TEC2014-53176-R HAVideo (2015-2017)

High Availability Video Analysis for People Behaviour Understanding

D3.1 Online adaptive people behaviour understanding based on contextual and quality information

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

AUTHORS LIST

Juan Carlos San Miguel Avedillo	juancarlos.sanmiguel @uam.es
Fulgencio Navarro	Fulgencio.navarro@uam.es

HISTORY

Version	Date	Editor	Description
0.1	05/03/2017	Juan C. SanMiguel	First version
0.2	21/03/2017	Fulgencio Navarro	Contributions
0.3	25/03/2017	Juan C. SanMiguel	Final Working Draft
1.0	28/03/2017	José M. Martínez	Editorial checking

CONTENTS:

1. Introduction

This document summarizes the work done so far for the task T3.1 “Adaptive approaches” (WP3 “Self-configurable approaches for long-term analysis”), whose goal is to analyze alternatives to include contextual and quality information in the developed algorithms to adapt their operation to the changing environment/conditions. Adaptation would be targeted at three different levels: model, algorithm configuration and processing strategy.

This task T3.1 depends upon developments within WP2 (T2.1 Analysis tools for human behavior understanding, T2.2 Contextual modeling and extraction and T2.3 Quality analysis). The results of this task T3.2 will provide self-configurable approaches for long-term analysis and WP4 Evaluation framework, demonstrators and dissemination.

Here we define *adaptation* where a single entity (e.g. algorithm) adjust some of its parameters according to various indicators based on quality signals or contextual information. We differentiate from *collaborative* where a process in which various entities (e.g. algorithms) interact to achieve a common goal.

1.1. Document structure

The document is structured in the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: description of the contributions
- Chapter 3: Conclusions and future work

2. Contributions

This chapter compiles the contributions developed in the scope of the task T3.1.

2.1. People detection based on context

We propose a novel approach for part-based people detection in images that uses contextual information. Two sources of context are distinguished regarding the local (neighbour) information and the relative importance of the parts in the model. Local context determines part visibility which is derived from the spatial location of static objects in the scene and from the relation between scales of analysis and detection window sizes. Experimental results over various datasets show that the proposed use of context outperforms the related state-of-the-art.

We include contextual information in DPMs [1]. Each part p is represented by a 3-tuple with the appearance model, the deformation model and the optimum location of the part. Detecting people in a MXN image I involves computing a score s for hypothesised locations of all parts, for each spatial position (x; y) and analysis scale a. This work extended DPMs to use the context of each hypothesis via contextual part scores where the scene knowledge of each part is defined. The score s for each hypothesis is computed by considering the scene knowledge (see the following figure) and the possible locations of the parts of the person.

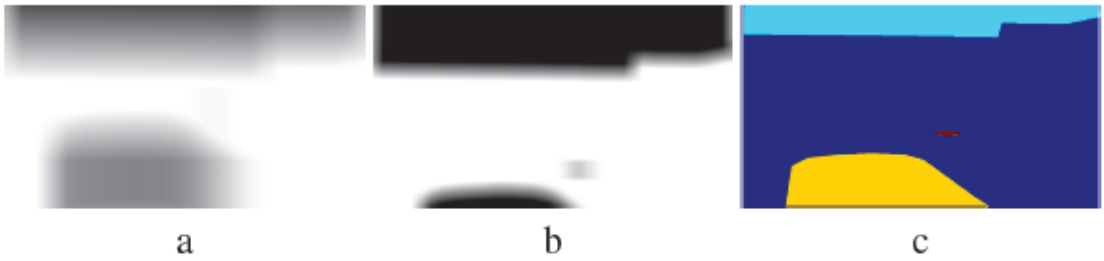

Figure 1. Block diagram of the proposed approach to combine four trackers. Examples of scene context for EDds dataset, using twice the original scale**[1]**. For the part maps, values range from 1 (white) to 0 (black). (a) Root body part (b) Head body part (c) The annotation of all stationary scene objects (each one as a unique colour)

We consider two local contexts that explore spatial neighbourhood to determine parts visibility and, therefore, their importance when combined in DPMs. First, we define context according to the detection scale a. Parts of the model may fall outside of the image I at certain locations and scales, thus decreasing detection performance as these parts are not visible. Second, we also estimate local context from domain knowledge descriptions such as the static scene objects [2][3], which are combined with spatial constraints into semantic rules in an ontology framework[2]. For example, some detections may be avoided such as for legs in the ceiling of a scene, heads in the floor of a scene or body parts occluded by tables. If we assume that this view-dependent context does not change over time, it can be applied to video monitoring with static cameras. Otherwise, context needs to be updated accordingly.

Moreover, a demonstrator of this work has been generated as part of the degree thesis of Carlos Chaparro Pozo (advisor: Álvaro García), for the Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

The following results are extracted from the associated publication:

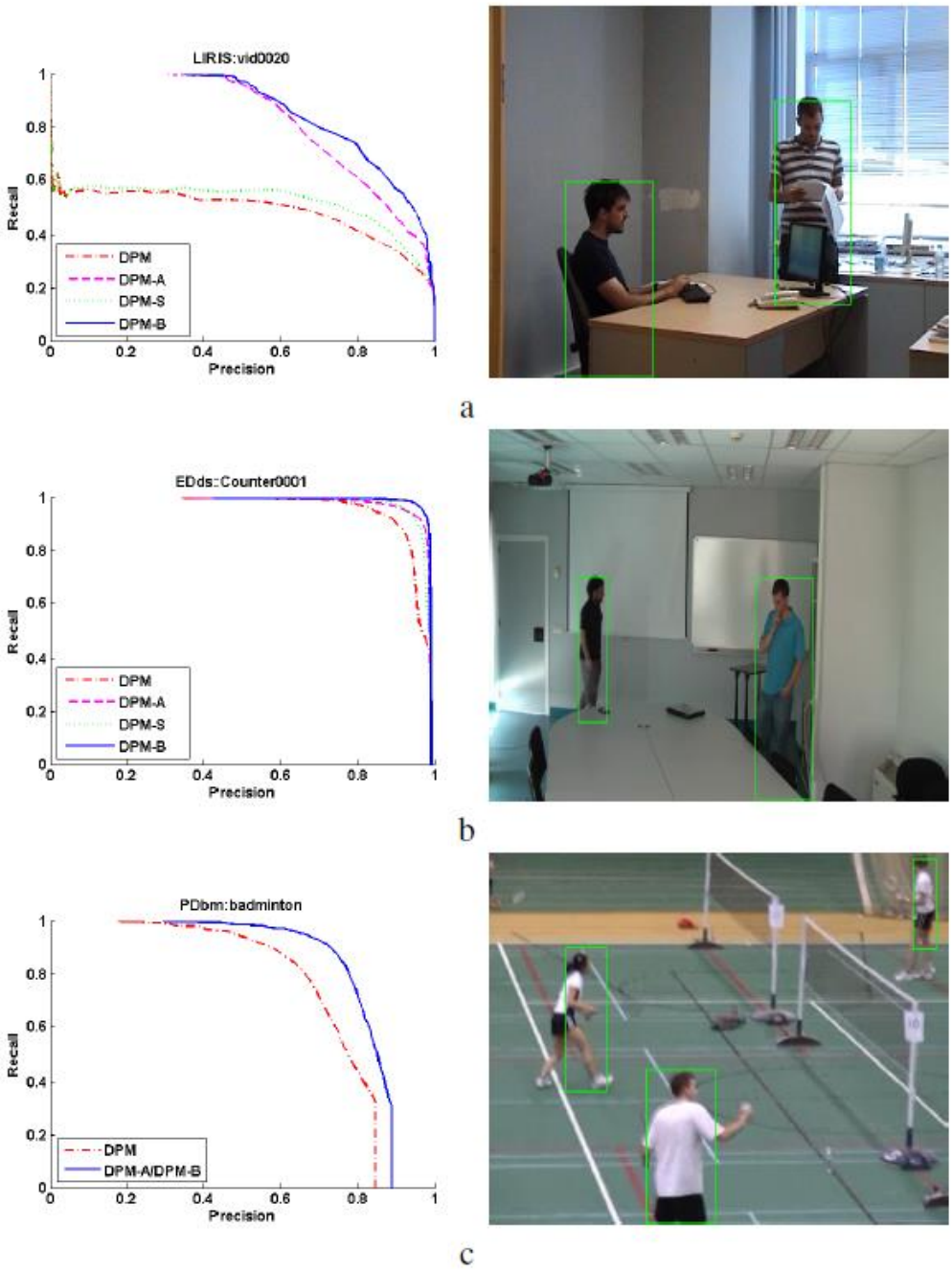

Figure 2. Block Comparative results between selected and proposed approaches using PR curves (left) and detected bounding boxes by DPM-B (right)
a Frame 18 of vid0020 sequence (LIRIS dataset)
b Frame 1238 of Counter0001 sequence (EDds dataset)
c Frame 19 of badminton sequence (PDbm dataset)

Dataset	HOG [1]	ACF-I [2]	ACF-C [2]	DPM [3]	DPM-B	%Δ
LIRIS	46.9	66.9	59.5	67.2	**86.1**	28.1
EDds	83.5	93.8	73.8	94.4	**98.3**	4.1
PDbm	48.2	73.4	60.5	75.1	**77.6**	3.3
Mean	59.5	78.0	64.6	78.9	**87.3**	10

Table 1. Detection results for each dataset in terms of AUC-PR. %A is the percentage increase of DPM-B against the best approach.

2.2. People detection based on adaptive scale selection

The main goal of this project consists of detecting the occluded persons in groups who are usually not detected. To achieve this goal, we use the previously proposed "Hierarchical detection of persons in groups" in the VPULab[3]. A hierarchy of persons in groups, where the detection of the most visible person could help to detect the occluded ones, and a hierarchy of body-parts, which main principle is to use the body-parts with the most useful information.

In addition, the main contribution of this project the design and implementation of a self-configurable variation of the original detector of persons in groups. Firstly, an exhaustive evaluation of the different configuration parameters of both proposed hierarchies have been done. After this evaluation, the most suitable parameters have been used to design and evaluate the performance of the proposed self-configurable approach. The chosen parameters were the scale of the persons in the group and the body parts configurations.

The results show clearly how the use of the proposed self-configuration approach can maintain similar results as the original approach but at the same time reducing the computational cost avoiding the computation of all the possible scales and body parts configuration in every frame. Figure 3 shows an example of scales distribution for a sequence, we can maintain the same results computing around 30% of the scales.

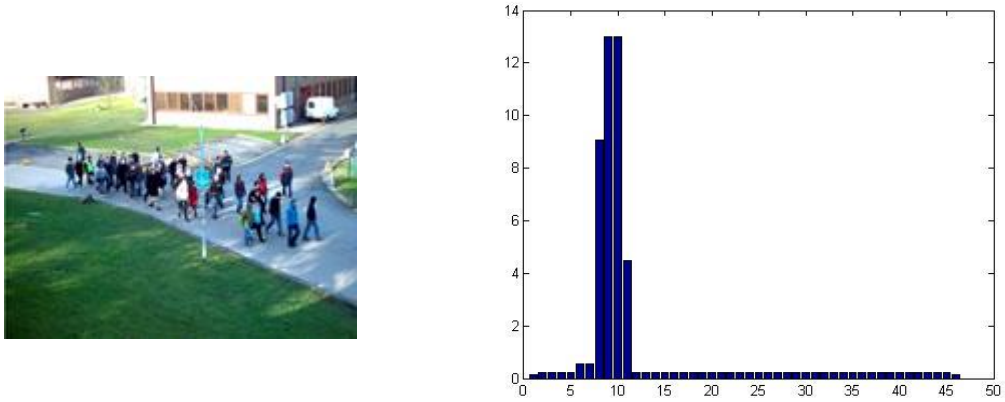

Figure 3. Example of scale distribution over sequence PETS2009-S2L3

2.3. Video tracking based on dual RGB-D models

Visual object tracking in wide baseline scenarios (VOT-WB) is a challenging task. As shown in recent surveys and contests [4], discriminative strategies are ranking top in VOT-WB.

However, the discriminative capacity of those algorithms is biased by the space where their features are built. Even algorithms able to overcome this limitation must maintain a trade-off between discriminativeness and repetitiveness to handle target self-variations.

Our approach, SP-D, is built on features extracted in low-correlated spaces, i.e. color (RGB) and depth. Self-variations on the target are less likely to be shown in both spaces simultaneously, so high-discriminative features are proposed, not at the cost of repetitiveness.

The proposal combines spatial-color characterized with superpixels, with spatial-depth information using weighted-confidence maps. Figure 4 shows the proposal overview scheme.

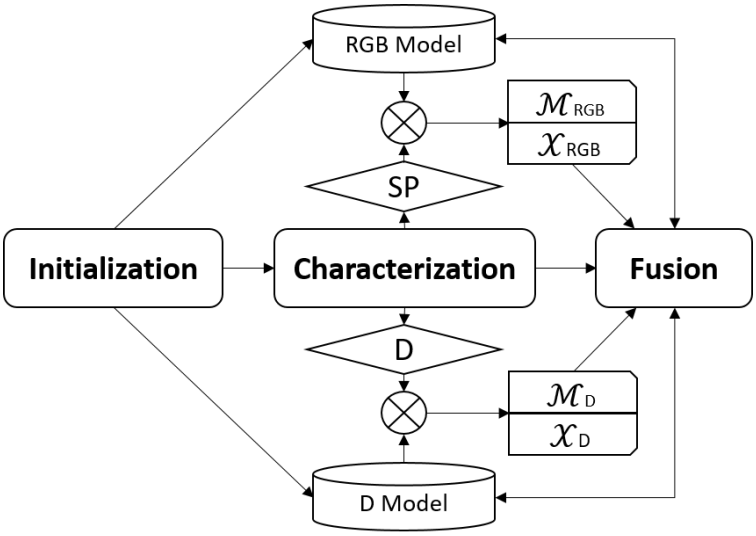

Figure 4: Algorithm overview. Three main stages, Initialization, Characterization and Fusion, are shown in bold in the mid row.

RBG model is a set of SLIC superpixels [5] extracted in the first frame via the spatial continuity theory of the background presented in [6]. The same technique is used to generate the D model using the gray-levels information of the depth channel.

In the characterization stage, superpixels and grey-levels information are extracted frame by frame, and a confidence map per space is generated using the models. An example of confidence map in the RGB spaces is show in Figure 5.

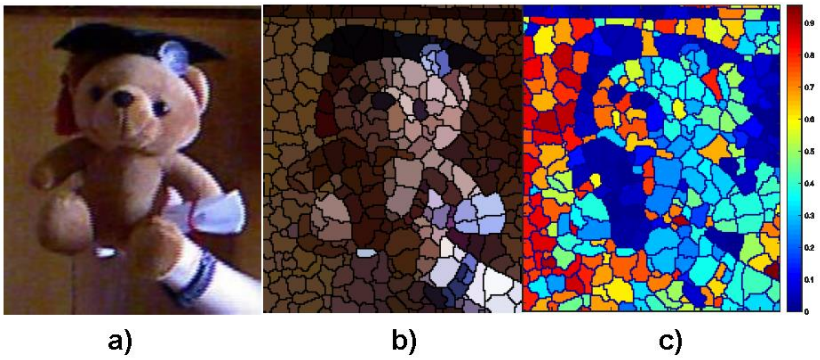

Figure 5. Characterization process in RGB space. a) RGB input. b) SLIC superpixels characterization. c) Confidence map.

Experimental evaluation sufficiently supports initial hypothesis even through most challenging situations, see Figure 6 for details of the dataset.

Figure 6. Proposed evaluation dataset.

Figure 7 shows results of the evaluation, where our proposal, SP-D, overcomes state-of-the-art tracking algorithm evaluated. Occlusion challenges are solved, Figure 7 b), using the RGB space, whereas other challenges as camouflage, Figure 7 a), are managed by depth space.

Obtained results demonstrates that combining non-correlated feature spaces results in a robust tracking algorithm.

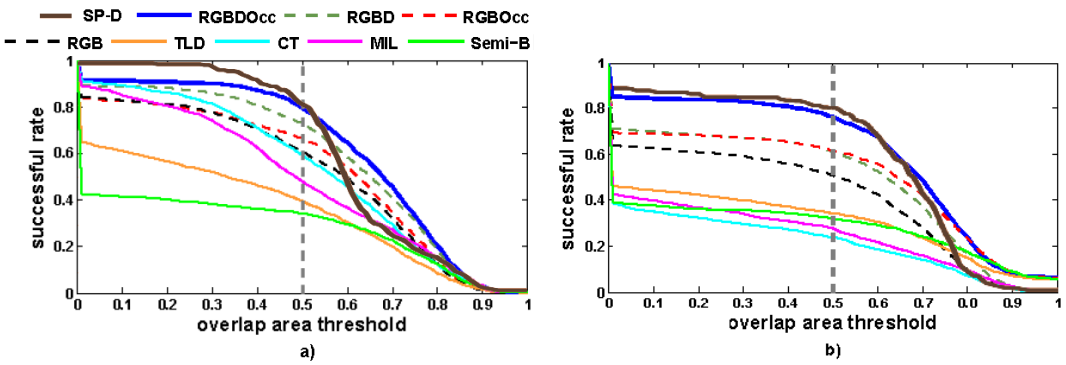

Figure 7. Success rate vs overlap area threshold for a) Non-occlusion sequences; and b) Occlusion sequences.

2.4. Abandoned object detection robust to illumination changes

We present a block-wise abandoned object detection algorithm to operate under sudden illumination changes. First, image blocks are grouped via statistical variation of pixels ratios, while discarding those blocks related to moving objects. Then, spatio-temporal stability changes of the most repeated clusters at regular sampling instants provide candidates for abandoned objects. Subsequently, entropy theory is used to detect sudden illumination changes and filter erroneously detected candidates. Finally, a People History Image is used to filter stationary pedestrians and refine the abandoned object set. Unlike previous work, robustness against sudden and gradual illumination variations and stationary pedestrians is achieved without foreground segmentation. The experimental work validates the performance of the proposed approach against related work.

This work has been published in:
Sergio López, Diego Ortego, Juan Carlos Sanmiguel, Jose M. Martinez, "Abandoned Object Detection under Sudden Illumination Changes", Actas del XXXI Simposium Nacional de la Unión Científica Int. de Radio - URSI 2016, Madrid, Spain, Sept. 2016

The proposed approach detects abandoned objects without using BS (see the following figure) based on [7][8]. A block-wise online clustering of the scene detects spatio-temporal stability changes at regular sampling instants. Those changes are exploited to identify abandoned object candidates. First, a Block Division stage decomposes each frame It into non-overlapping NxN blocks B^b_t (N = 16) at each instant t, where b denotes the block location. Second, an Online Block Clustering stage robust to gradual scene changes models each location b over time, updating a cluster partition Lb that groups each incoming non-moving block B^b_t . Third, an Abandoned Object candidates stage computes an initial set Ds of abandoned objects, where s defines the sampling instant each k = 50 frames. Data associated to the last stable cluster Sb, old stable clusters Ob and the alarm time T is used to respectively detect the spatio-temporal stability changes, discard those changes caused by previously visualized clusters (i.e. empty scene or previous detections) and detect potential abandonment for changes longer than the alarm time. Fourth, as the Online Block Clustering is not robust no sudden illumination changes, image luminance entropy Ht variation along time is used to handle such situations. Finally, a pedestrian detector is used to compute a Pedestrian History Image PHIp, where p denotes a pixel location, to determine stationary pedestrians and refine the abandoned object set Ds. The last two stages improve the state-of-the-art by refining the abandoned object candidates and provide a result image As.

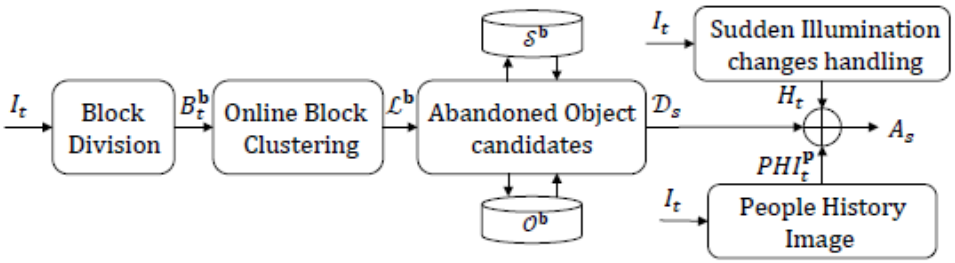

Figure 8. Block diagram of the proposed approach to detect abandoned objects.

2.4.1. Improvement 1: sudden Illumination Changes handling

Inspired by [9], we make use of entropy theory to recognize sudden illumination changes and avoid the detection of erroneous abandoned candidates (see Abandoned candidates refinement I from Figure 2). Based on this theory, dark (bright) images are characterized by low (high) entropy values due to poor (enough) luminance in pixel values. Therefore, image entropy over time is suitable to detect sudden changes in the illumination. The entropy Ht is defined as:

$$H_t = -\sum_{l=l_{min}}^{l_{max}} pdf\,(l) \cdot \log\,(pdf\,(l))\,,$$

where lmin (lmax) and pdf (l) are, respectively, the minimum (maximum) luminance level and the probability density function of each luminance level l in frame It. Note that pdf (l) is computed as the normalized histogram of the image luminance. Given the entropy value of each frame, temporal variation of such value is used to detect sudden illumination changes as:

$$\mathbb{I}_t = \begin{cases} 1 & if \quad |H_t - H_{t-1}| > \alpha \\ 0 & otherwise \end{cases},$$

where α is a threshold set to 0.05 as in [9]. Therefore, in case of sudden illumination change It = 1, whereas when no change occurs It = 0. The following figure (a) shows a sudden variation in the illumination and (b) depicts the associated entropy value that experiments a high variation in such instant (before frame 1000). As sudden illumination changes may spread across several frames, it is automatically set to 1 for the following k frames of a sudden illumination change detection.

Then, It = 1 is used to discard all detections from Ds when this condition is obtained from the last sampling instant to the current one, thus avoiding triggering false alarms induced by sudden illumination changes.

Figure 9. Example of sudden illumination change. (a) Image captures before and after the change (frame 1000) and (b) associated entropy value Ht.

2.4.2. Improvement 2: Pedestrian History Image

Abandoned object candidates 4 contain false detections produced by stationary pedestrians. To filter such detections, we apply a History Image framework [7] based on a pedestrian detector. First, a pedestrian map (PM) is computed using [1], where bounding boxes of people are marked as 1 and the remaining areas as 0. Note that bounding boxes are extended in all directions by a factor of 0.5 as objects close to pedestrians are not considered of interest. Then, the pedestrian map is accumulated over time to compute the Pedestrian

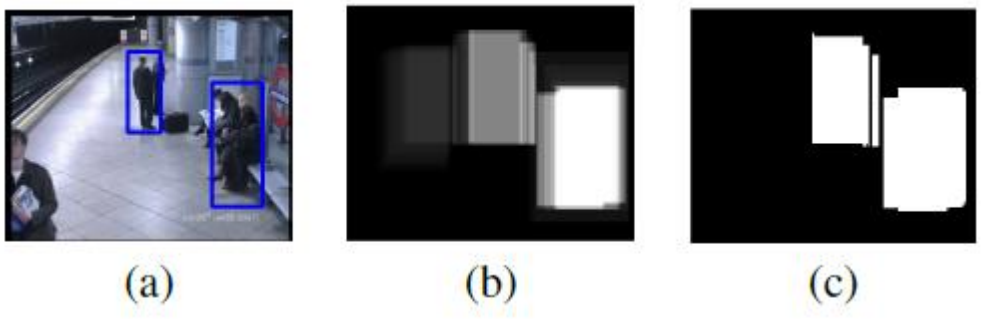

Figure 10 Example of stationary pedestrian computation. (a) Output from [1], (b) PHIp t and (c) Pp_t

2.4.3. Results

We perform two evaluations to validate both the robustness against sudden illumination changes and the capabilities to filter stationary pedestrians in typical sequences from the state-of-the-art. To evaluate the results, we compute TP and AFP that denote, respectively, correct detections and accumulated error pixels. For TP we consider every detected block that overlaps an abandoned object, while for AFP we accumulate the erroneously detected pixels over time. Moreover, the alarm time T is set 10, 30 and 60 according to the nature of each sequence.

The obtained results are extracted from the published paper:

Algorithm		Sudden Illumination Changes sequences					Stationary Pedestrians sequences					
		ABODA		I2R	LIMU	Wallflower	AVSS07					PETS06
		Video6	*Video7*	*Lobby*	*LightSwitch*	*LightSwitch*	*AB_E*	*AB_M*	*AB_H*	*PV_E*	*PV_H*	*Cam3*
[13]	GT/TP/AFP	1/1/33377	1/1/280980	0/0/20480	0/0/62768	0/0/15010	1/1/0	1/1/5632	1/1/5632	1/1/0	1/1/10	1/1/0
Proposed	GT/TP/AFP	1/1/0	1/1/90368	0/0/0	0/0/0	0/0/0	1/1/0	1/1/0	1/1/3584	1/1/0	1/1/10	1/1/0

Table 2. Comparative evaluation. GT, TP and AFP denote, respectively, Ground-truth, Correct and Accumulated error pixels. The proposed approach achieves best results against both sudden illumination changes and stationary pedestrians

The following figure depicts some examples of the detections achieved.

Figure 11. Example of result images for sequences containing pedestrians.

3. Conclusions and future work

3.1. Achievements

As summary, the achievements, so far, of task 3.1 are:

- Development of a people detection algorithm based on contextual information (scene knowledge about object types and their locations).
- Development of a people detection algorithm based on adaptive selection of scales to detect people. Such selection is based on previous information.
- Development of a single-target video tracking algorithm able to quantify the importance of features to adapt to target or scene changes over time.
- Development of an approach for abandoned object detection able to counteract stationary people and adapt to illumination changes.

3.2. Future work

As future work, we will focus on the following:

- Improvement of Background Subtraction algorithms based on stand-alone evaluation
- Improvement of multi-target tracking algorithms based on stand-alone evaluation

4. References

[1] Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D.: 'Object detection with discriminatively trained part-based models', IEEE Trans. Pattern Anal. Mach. Intell., 2010, 32, (9), pp. 1627-1645

[2] SanMiguel, J. and Martínez, J.: 'An ontology for event detection and its application in surveillance video', IEEE Int. Conf. on Advanced Video and Signal-based Surveillance, 2009, pp. 220-225

[3] Alvaro Garcia-Martin, Ricardo Sanchez, Jose M. Martinez: "Hierarchical detection of persons in groups", Signal, Image and Video Processing, (Accepted February 2017), ISSN 1863-1711.

[4] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," IEEE Transactions on PAMI, vol. 36, no. 7, pp. 1442–1468, 2014.

[5] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," IEEE Transactions on PAMI, vol. 34, pp. 2274–2282, 2012.

[6] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking," in IEEE In ICCV. IEEE, 2011, pp. 1323–1330.

[7] D. Ortego and J. SanMiguel, "Multi-feature stationary foreground detection for crowded video-surveillance," in Proc. of IEEE Int. Conf. on Image Processing, 2014, pp. 2403–2407.

[8] D. Ortego, J. SanMiguel, and J. Mart´ınez, "Long-term stationary object detection based on spatio-temporal change detection," IEEE Signal Processing Letters, vol. 22, no. 12, pp. 2368–2372, 2015.

[9] F. Cheng, S. Huang, and S. Ruan, "Illumination-sensitive background modeling approach for accurate moving object detection," IEEE Trans. on Broadcasting, vol. 57, no. 4, pp. 794–801, 2011