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 (December 2017)

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
Diego Ortego	Diego.ortego@uam.es
Marcos Escudero	Marcos.escudero@uam.es
Alvaro Garcia Martín	Alvaro.garcia@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
1.1	28/11/2017	Juan C. SanMiguel	First Working Draft v2
1.2	01/12/2017	Diego Ortego	Contributions
1.3	10/12/2017	Álvaro García	Contributions
1.4	15/12/2017	Marcos Escudero	Contributions
1.5	19/12/2017	Juan C. SanMiguel	Final Working Draft v2
2.0	23/12/2017	José M. Martínez	Editorial checking

CONTENTS:

1. Introduction

This document summarizes the work during the first and half year(s) of the project for the task T3.1 “Adaptive approaches” (WP3 “Self-configurable approaches for long-term analysis”).
whose goal is to analyse 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.

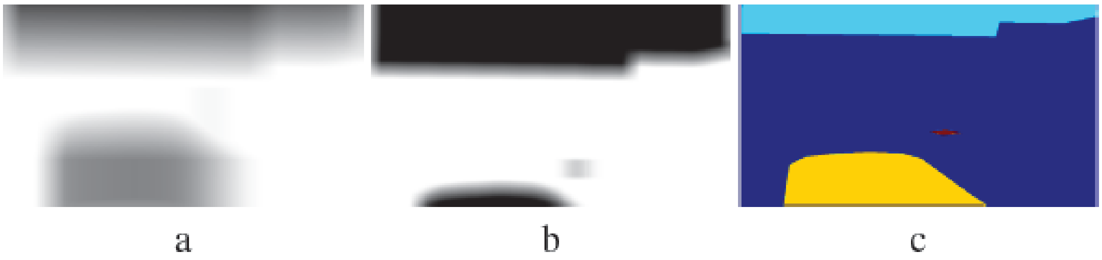


Figure 1. Block diagram of the proposed approach to combine four trackers. Examples of scene context for EDds dataset, using twice the original scale**[1]**. For the part maps, values range from 1 (white) to 0 (black). (a) Root body part (b) Head body part (c) The annotation of all stationary scene objects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

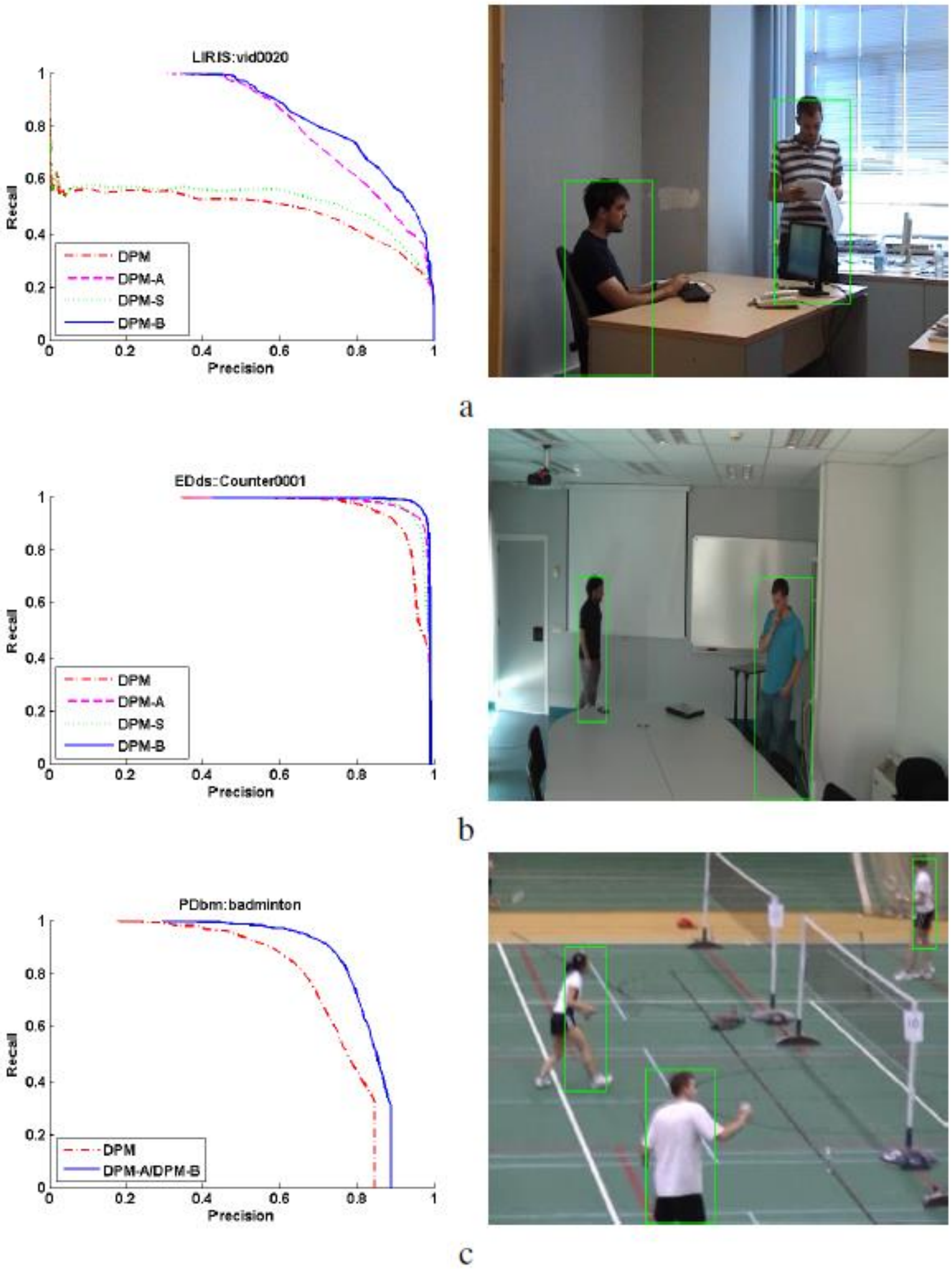


Figure 2. Block Comparative results between selected and proposed approaches using PR curves (left) and detected bounding boxes by DPM-B (right)
a Frame 18 of vid0020 sequence (LIRIS dataset)
b Frame 1238 of Counter0001 sequence (EDds dataset)
c Frame 19 of badminton sequence (PDbm dataset)

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.

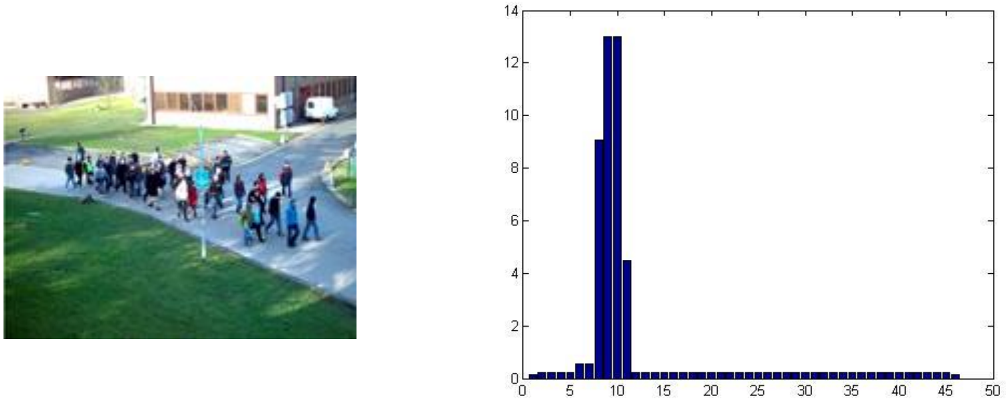


Figure 3. Example of scale distribution over sequence PETS2009-S2L3

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.

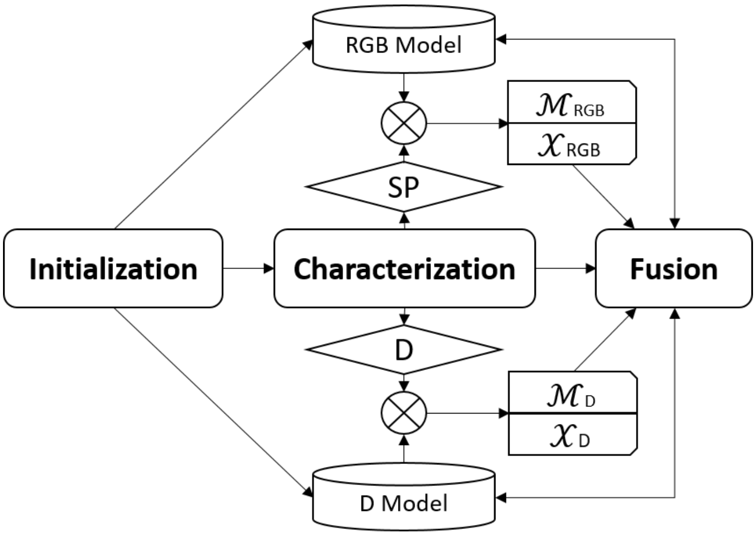


Figure 4: Algorithm overview. Three main stages, Initialization, Characterization and Fusion, are shown in bold in the mid row.

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.

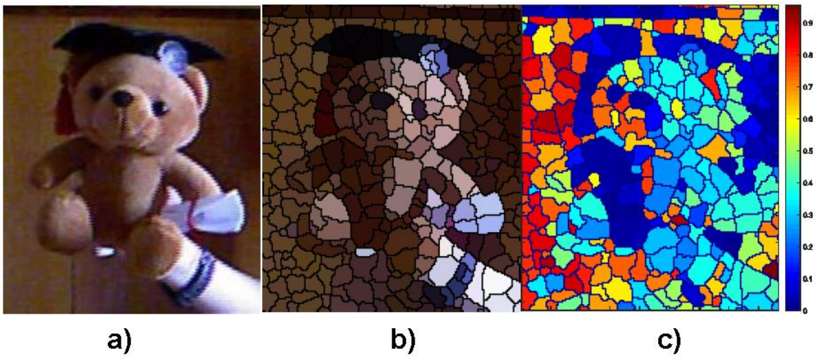


Figure 5. Characterization process in RGB space. a) RGB input. b) SLIC superpixels characterization. c) Confidence map.

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

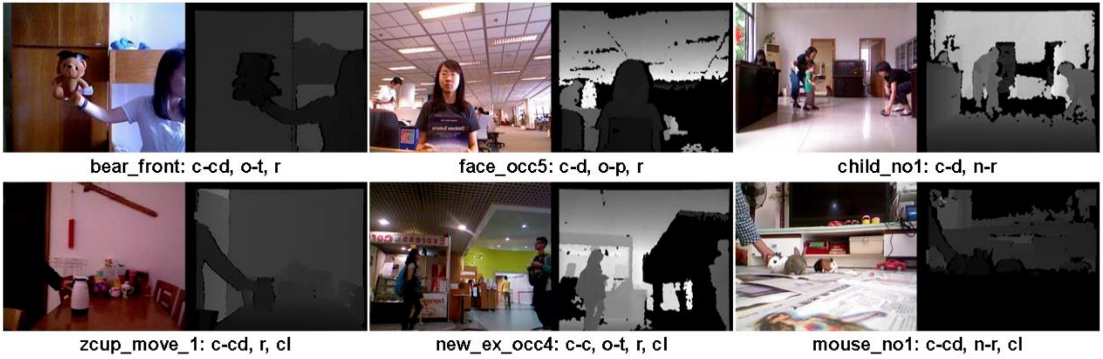


Figure 6. Proposed evaluation dataset.

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.

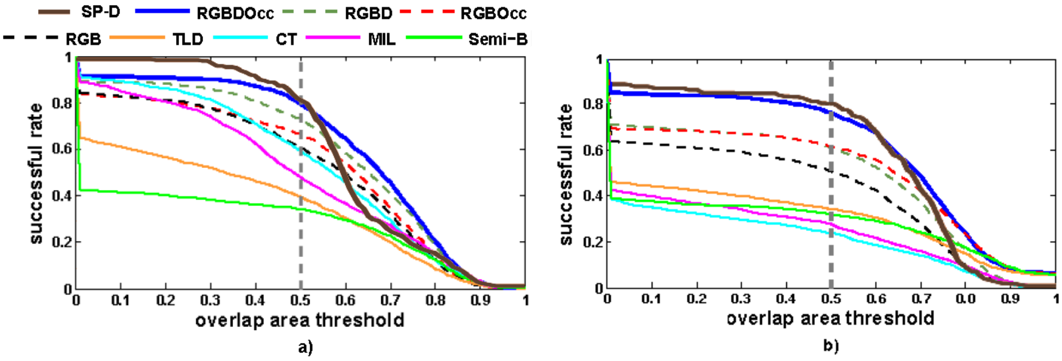


Figure 7. Success rate vs overlap area threshold for a) Non-occlusion sequences; and b) Occlusion sequences.

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.

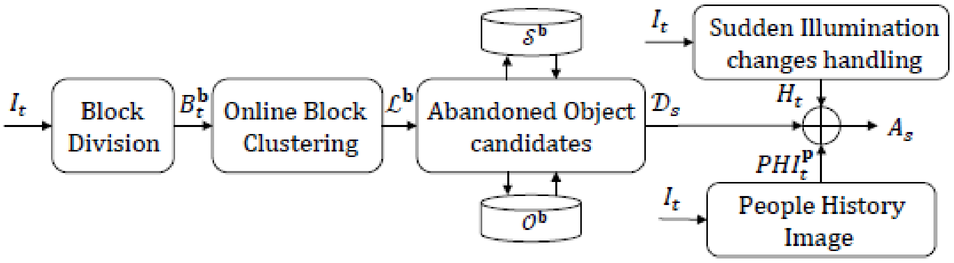


Figure 8. Block diagram of the proposed approach to detect abandoned objects.

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.

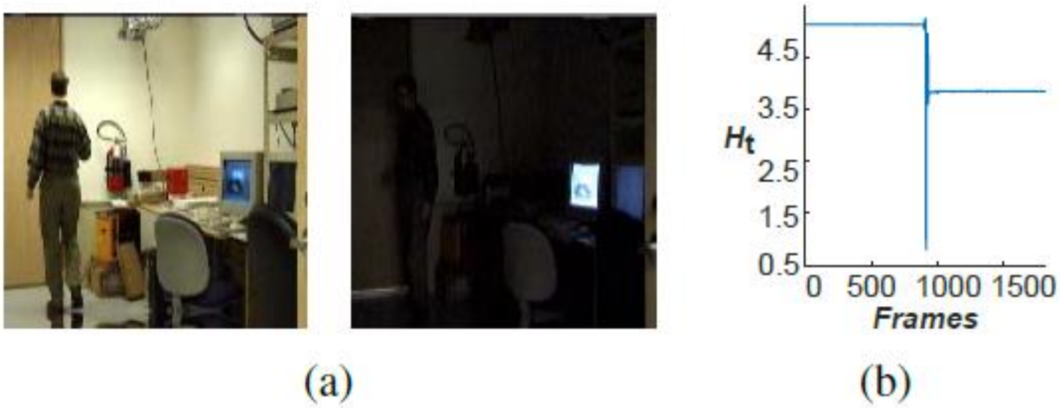


Figure 9. Example of sudden illumination change. (a) Image captures before and after the change (frame 1000) and (b) associated entropy value Ht.

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

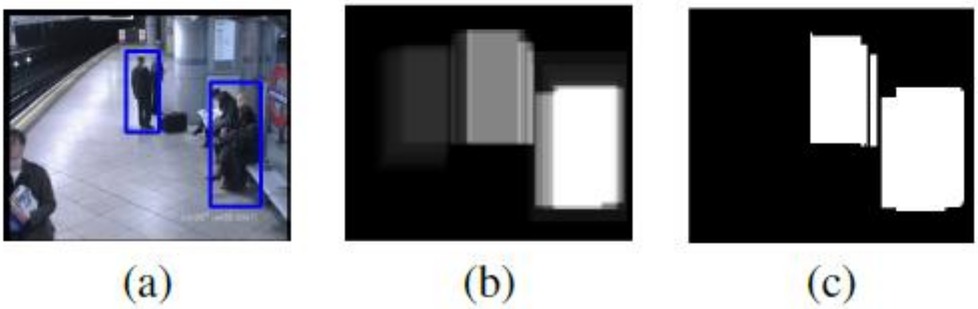
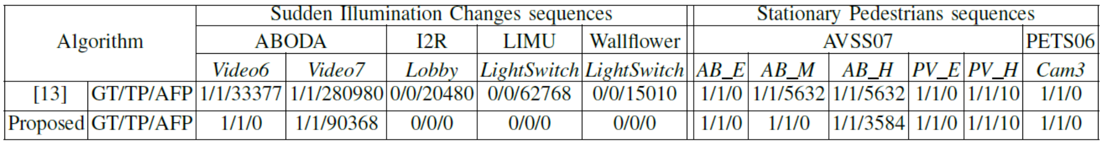


Figure 10 Example of stationary pedestrian computation. (a) Output from [1], (b) PHIp t and (c) Pp_t

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.

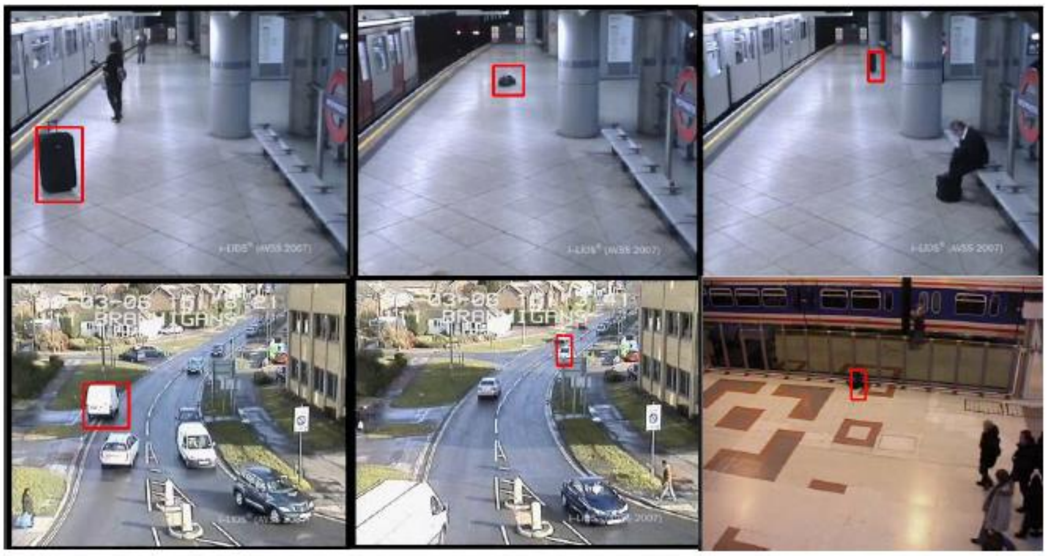


Figure 11. Example of result images for sequences containing pedestrians.

2.5. Foreground segmentation post-processing through quality information

Video object segmentation is a popular low-level task in computer vision which aims to segment the objects of interest or foreground in a video sequence. In particular, we focus on scenarios with a relative control of camera motion, where video object segmentation is tackled through background subtraction algorithms [10].

Improving background subtraction performance has been mainly addressed either by selecting a better model or better features. Alternatively, foreground segmentation masks can be also improved adopting post-processing techniques to either remove false positives or recover false negatives [11]. These techniques represent a very interesting alternative as they can be performed independently of the algorithm, thus avoiding the complex task of modifying features or models that are inherent to each algorithm.

In the literature, performance improvement through post-processing has been mainly addressed using morphological operations [12] and inspecting generic foreground mask properties [13][14] to filter erroneous foreground and expand to undetected areas. Among these approaches, post-processing ones provide independence of specific phenomena (e.g. illumination or shadows) and, unlike morphological operations, introduces complementary information to the foreground mask. In this context, in our previous work [15] we analysed a set of foreground mask properties to estimate ground-truth performance or quality, concluding that the fitness between foreground mask and segmented image regions (fitness-to-regions) has a great potential for such task. Therefore, in this work we improve foreground segmentation masks based on this finding.

The post-processing framework proposed receives an image and the foreground mask computed by an algorithm and computes an improved foreground mask. To that end, we first compute a hierarchy of image segmentations (i.e. a set of segmentations with different degrees of detail). The coarser the level the higher the regions, thus enabling to cover complete or large object areas. However, to prevent merging foreground and background regions in each hierarchy level, we introduce motion constraints through the optical flow. Then, we estimate a foreground quality image for each hierarchy level using a fitness-to-region property, thus obtaining a hierarchy of qualities. Subsequently, we combine all the quality images, thus obtaining a unique collaborative quality across image levels. Finally, we use a Conditional Random Field (CRF) to obtain the improved foreground mask through an optimal labelling process that considers both foreground quality and spatial information. Please, see Figure 12 where we show an example of improvement.

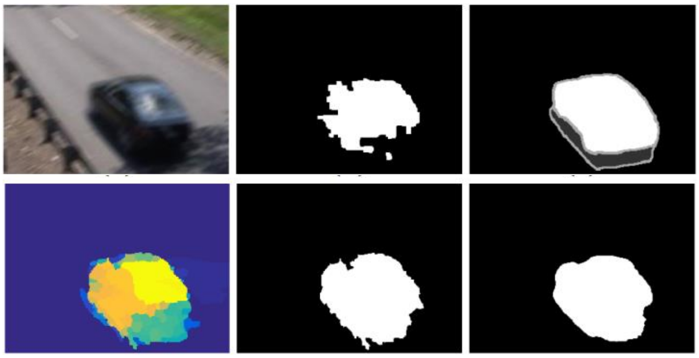


Figure 12. Example of the post-processing framework. From top to bottom and left to right: Image under analysis, foreground segmentation mask to improve, ground-truth, collaborative quality, improved foreground obtained thresholding the quality and the improved foreground mask obtained by our proposed framework after using a CRF.

To validate the effectiveness of the proposed approach we use four well-known datasets from the literature [16][17][18][19] and 17 different algorithms. We do not provide more details as this work is currently under review in a high impact journal.

2.6. Guided video object segmentation using convolutional neural networks

In contrast with the previous work, we focus on unconstrained scenarios to perform video object segmentation (VOS). In these environments [20], VOS is cast as detecting spatio-temporal relevant objects, propagating initially segmented objects or using frame-by-frame human intervention, respectively, for unsupervised, semi-supervised and supervised algorithms. These situations present challenges related to camera motion, shape deformations of objects or motion blur.

The VOS task can be formulated as a pixel-wise labelling process where the aim is to assign to each pixel of a video frame a foreground or background label in each temporal instant. In this sense, we overcome the labelling process through convolutional neural networks (CNNs) by defining an architecture that uses a RGB colour frame together with its corresponding optical flow and a foreground segmentation of an external algorithm (from now on, the external foreground mask), to compute a new foreground segmentation (see Figure 13).

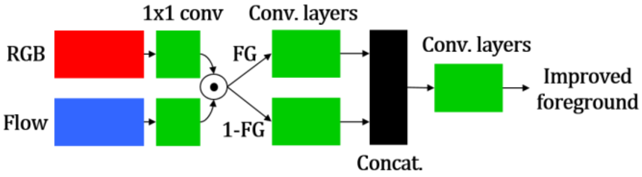


Figure 13. Proposed arquitecture for video object segmentation using external algorithm results (FG) to compute an improved foreground. There a

The novelty of our proposed approach lies in the use of pre-computed results from an external VOS algorithm (FG) as a powerful information to complement a VOS-CNN model to both improve the external foreground mask that the CNN-VOS model would compute without making use of FG. Our CNN-VOS approach is comprised by three elements: an appearance network that encodes the video frame (red in Figure 13), a motion network that encodes the optical flow (blue in Figure 13) and a decoder (green in Figure 13) that takes both encodings and uses the external foreground mask to weight them, thus introducing an attention mechanism to enforce certain responses. The intuition behind using two networks in the encoding is to benefit from the complementary information that appearances and motion reveals about moving objects [21]. Note that the architecture used for both the appearance and the motion network is PSPNet [22] using ResNet-101. In particular, we select the 512 features after the convolutional layer that follows the pyramid parsing module.

We train the network in three steps. First, we train the appearance network in PASCAL VOC dataset to perform semantic segmentation. Second, we train the motion network in the data provided at [23] to perform segmentation based on optical flow. Third, we train the decoder in a subset of DAVIS dataset freezing the appearance and the motion network.

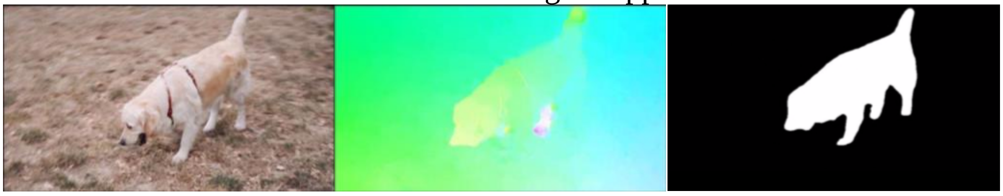


Figure 14. Example of video object segmentation. From left to right: colour image, color-coded optical flow and the improved foreground mask.

We evaluate the proposed approach in the evaluation set of DAVIS 2016 dataset [20]. We do not provide more details as this work is publication pending

2.7. Semantic-Constrained Multi-Camera Pedestrian Detection

We have developed an approach for multi-camera multi-target detection that employs semantic maps to collaboratively improve detections across multiple cameras viewing the same area. These semantic maps provide useful information of the monitored environment (e.g. location of areas such as the floor, road, …) which is applied to define and apply constraints over provided detections by each camera.

Automatic people detection can be considered a solid and mature technology able to operate on detection rates over 90% in most video surveillance scenarios. However, the handling of severe-occlusions is still a major challenge to be addressed. Occlusions between people occur due to the projection of the 3D world onto a 2D representation. Although recent deep-learning based methods are able to cope with moderate occlusions, the detection process fails when only a small part of the person is visible.

An interesting alternative for some scenarios is the use of additional cameras, moving from a mono-camera to a multi-camera scenario. If the cameras are adequately positioned, the additional cameras might provide disambiguation in severe occlusion situations. However, the use of additional cameras implies new challenges. Among the striking ones are calibration problems, persons' self-occlusions and the requirement to set—usually manually annotated— operational areas on which to fuse detections.

To face these challenges, in this study we propose a novel multi-camera approach to fuse detections from several cameras on a common plane based on graph theory. It is also combined with a modification of and height adaptive algorithm from the state-of-the-art which uses semantic segmentation to improve the location of people detections. Additionally, semantic information is also used to automatically generate the operational area given a scene. The flowchart of the proposed approach is depicted in the following Figure. Experimental results suggest that the proposed fusion method outperforms existing strategies in widely used multi-camera data-sets.

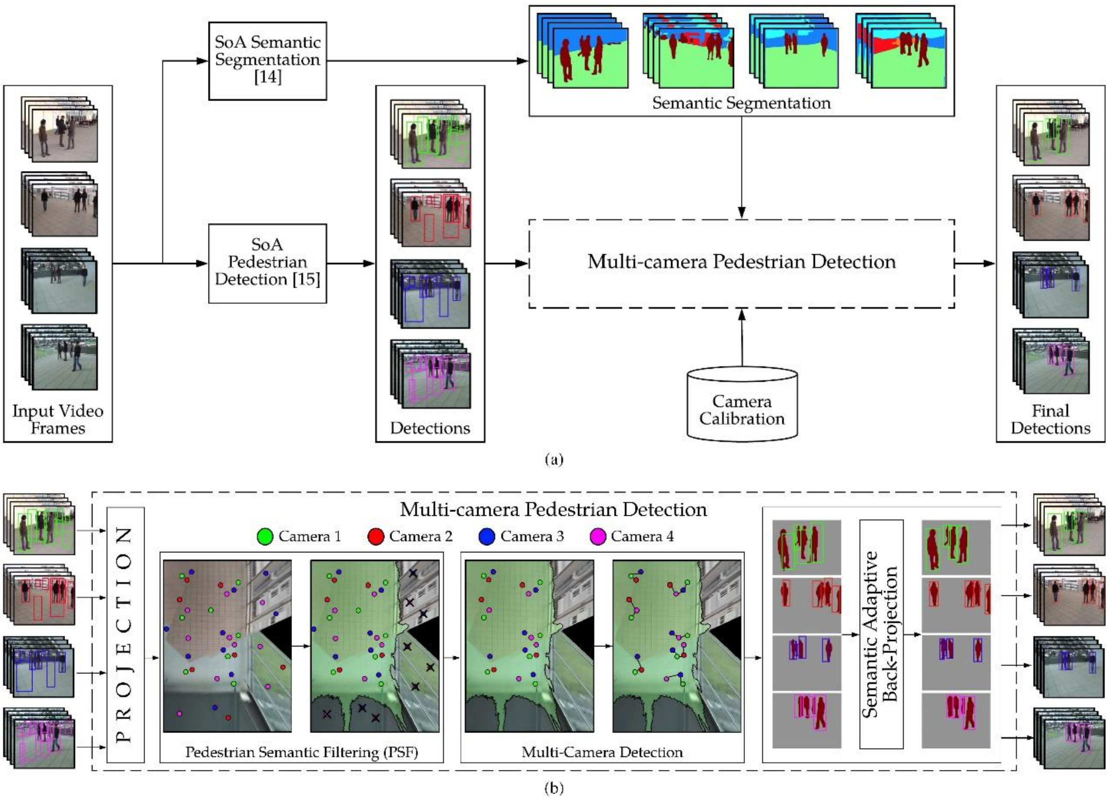

SoA Semantic Segmentation [14]
Semantic Segmentation
Input Video Frames
SoA Pedestrian Detection [15]
Detections
Multi-camera Pedestrian Detection
Camera Calibration
Final Detections
(a)
Multi-camera Pedestrian Detection
Camera 1
Camera 2
Camera 3
Camera 4
PROJECTION
Pedestrian Semantic Filtering (PSF)
Multi-Camera Detection
Semantic Adaptive Back-Projection
(b)

Figure 15. Flowchart of the proposed method. (a): Processing starts by extracting both semantic segmentation information and pedestrian detections from a set of different overlapping cameras. Mono-camera people detections, semantic segmentation, and camera calibration feed the Multi-camera Pedestrian Detection module detailed in part (b). (b): Detections are projected to a common ground plane. An automatically generated evaluation area is derived to filter detections through a Pedestrian Semantic Filtering (PSF) module. The Multi-Camera Detection module is used to merge detection by means of the use of spatially-constrained graphs. The Semantic Adaptive Back-Projection module is then applied to refine detections by an iterative height and position adjusting on the semantic pedestrian masks.

3. Conclusions and future work

3.1. Achievements

As summary, the achievements 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.
- Development of algorithms to filter and improve foreground segmentation masks from background subtraction algorithms
- Development of a filtering approach for detecting pedestrians using multiple-cameras.

3.2. Future work

As future work, we will focus on the following:

- Keep with the 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

[10] T. Bouwmans, “Traditional and recent approaches in background modelling for foreground detection: An overview”, Comput. Sci. Rev., vol. 11–12, pp. 31–66, 2014.

[11] D. H. Parks and S. S. Fels, "Evaluation of Background Subtraction Algorithms with Post-Processing", in Proc. Of IEEE Int. Conf. on Adv. Video and Signal Based Surv. (AVSS), pp. 192-199, 2008.

[12] P.-L. St-Charles, G.-A. Bilodeau and R. Bergevin, "SuBSENSE: A Universal Change Detection Method With Local Adaptive Sensitivity", IEEE Trans. Image Process., pp. 359-373, 2015.

[13] A. Schick, M. Bauml and R. Stiefelhagen, "Improving foreground segmentations with probabilistic superpixel Markov random fields", in Proc. of IEEE Conf. on Comp. Vis. And Pattern Recogn. Worshops (CVPRW), pp. 27-31, 2012.

[14] D. Giordano, I. Kavasidis, S. Palazzo and C. Spampinato, "Rejecting False Positives in Video Object Segmentation", in Proc. of Int. Conf. on Comp. Anal. of Images and Patterns (ICIAP), pp. 100-112, 2015.

[15] D. Ortego and Juan C. SanMiguel and José M. Martínez, "Stand-alone quality estimation of background subtraction algorithms", Comp. Vis. Image Underst. pp. 87-102, 2017.

[16] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth and P. Ishwar, "CDnet 2014: An Expanded Change Detection Benchmark Dataset", in Proc. of IEEE Conf. on Comp. Vis. and Pattern Recogn. Workshops (CVPRW), pp. 393-400, 2014.

[17] C. Cuevas and E.M. Yáñez and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA", Comp. Vis. Image Underst., pp. 103-117, 2016.

[18] S. Brutzer, B. Hoferlin and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance", ", in Proc. of IEEE Conf. on Comp. Vis. and Pattern Recogn. (CVPR), pp. 1937-1944, 2011.

[19] A. Vacavant, T. Chateau, A. Wilhelm and L. Lequièvre, "A Benchmark Dataset for Outdoor Foreground/Background Extraction", in Proc. of Asian Conf. on Comp. Vis., pp. 291-300, 2013.

[20] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation", in Proc. of IEEE Conf. on Comp. Vis. and Pattern Recogn. (CVPR), pp. 724-732, 2016.

[21] P. Tokmakov, K. Alahari and C. Schmid; in Proc. of IEEE Int. Conf. on Comp. Vis. (ICCV), 2017, pp. 4481-4490.

[22] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network", in Proc. of IEEE Conf. on Comp. Vis. and Pattern Recogn. (CVPR), pp. 6230-6239, 2017.

[23] S. D. Jain, B. Xiong and K. Grauman, "FusionSeg: Learning to Combine Motion and Appearance for Fully Automatic Segmentation of Generic Objects in Videos", in Proc. of IEEE Conf. on Comp. Vis. and Pattern Recogn. (CVPR), pp. 2117-2126, 2017.