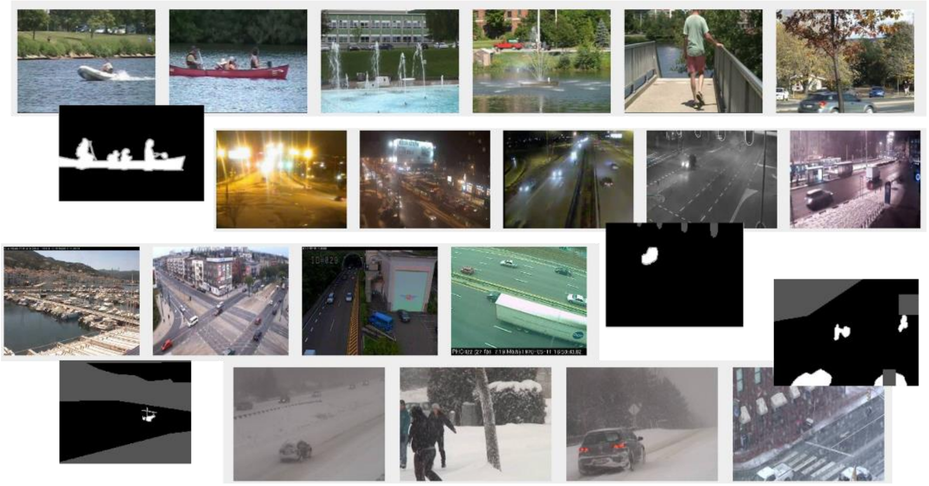


Figure 17. Sample frames for the CDNET dataset.

7.2. People detection

7.2.1. Person Detection dataset – PDds

The PDds corpus or dataset [22] consists of a set of video and associated ground-truth, for the evaluation of people detection algorithms in surveillance video scenarios. 91 sequences from scenes with different levels of complexity have been manually annotated. Each person present at a scene has been labeled frame by frame, in order to automatically obtain a people detection ground-truth for each sequence. Sequences have been classified into different complexity categories depending on critical factors that typically affect the behavior of detection algorithms. The resulting corpus exceeds other public pedestrian datasets in the amount of video sequences and its complexity variability (see Figure 18).

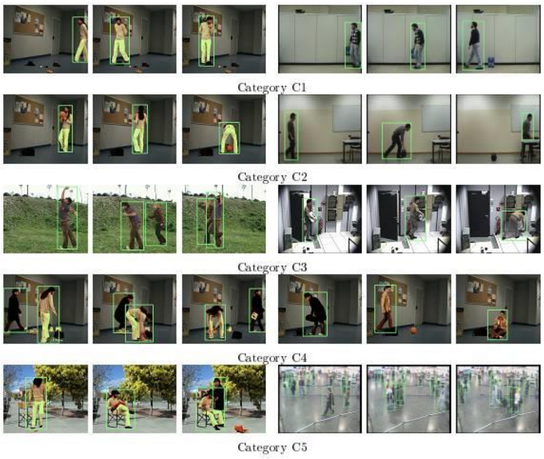


Figure 18. Sample frames for the PDds dataset.

7.2.2. TUD-Pedestrians

The TUD Pedestrians dataset [23] from Micha Andriluka, Stefan Roth and Bernt Schiele consists of training images and test sequences. The TUD pedestrian dataset consists of 250 images with 311 fully visible people with significant variation in clothing and articulation and 2 video sequences with highly overlapping pedestrians with significant variation in clothing and articulation (see Figure 19).

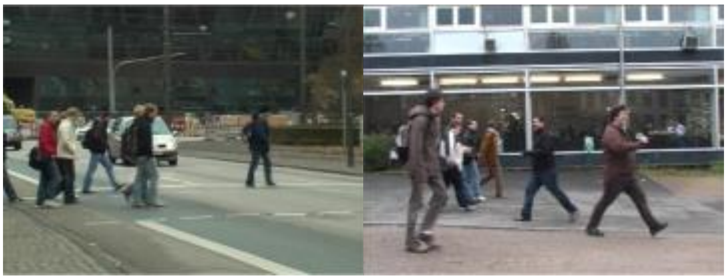


Figure 19. Sample frames for the TUD-Pedestrians dataset.

7.3. Object tracking

7.3.1. SPEVI

The Surveillance Performance EValuation Initiative (SPEVI) [25] is a set of links of publicly available datasets for researches. The videos can be used for testing and evaluating video tracking algorithms for surveillance-related applications. Two datasets are especially interesting regarding the tracking evaluation and they are described as follows. The Single Face Dataset for single person/face visual detection and tracking. And the Multiple Face Dataset for multiple people/faces visual detection and tracking (see Figure 20).

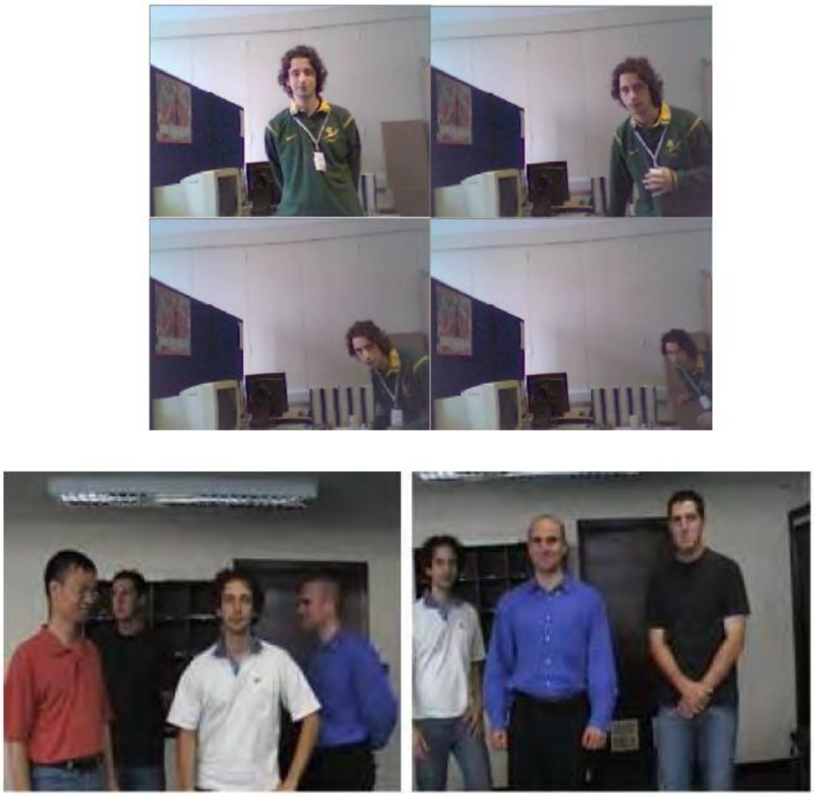


Figure 20. Sample frames for the TUD-Pedestrians dataset.

7.3.2. CAVIAR

The main objective of CAVIAR dataset [26] includes sequences of people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place. All video clips were filmed with a wide-angle camera lens, and some scenarios were recorded with two different points of view, see Figure 21.

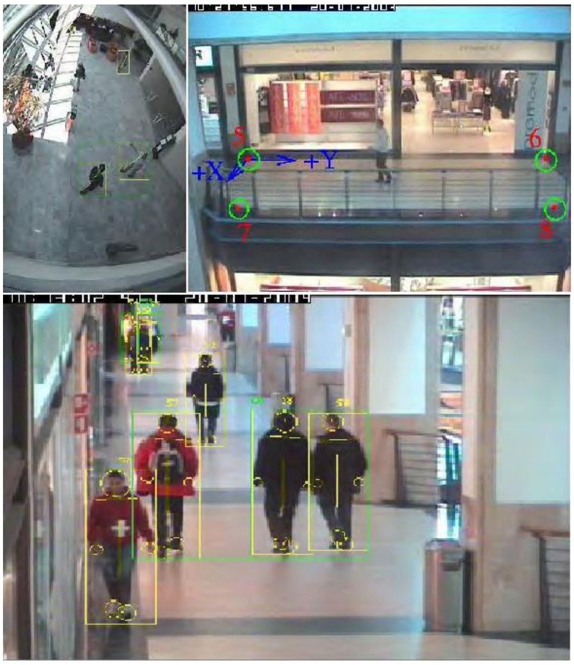


Figure 21. Sample frames for the CAVIAR dataset.