Computer Vision and Image Understanding 133 (2015): 76-89

DOI: http://dx.doi.org/10.1016/j.cviu.2014.09.010

Post-processing approaches for improving people detection performance

Álvaro García-Martín and José M. Martínez

Abstract—Nowadays, people detection in video surveillance environments is a task that has been generating great interest. There are many approaches trying to solve the problem either in controlled scenarios or in very specific surveillance applications. We address one of the main problems of people detection in video sequences: every people detector from the state of the art must maintain a balance between the number of false detections and the number of missing pedestrians. This compromise limits the global detection results. In order to reduce or relax this limitation and improve the detection results, we evaluate two different post-processing subtasks. Firstly, we propose the use of people-background segmentation as a filtering stage in people detection. Then, we evaluate the combination of different detection approaches in order to add robustness to the detection and therefore improve the detection results. And, finally, we evaluate the successive application of both post-processing approaches. Experiments have been performed on two extensive datasets and using different people detectors from the state of the art: the results show the benefits achieved using the proposed post-processing techniques.

Index Terms—People detection, people-background segmentation, segmentation confidence map, segmentation mask, decision-level fusion, fusion methods.

1 Introduction

Within the computer vision field, particularly in the research area of digital image and video processing, there exists a rich variety of algorithms for segmentation, object detection, event recognition, etc, which are being used in security systems. People detection is one of the most challenging problems in this field. The complexity of the people detection problem is mainly based on the difficulty of modeling persons because of their huge variability in physical appearances, articulated body parts, poses, movements, points of view and interactions among different people and objects. This complexity is even higher in real world scenarios such as airports, malls, etc, which often include multiple persons, multiple occlusions and background variability.

The main contribution presented in this paper is the application of two subtasks of people detection post-processing. The first one is based on the people-background segmentation. People-background segmentation gives us information about where there are not people in the scene. We can use this information to eliminate or, at least, reduce the number of false positives. The second one is based on the combination at decision-level of multiple people detectors from the state of the art in order to take advantage of their independent strengths and at the same time reduce their drawbacks and limitations. And, finally, we also evaluate the successive application of both post-processing approaches in order to combine both improvements.

The remainder of this paper is structured as follows: Section 2 describes the related state of the art; Sections 3 and 4 describe the two different proposed approaches; Section 5 discusses the experimental results. Finally, Section 6 summarizes the main conclusions and future work.

A. García-Martín and J. M. Martínez are with the Department of Electronics and Communications Technology, Video Processing and Understanding Lab, Escuela Politécnica Superior, Universidad Autónoma de Madrid, E-28049 Madrid, Spain. E-mail: {alvaro.garcia, josem.martinez}@uam.es

2 State of the art

As discussed previously, this article is focused on people detection post-processing approaches. For this reason, the following sections include a brief state of the art of people detection and selected post-processing approaches.

2.1 People detection

Every people detection approach consists mostly of two phases: firstly, the design and training (if training is required) of a person model based on characteristic parameters such as motion, dimensions, silhouette, etc. Secondly, the adjustment of this person model to the candidates to be person in the scene. All candidates that adjust to the model will be detected or classified as person, whilst all the others will not be detected or classified as person. Therefore, these two main critical tasks of people detection (object detection and person model) determine the global detection performance.

There are two main conventional object detection approaches: one based on some kind of segmentation of the scene in foreground (objects) and background [1]–[13] and one based on an exhaustive scanning approach [14]–[42]. There are also some approaches that try to combine both approaches together [43], [44]. In any case, the result of this stage is the location and dimension (bounding

box or blob) of the different objects candidates to be a person.

In relation to the chosen person model, there are two main discriminative information sources to characterize the people model: appearance and motion. Nowadays in the existing literature, most methods are only based on appearance information or they add robustness to the detection with motion information through tracking algorithms. However, human appearance varies due to environmental factors such as light conditions, clothing, contrast, etc, apart from the huge intrinsic people variability such as different heights, widths, poses, etc. For these reasons, there are some approaches which try to avoid these factors using only motion information [1], [16].

There are many approaches that use appearance information to define the person model. This is because appearance is more discriminant than motion. We classified the appearance models according to simplified human models or complex models. There are simple person models that define the person as a region or shape, i.e., holistic models [2]–[10], [14], [15], [17]–[20], [23]–[38], [42] and more complex models that define the person as combination of multiple regions or shapes, i.e., part-based models [11]–[13], [21], [22], [25], [39]–[41], [43], [44]. Although the vast majority of approaches are mainly based on appearance information, there are some approaches that combine appearance and motion information in order to improve the detection results. Some authors combine appearance and motion expanding previous detectors based on appearance to more than one frame [14], [17], [19]; in this way they are able to easily introduce motion information in the person model and add robustness to the detector.

Lately, the most popular approaches (detection-by-tracking approaches) are those that combine detection and tracking in order to improve the detection results [2], [15], [18], [20]–[30]. In this case, the motion information is not implicitly part of the person model but it is still useful in order to filter or extrapolate detections over time. On the other hand, [28], [30] not only combine detection and tracking information but also propose the combination of two independent and implicit person models: one model based on appearance and another model based on motion.

2.2 People detection post-processing

Traditionally, the typical additional preprocessing subtasks in people detection are not oriented to one specific processing task, i.e., they are oriented to enhance/adapt/reduce the video information before being analyzed. For example: camera motion compensation, camera calibration, noise removal, etc. In turn, the typical additional post-processing subtasks are applied over the detection outcome. They are oriented to filter or verify the final detections using any additional information source. The most typical ones are those based on tracking information [25], [30], [45] which study the detections evolution over time. Other approaches use some kind of scene/contextual restriction (spatial, people size, symmetry, etc) or motion restrictions. In relation to scene restrictions, Geronimo et al. [46] describes different preprocessing subtasks with a clear focus on driver assistance systems such as exposure time, gain adjustments and camera calibration. On the other side, Eiselein et al. [47] proposes the use of motion restrictions combining people detection and optical flow in order to reduce the number of missing detections in a tracking system.

Any fusion technique attempts to combine the information from all available sources into a unified representation. This provides better information for human or machine perception as compared to any of the input sources. One of the data fusion models most commonly used in image processing applications is the three-level fusion model. It is based on the levels at which information is represented [48]. This model classifies data fusion into three levels: data or pixel-level fusion, feature fusion and decision fusion. At the lowest level, the fused pixel is derived from a set of pixels from the multiple input sources. At the intermediate level, the features for each object are independently extracted in each information source; these features crate a common feature space for object classification. Finally, at the highest level, decision-level fusion corresponds to combining decisions from several experts.

In the case of people detection, every people detector must build up some form of dense confidence map [26] (explicitly or implicitly), which consists of the continuous detection confidence score for each location and scale. Felzenszwalb et al. [41] combines or fuses the confidence map of several independent body parts at pixel-level in order to obtain a final confidence map. There are some approaches that combine or fuse more than one feature at feature-level in order to improve the detection results: [14], [17] combine appearance and motion expanding previous features based on appearance to more than one frame, whilst Gan and Cheng [49] uses the feature HOG-LBP (combination of the HOG [33] and LBP [50] features). Finally, every people detector must compare the previously defined/trained person model with the input image and make a final decision according to a similarity criterion. There are some approaches that combine or fuse multiple detectors at decision-level using multi body part detectors [39] or detectors [28].

In this work, we evaluate two new subtasks of people detection post-processing and their successive application in typical video surveillance environments (see Figure 1). The first one is based on the people-background segmentation [51]. The second one is the combination or fusion of up to six independent, appearance based people detectors at decision-level and their combination with a motion based people detector. In any case, the proposed post-processing subtasks are based only on some kind of people detection information. Therefore, they can be considered as any other additional

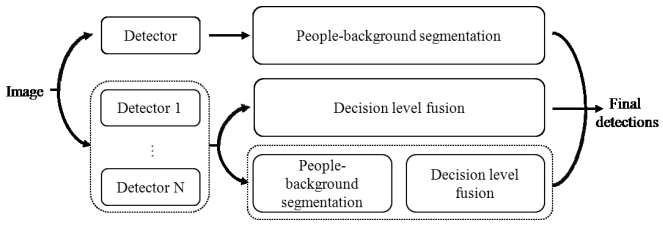


Figure 1. Block diagram of the proposed people detection post-processing configurations.

post-processing step in any people detection system, i.e., they do not interfere or are independent of any other additional improvement using tracking information, scene/contextual restrictions, etc.

3 People detection using people-background segmentation confidence

As already mentioned, every people detector from the state of the art must keep a balance between Precision and Recall rates. For this reason, the global detection performance is mainly limited by the number of possible false detections. Our main idea consists in reducing or relaxing this limitation using people-background segmentation. The proposed filtering approach has been implemented as a post-processing, but it can be used as either a preprocessing or post-processing stage.

3.1 People-background segmentation

People-background segmentation [51] is a two-class segmentation ensuring that no people or body parts are appearing in the background class. This type of segmentation is useful not only as a people detection preprocessing or post-processing step, but also for other video analysis processes such as tracking and people density estimation. While the focus of person detection approaches are on obtaining a high detection performance and on reducing false positive detections. People-background segmentation aims at determining the areas without people in the scene by giving a higher penalty to pixels incorrectly classified as background. This results in a segmentation mask with a bias on the background as opposed to a segmentation with a bias on people.

The chosen people-background segmentation method [51] uses the DTDP detector [41] in order to detect different body parts and extends this representation by appropriately grouping them. Then, they fuse detection confidence maps according to regions that are expected to be covered by the body parts. The corresponding background segmentation mask is finally generated after binarization and post-processing. Therefore, although the people-background segmentation [51] is based on [41], the objective and result are fairly different from the traditional people detection approaches.

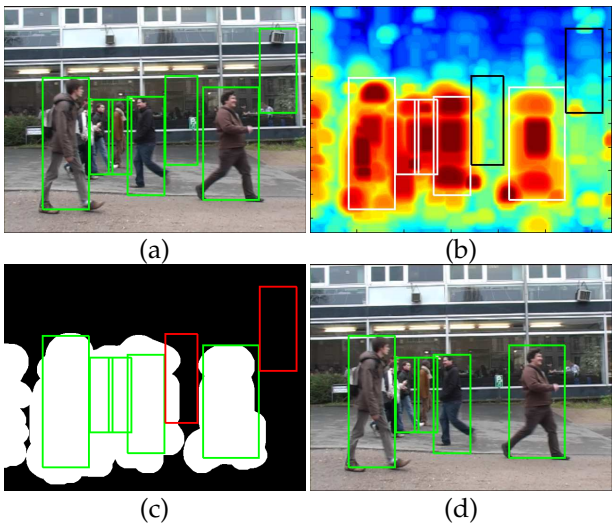


Figure 2. People detection system example: (a) people detections; (b) people detections over the DEBP segmentation confidence map; (c) people detections over the DEBP-P segmentation mask; and (d) final people detections.

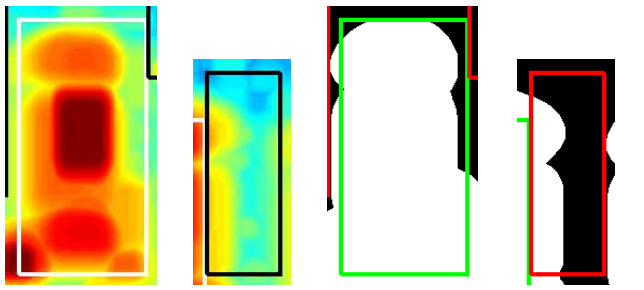
3.2 People detection post-processing based on people-background segmentation

In this section, we describe the people detection system that includes a post-processing or filtering stage using the people-background segmentation (see Figure 2). Firstly, people detections could be obtained using any people detector from the state of the art and the people-background segmentation is obtained with the approach proposed in [51] (see previous section 3.1). Then, both information sources are combined with the aim of eliminating or reducing the number of false detections while keeping, as much as possible, the number of positive detections. The combination of human detection and people-background segmentation is made with the detections and the people-background confidence map (Dependent Extended Body Parts, DEBP, confidence map [51]) or the binarized and post-processed segmentation mask (Dependent Extended Body Parts Post-processed, DEBP-P, segmentation mask [51]).

Our main objective is to demonstrate the utility of combining people-background segmentation instead of traditional foreground-background segmentation techniques. The evaluation with other combination techniques or strategies is out of the scope of this paper, but is part of the extensions of this work in the future.

Figure 2 shows one example where two false positives are eliminated using the people-background segmentation map (black blobs in Figure 2-b) and the people-background segmentation mask (red blobs in Figure 2-c).

In general, any people detection outcome always consists of a list of N detections in each frame t. Each detection n ($n = 1, ..., N$) is represented by its position (x, y) and dimensions (w, h), i.e., bounding box (or blob) $B_n(x, y, w, h)$ and a People-detection Confidence PC_n



$SC^C = 0.87$ (a) $SC^C = 0.50$ (b) $SC^M = 0.93$ (c) $SC^M = 0.38$ (d)

Figure 3. Examples of segmentation confidence $SC^{C/M}$ associated with a positive and a false detection: (a) and (b) using the DEBP confidence map SC^C; (c) and (d) using the DEBP-P segmentation mask SC^M.

$(0 \leq PC_n \leq 1)$. In order to process every detection, it has been defined a People Segmentation Confidence associated with every detection SC_n $(0 \leq SC_n \leq 1)$. This associated confidence is the averaged segmentation confidence over the corresponding blob. In the case of the DEBP confidence map $C(x, y)$, it is the averaged of the confidence values SC_n^C.

$$SC_n^C = \frac{1}{w \cdot h} \sum_{x,y \in B_n} C(x, y) \tag{1}$$

However, in the case of the DEBP-P segmentation mask $M(x, y)$ (a binarized and post-processed version of the DEBP confidence map), the segmentation confidence corresponds to the percentage of pixels classified as people versus the number of pixels classified as background SC_n^M.

$$SC_n^M = \frac{1}{w \cdot h} \sum_{x,y \in B_n} M(x, y) \tag{2}$$

Figure 3 shows SC^C and SC^M examples over a positive and a false detection.

Then the final list of detections consists of the initial N detections with a new associated confidence. This new confidence is the combination of the detection and segmentation confidences PSC_n $(0 \leq PSC_n \leq 1)$:

$$PSC_n = PC_n \cdot SC_n \tag{3}$$

Figure 4 shows one additional experimental example where it is shown the people detection performance with and without the proposed post-processing step. Figure 4(a) shows the Precision-Recall curve, whilst Figure 4(b) shows the relation between the true positive and false positive detection rate. In both cases, we can see how the use of the proposed post-processing step improves detection performance. In the first case, the Precision-Recall curve is significantly improved. In the second case, according to the selected threshold: (1) the number of true positives are maintained while reducing significantly false positives (straight line), or (2) the number of true positives is slightly reduced but reducing more the number of false positives (dotted line).

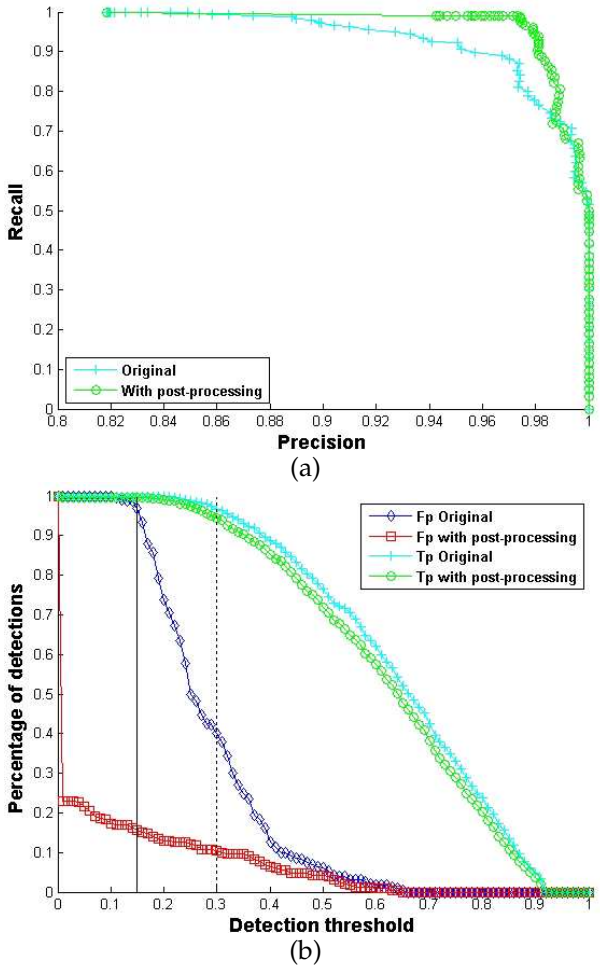


Figure 4. People detection performance example with and without the proposed post-processing step: (a) Precision-Recall curve; and (b) Percentage of false positive (Fp) and true positive (Tp) detections with and without the proposed post-processing. According to the selected threshold, the number of true positives are maintained while reducing false positives a 81% (straight line or 0.15 threshold), or the number of true positives is reduced a 3% but reducing the number of false positives a 29% (dotted line or 0.3 threshold).

4 Decision-level fusion of people detectors

In this section, we evaluate the decision-level fusion of independent appearance based people detectors. All detectors or experts are run in parallel, and the final decision is obtained as a combination of local expert responses using fusion methods widely studied in the literature. However, they are adapted to the particular case of people detection fusion at decision-level [52]: average, product, minimum, maximum, median and majority vote.

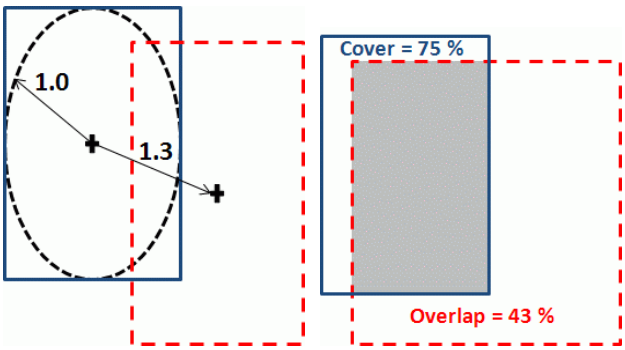


Figure 5. Evaluation criteria for comparing bounding boxes [53]: (left) relative distance; (right) cover and overlap.

Every people detector has its advantages and disadvantages, mainly because, each of them is based on different object extraction approaches and/or person models. The objective of this work is neither to evaluate individual detectors nor to analyze the correlation among them, but to evaluate that the fusion improves results. Frame by frame, every detector has different results; the main idea consists in keeping the correct true positive detections selected by a certain number of detectors and, at the same time, eliminating those false positive detections selected by only one or a smaller number of detectors.

In relation to the selected fusion techniques, our main objective is to validate the utility of combining multiple detectors in order to improve the final results. Therefore, the detectors and fusion techniques can be replaced by others without great difficulty. The use of different modules (detectors or fusion techniques) will vary the overall performance of the system, but the combination of detectors will be useful for improving the system. The evaluation with other detectors or more complex fusion techniques or strategies is out of the scope of this paper, but is part of the extensions of this work in the future.

In order to combine or fuse the different detectors, firstly, it is necessary to find matches or correspondences between every detector with the other detectors; the chosen matching criterion is the Multiple Hypotheses Simplification Criteria (MHSC) [30]. The MHSC allows us to compare hypotheses at different scales using the three evaluation criteria defined by Leibe et al. [53]: relative distance, cover and overlap. The relative distance (dr) measures the distance between the bounding box centers in relation to their size; cover and overlap measure how much of one bounding box hypothesis is covered by the other and vice versa (see Figure 5). A matching is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%.

Every people detector l has generally a different outcome N^l in each frame t. The number of detections and the detections themselves are not always matched between approaches (there is no unequivocal relationship between detectors' outcomes), so we are not able to apply directly the traditional fusion techniques [52]: average, product, minimum, maximum, median and majority vote. Instead, we evaluate the use of the five first mentioned fusion techniques but taking into account the minimum number of matches required in the fusion (variation of majority vote) in order to validate the fusion. Therefore, we perform the fusion and evaluate the five fusion techniques for each possible number of matches m ($m = 1, ..., L$). Assuming that one match corresponds actually to no matching, i.e., the detection is presented in only one detector. The final outcome is again a list of N^{out} detections, where each detection n ($n = 1, ..., N^{out}$) is represented by three components: (1) the matched averaged bounding box B_n^{out}, (2) the People-detection Confidence resulting to apply the corresponding fusion technique PC_n^{out} ($0 \leq PC_n^{out} \leq 1$) and (3) the corresponding number of matches m_n^{out}. Each final bounding box B_n^{out} is obtained as the average of the respective matched bounding boxes, whilst each final People-detection Confidence PC_n^{out} is obtained applying the corresponding fusion technique over the People-detection Confidence of the respective matched bounding boxes.

Figure 6 shows a visual fusion example with three detectors, whilst Algorithm 1 shows the corresponding fusion example pseudo-code. Following the example, we have three different people detectors outputs $L = 3$ ($l = 1, 2, 3$) in Figure 6 (a), (b) and (c); therefore, there are five fusion techniques for each possible number of matches ($m = 1, 2, 3$) among detectors (each detector has $N^l = 5, 4, 4$ detections respectively). The final l_{out} in Figure 6 (d) is the list of matched detections between the three detectors outputs. For example, the final detection number 7 is the result of matching the detections 5 and 4 from detectors 1 and 3 respectively. The final bounding box B_7^{out} is the average bounding box between both of them

$$B_7^{out}(x, y, w, h) = \frac{\left(B_5^1(x, y, w, h) + B_4^3(x, y, w, h)\right)}{2} \quad (4)$$

and the final People-detection Confidence PC_n^{out} is the corresponding fusion technique over them

$$PC_n^{out} = fusion\left(PC_5^1, PC_4^3\right) \quad (5)$$

5 Experimental Results

In order to evaluate our people detection approach, we compare in this section the original performance and the post-processed performance over seven people detection approaches. They have been chosen in order to cover the state of the art classification for people detection that we propose (see section 2.1): Edge [44], HOG [33], ISM [53], TUD [40], DTDP [41], ACF [42] and IMM [28].

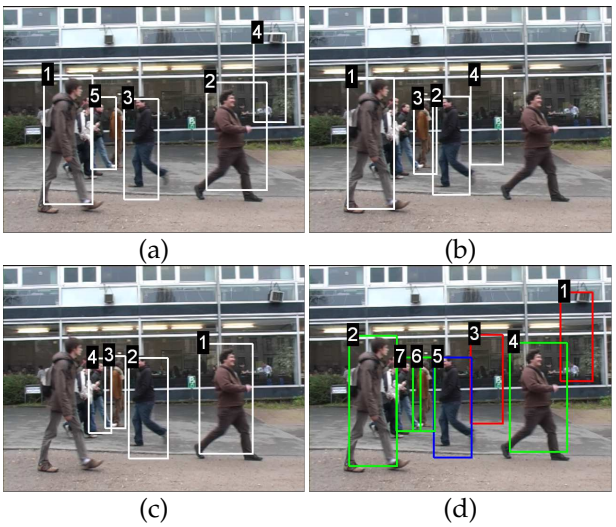


Figure 6. Visual people detection fusion example: (a) people detector outcome l_1; (b) people detector outcome l_2; (c) people detector outcome l_3; and (d) final people detection fusion outcome l_{out} (see Algorithm 1). Blue color corresponds to $m_5^{out} = 3$, green color corresponds to $m_{2,4,6,7}^{out} = 2$ and red color corresponds to $m_{1,3}^{out} = 1$.

Algorithm 1 People detection fusion example pseudo code.

- $L = 3$ $(l = 1, 2, 3)$.
- $l_{out} = fusion \begin{cases} l = 1, N^1 = 5 & \{B_1^1, PC_1^1\}, ..., \{B_5^1, PC_5^1\}. \\ l = 2, N^2 = 4 & \{B_1^2, PC_1^2\}, ..., \{B_4^2, PC_4^2\}. \\ l = 3, N^3 = 4 & \{B_1^3, PC_1^3\}, ..., \{B_4^3, PC_4^3\}. \end{cases}$
- $N^{out} = 7, l_{out} = \{B_1^{out} = B_4^1, PC_1^{out} = PC_4^1, m_1^{out} = 1\}, ...$ $\left\{B_7^{out} = \frac{(B_5^1 + B_4^3)}{2}, PC_7^{out} = fusion^* \left(PC_5^1, PC_4^3\right), m_7^{out} = 2\right\}$.

$^*average, product, minimum, maximum$ or $median$.

According to the chosen object detection approach, Edge combines segmentation and exhaustive search and the rest of them are based on exhaustive search. According to the chosen person model, the IMM includes the use of motion, appearance and their combination, the rest of them are based only on appearance: holistic (HOG, ISM, ACF) or part-based (Edge, TUD, DTDP).

Despite the fact that all algorithms performance depends on the hit rate, or confidence level of the decision, we only classify objects detected in previous stages as person or non-person. Consequently, the maximum or minimum Recall and Precision will be limited by previous stages. Edge is mainly limited by the segmentation step. Moreover, HOG, ISM, TUD, DTDP, ACF and IMM, are limited by the image scanning.

5.1 Experimental datasets

We use two different available people detection datasets from the state of the art. With the first one, the PDds dataset [54], we will make a deep analysis of the results over typical people detection scenarios and complexities. And with the second one, we will also include a brief evaluation over a typical surveillance setup for crowd analysis PETS 2009/2010 benchmark.

5.1.1 PDds dataset

The Person Detection dataset (PDds) [54] includes five different complexity categories depending on two people detection critical factors: classification and background complexity (low, medium and high). It mainly excels other datasets from the state of the art in the amount of sequences (90 videos) and variability of sequences. The experimental dataset includes both non-rigid and rigid people/objects differing in size, motion and textural appearance. These people/objects are involved in a number of interactions and in different contexts, like typical every-day situations or surveillance video scenarios. Regarding the backgrounds, it includes in-door and out-door scenarios with different background complexities (textural, lighting changes, multimodal, etc.). On the one hand, the background complexity is defined as the difficulty to detect the initial objects candidate to be person. This is due to the presence of edges, multiple textures, lighting changes, reflections, shadows and any kind of background variation. On the other hand, the people classification complexity is defined as the difficulty to verify every object candidate to be person. It is related to the number of objects, their velocity, partial occlusions, pose variations and interactions between different people and/or objects.

The dataset has been divided in two datasets (A and B). Dataset A includes 29 sequences including the five different complexity categories, whilst B includes 61 sequences of the highest complexity category (C5). Following [28], this dataset B has been divided in train and test sequences in order to evaluate the motion approach and appearance-motion combinations. The test dataset is composed of 36 sequences[1]. The training dataset is composed of the other 25 sequences. The dataset B includes the 61 sequences (train and test) but the named dataset B with motion only includes the 25 test sequences.

A summary of the complexity levels of the selected experimental sequences is shown in Table 1. In addition, Figure 7 shows some example frames from several sequences of the experimental datasets A and B, including the annotated ground truth.

5.1.2 PETS 2009 dataset

This second chosen dataset consists of sequences extracted from the PETS 2009/2010 benchmark[2]. It includes several sequences recorded outdoors from an elevated viewpoint, of the same location, corresponding to a typical surveillance setup for crowd analysis. The sequences are classified originally according to three scenarios (S1, S2 and S3) and three progressive difficulty levels (L1, L2

1. Test sequences (referring to PDds numbering): 2-5, 7-8, 12, 14, 18, 32, 34, 36-38 and 40-61.
2. http://www.cvg.rdg.ac.uk/PETS2009/

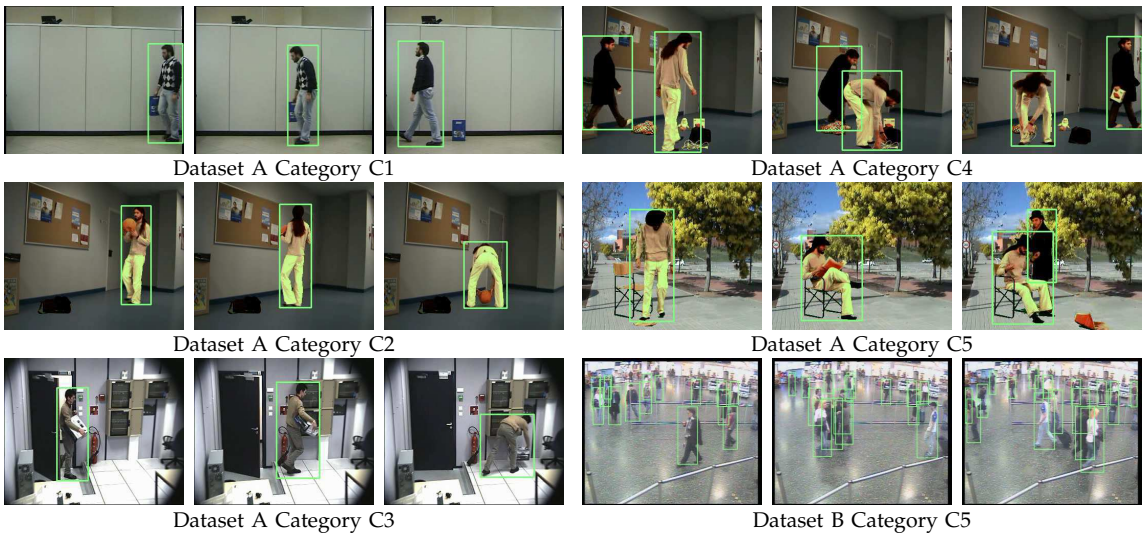


Figure 7. Experimental dataset examples. Every example shows three random frames from a sequence.

Category	#Sequences		#Frames	Complexity	
	Dataset A	Dataset B		Classification	Background
C1	6	0	1824	Low	Low
C2	6	0	2649	Medium	Low
C3	4	0	3143	Medium	Medium
C4	5	0	5301	High	Low
C5	8	61	15441	High	High

Table 1
Sequences categorization from evaluation dataset PDds [54].

Sequence	Up to #	Occupation	#Frames
PETS2009-S2L1	8	Low	795
PETS2009-S3L1	7	Low	107
PETS2009-S1L1-1	34	Medium	221
PETS2009-S1L1-2	26	Medium	241
PETS2009-S2L2	35	Medium	436
PETS2009-S1L2-1	42	High	201
PETS2009-S1L2-2	40	High	131
PETS2009-S2L3	42	High	240

Table 2
Sequences categorization from evaluation dataset PETS2009. Occupation in terms of number of pedestrians present simultaneously. Complexity classification.

and L3) for each scenario. These scenarios include high complexity in terms of crowds and occlusions (generally more than 10 pedestrians are present simultaneously).

In particular, [55] provides the ground truth of eight sequences of the PETS dataset, namely S1L1 (1 and 2), S1L2 (1 and 2), S2L1, S2L2, S2L3 and S3L1. The annotations only include the first view of each sequence. The main difference among the eight sequences is the number of pedestrians. We classify the whole set of sequences independently of the original scenario purpose (S1 for person count and density estimation, S2 for people tracking and S3 for flow analysis and event recognition). In our experiments, we classify the sequences according to the number of people present simultaneously and, therefore, the degree of occupation of the scene (low, medium or high). Table 2 includes a description of each sequence and complexity classification in terms of occupation. Figure 8 shows sample images of the used sequences.

5.1.3 Experimental setup

In order to evaluate different people detection approaches, we need to quantify the performance results. In the state of the art, performance can be evaluated at two levels: sequence sub-unit (frame, window, etc) or global sequence. Sub-unit performance is usually measured in terms of Detection Error Tradeoff (DET) [33], [56] or Receiver Operating Characteristics (ROC) [57], [58] curves. Global sequence performance is usually measured in terms of Precision-Recall (PR) curves [22], [37], [59]. The first level gives us information about the classification stage, while the second one provides overall system performance information. In order to evaluate a video surveillance system, it is more interesting to compare the overall performance. In both cases the detectors output is a confidence score for each person detection, where larger values indicate higher confidence. Both evaluation methods compute progressively the respective parameters such as the number of false positives, Recall rate or Precision rate, from the lowest possible score to the highest possible score. Each score threshold iteration provides a point on the curve.

The integrated Average Precision (AP) has been used to summarize the overall people detection performance, represented geometrically as the area under the Precision-Recall curve (AUC-PR). In order to take into account not only the yes/no detection decision but also the precise persons locations and extents, we validate the detection hypotheses with the annotated ground-truth applying also the same matching criteria MHSC used on the decision-level fusion process (see section 4). Only

(a) (b) (c)

Figure 8. Experimental sequences examples: (a) PETS2009-S2-L1, (b) PETS2009-S1-L1-2 and (c) PETS2009-S1-L2-1.

one hypothesis per object is accepted as correct, so any additional hypothesis on the same object is considered as a false positive. All the approaches use the default settings proposed by their respective authors: the Edge and IMM results have been obtained with the original code, the HOG results have been obtained using the available binaries[3], the ISM results have been obtained using the available code and binaries[4], the TUD results have been obtained using the available code[5], the DTDP results have been obtained using the available code [60] and the ACF results have been obtained using the available code and INRIA person model[6]. In the case of the people-background segmentation, it has been obtained using the original code (following [51], the chosen empirical binarization threshold is 0.8).

In order to apply any of the proposed post-processing approaches, the People Segmentation Confidence and People-detection Confidence are assumed to be normalized, SC_n $(0 \leq SC_n \leq 1)$ and PC_n $(0 \leq PC_n \leq 1)$ (see sections 3.2 and 4). The People Segmentation Confidence is already by definition normalized SC_n $(0 \leq SC_n \leq 1)$ [51]. However, every people detector use the default settings proposed by their respective authors and therefore, has different People-detection Confidence output space or range (see examples in Figure 9). We normalize every detector under evaluation PC_n $(0 \leq PC_n \leq 1)$. The normalization is performed according to the probability density function of the People-detection Confidence, which has been learnt or estimated using the respective detectors outputs over the INRIA dataset [33]. Figure 9 shows examples of different People-detection Confidence density functions (empirical and approximation).

5.2 Original detectors

5.2.1 Evaluation dataset PDds

Table 3 shows the original people detection results. The results over dataset A show clearly that all algorithms perform worse at higher complexity categories (from C1 to C5). All detectors provide similar or comparable results per each category. Due to the greater complexity of the dataset B, the results are worse than those obtained in the dataset A. In this case, the Edge and TUD approaches provide significantly worse results than the other appearance based approaches. Again, all the appearance based detectors provide similar results in dataset B with motion. Moreover, the results show how the combination of appearance and motion provide better results than the single appearance versions.

5.2.2 Computational cost

According to the computational cost, each detector's results has been obtained with the available code, implemented with different tools and programming languages, so a fair comparison is not possible. For this reason and according to the original implementations, we have decided to classify them in three categories: real time (Edge and ACF), near real time (HOG and DTDP) or no-real time (ISM, IMM and TUD). The tests have been performed on a Pentium IV with a CPU frequency of 2.4 GHz and 3GB RAM.

The Edge detector [44] combines segmentation and exhaustive search in order to achieve robustness and real time operation. It is a real time adaptation of the people detection approach [39]. The Edge approach [44] is implemented in C++ (OpenCV) and the computational cost is around 0.02 seconds per frame with 352x288 images.

The ACF detector proposes a very fast exhaustive search and a holistic person model using aggregate channel features. The ACF approach [42] is implemented in Matlab and the computational cost is around 0.02 seconds per frame with 352x288 images.

The HOG detector [33] is based on exhaustive search and a holistic person model using the Histogram of Oriented Gradients. It consists in scanning the full image looking for similarities with the chosen person model, evaluating different detection windows with a classifier at multiple scales and locations. The HOG approach [33] is implemented in C++ and the computational cost is around 1 second per frame with 352x288 images (there is a faster implementation in OpenCV that runs around 0.1 seconds per frame).

The DTDP detector [41] is based on exhaustive search and a part-based person model. The DTDP approach [41] is implemented with Matlab and the computational cost is around 2 seconds per frame with 352x288 images

3. http://pascal.inrialpes.fr/soft/olt/
4. http://www.vision.ee.ethz.ch/~bleibe/index.html
5. http://www.d2.mpi-inf.mpg.de/andriluka_cvpr09
6. http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html

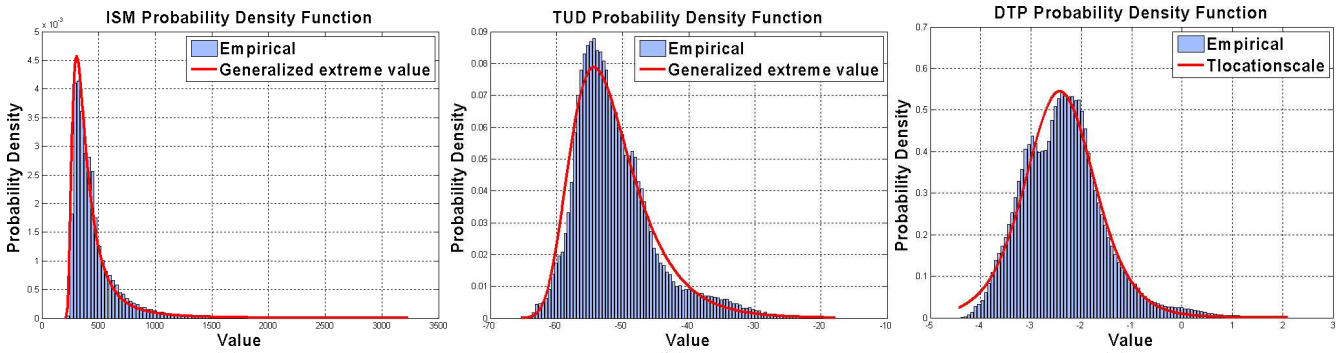


Figure 9. People-detection Confidence output distributions empirical and approximation examples: (a) ISM detector, (b) TUD detector and (c) DTDP detector.

Dataset	Category	Edge	HOG	ISM	TUD	DTDP	ACF	IMM
A	C1	0.98	0.92	0.95	0.93	0.96	0.94	-
A	C2	0.93	0.86	0.91	0.88	0.92	0.88	-
A	C3	0.85	0.74	0.80	0.75	0.81	0.80	-
A	C4	0.89	0.82	0.84	0.84	0.86	0.84	-
A	C5	0.70	0.71	0.71	0.67	0.74	0.78	-
B	C5	0.59	0.66	0.69	0.56	0.68	0.72	-
B	C5 motion	0.58	0.66	0.64	0.56	0.67	0.70	0.60
Dataset	Category	Edge+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	ACF+IMM	-
B	C5 motion	0.62	0.68	0.67	0.62	0.70	0.72	-

Table 3
Original results over dataset PDds in terms of AUC-PR.

(there is a faster implementation in OpenCV that runs around 1 second per frame).

The ISM people detector [53] is based on exhaustive search and a holistic person model. It consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations by local features matching. The chosen person model is based on appearance information using the SIFT features. On the second hand, the IMM detector [30] is a variation of the ISM detector where the chosen person model is based in the characteristic movements of people using the MoSIFT features. Both approaches have been implemented with C++ and have similar computational cost between 4-7 seconds per frame with 352x288 images.

The TUD people detector [40] is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original ISM detector [53] using pictorial structures. The TUD approach [40] is implemented with Matlab subroutines and C++, the computational cost is several orders of magnitude greater than the other approaches.

5.3 Results using people-background segmentation confidence

5.3.1 Evaluation dataset A

Tables 4 and 5 show the people detection results using the DEBP confidence map and the DEBP-P segmentation mask respectively. The use of the people-background segmentation allows us to reduce the number of false detections and, therefore, in almost all the cases we improve the global detection results. The use of DEBP confidence map provides good segmentation results but the DEBP-P segmentation mask provides better results thanks to the use of a segmentation post-processing. For this reason, the detection improvements obtained with the DEBP-P segmentation mask (average improvement of 3.8%) are significantly better than the ones obtained with the DEBP confidence map (average improvement of 2.8%) with the inconveniences of binarization (see section 3 for more details).

According to the experimental dataset, the results show that in both cases (DEBP and DEBP-P) the highest improvements are obtained in categories C3 (average improvement of 6.8 and 9.3% respectively) and C5 (average improvement of 3.1 and 4.3% respectively). It is due mainly to the background complexity: these two categories (C3 and C5) present medium or high background complexity, being the background complexity one of the main factors that produce false detections. For the same reason, the lowest improvements (except the simplest category C1) are obtained in categories C2 (average improvement of 1.7 and 2.0%) and C4 (average improvement of 1.4 and 2.2%). This is because the complexity of these categories lies on the classification.

According to thedetection approach, in general the results show that in both cases (DEBP and DEBP-P) there are detection improvements. In particular the HOG approach presents negative results in category C1 because it generates bigger blobs than the other detectors. Every blob always contains the person and a small part of background around, but in this case the background

	Edge	%Δ	HOG	%Δ	ISM	%Δ	TUD	%Δ	DTDP	%Δ	ACF	%Δ	Total	%ΔTotal
A.C1	0.99	+1.0	0.91	-1.1	0.98	+3.2	0.96	+3.2	0.96	+0.0	0.94	+0.0	0.96	+1.1
A.C2	0.95	+2.2	0.86	+0.0	0.93	+2.2	0.91	+3.4	0.92	+0.0	0.90	+2.3	0.91	+1.7
A.C3	0.90	+5.9	0.79	+6.8	0.87	+8.8	0.83	+10.7	0.85	+4.9	0.83	+3.8	0.85	+6.8
A.C4	0.89	+0.0	0.83	+1.2	0.87	+3.6	0.85	+1.2	0.87	+1.2	0.85	+1.2	0.86	+1.4
A.C5	0.73	+4.3	0.73	+2.8	0.74	+4.2	0.70	+4.5	0.74	+0.0	0.80	+2.6	0.74	+3.1
Total	0.89	-	0.82	-	0.88	-	0.85	-	0.87	-	0.86		0.86	-
%ΔTotal	-	+2.7	-	+1.9	-	+4.4	-	+4.6	-	+1.2		+2.0	-	+2.8

Table 4
People detection performance using the DEBP confidence map over dataset A. Percentage increase (%Δ) over the original performance (see section 5.2.1).

	Edge	%Δ	HOG	%Δ	ISM	%Δ	TUD	%Δ	DTDP	%Δ	ACF	%Δ	Total	%ΔTotal
A.C1	0.99	+1.0	0.91	-1.1	0.98	+3.2	0.97	+4.3	0.96	+0.0	0.94	+0.0	0.96	+1.2
A.C2	0.95	+2.2	0.86	+0.0	0.94	+3.3	0.91	+3.4	0.92	+0.0	0.91	+3.4	0.92	+2.0
A.C3	0.92	+8.2	0.78	+5.4	0.89	+11.3	0.87	+16.0	0.87	+7.4	0.86	+7.5	0.87	+9.3
A.C4	0.90	+1.1	0.84	+2.4	0.88	+4.8	0.86	+2.4	0.87	+1.2	0.85	+1.2	0.87	+2.2
A.C5	0.74	+5.7	0.73	+2.8	0.75	+5.6	0.72	+7.5	0.75	+1.4	0.80	+2.6	0.75	+4.3
Total	0.90	-	0.82	-	0.89	-	0.87	-	0.87	-	0.88		0.87	-
%ΔTotal	-	+3.6	-	+1.9	-	+5.6	-	+6.7	-	+2.0		+2.9		+3.8

Table 5
People detection performance using the DEBP-P segmentation mask over dataset A. Percentage increase (%Δ) over the original performance (see section 5.2.1).

around is bigger, so the associated confidence computed over the corresponding blob (see equation 1) is affected negatively.

5.3.2 Evaluation dataset B

Table 6 shows the people detection results using the DEBP confidence map and the DEBP-P segmentation mask. As in the evaluation of dataset A, in almost all the cases we improve the global detection results: we can see how the improvements obtained with the DEBP-P segmentation mask (average improvement of 2.5%) are significantly better than the ones obtained with the DEBP confidence map (average improvement of 1.4%). In general the improvements obtained with dataset B are smaller than the ones obtained with dataset A. The results are comparable with the results obtained in categories C2 and C4 of dataset A. The main reason for this is that the complexity of dataset B lies not only on background complexity, but also on the classification complexity.

	Edge	HOG	ISM	TUD	DTDP	ACF	Total
B.C5*	0.60	0.66	0.69	0.58	0.69	0.73	0.66
B.C5 (%Δ)*	+1.7	+0.0	+0.0	+3.6	+1.5	+1.4	+1.4
B.C5**	0.61	0.66	0.70	0.60	0.69	0.73	0.67
B.C5 (%Δ)**	+3.4	+0.0	+1.4	+7.1	+1.5	+1.4	+2.5

Table 6
People detection performance using the DEBP confidence map* or DEBP-P segmentation mask** over dataset B. Percentage increase (%Δ) over the original performance (see section 5.2.1).

5.3.3 Evaluation dataset B with motion

Table 7 shows the people detection results using the DEBP confidence map and the DEBP-P segmentation mask. As in the evaluation of dataset A and dataset B without motion, in almost all the cases we improve the global detection results: we can see how the improvements obtained with the single appearance versions with the DEBP-P segmentation mask and with the DEBP confidence map (average improvement of 3.0 and 1.9% respectively), or motion versions (average improvement of 2.5 and 1.5% respectively) are quite similar to the ones obtained with the dataset B without motion (see previous section 5.3.2). However, the results show how the use of motion and the proposed post-processing obtains the best final results, with the DEBP-P segmentation mask and with the DEBP confidence map (AUC-PR Total average of 68∼69%).

5.3.4 Computational cost

The proposed post-processing approach includes two main tasks, the computation of the people-background segmentation and the combination of detection and segmentation confidences. According to the original computational cost (see section 5.2.2), the additional computational cost of the second task is almost insignificant (averaged of the dense confidence or percentage of foreground pixels -see section 3.2-). However, the first step introduces a considerable additional computational cost. The people-background segmentation is based on the DTDP detector [41] and has a comparable computational

	Edge	HOG	ISM	TUD	DTDP	ACF	Total	IMM
B.C5 motion*	0.59	0.65	0.66	0.58	0.68	0.72	0.65	0.62
B.C5 motion (%Δ)*	+1.7	-1.5	+3.1	+3.6	+1.5	+2.9	+1.9	+3.3
B.C5 motion**	0.60	0.66	0.66	0.60	0.68	0.72	0.66	0.62
B.C5 motion (%Δ)**	+3.4	+0.0	+3.1	+7.1	+1.5	+2.9	+3.0	+3.3

	Edge+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	ACF+IMM	Total
B.C5 motion*	0.63	0.68	0.69	0.64	0.71	0.72	0.68
B.C5 motion (%Δ)*	+1.6	+0.0	+3.0	+3.2	+1.4	+0.0	+1.5
B.C5 motion**	0.64	0.68	0.69	0.65	0.71	0.74	0.69
B.C5 motion (%Δ)**	+3.2	+0.0	+3.0	+4.8	+1.4	+2.8	+2.5

Table 7
People detection performance using the DEBP confidence map* or DEBP-P segmentation mask** over dataset B with motion. Percentage increase (%Δ) over the original performance (see section 5.2.1).

cost, i.e., the computational cost is around 2 seconds per frame with 352x288 images in Matlab (using the faster implementation of DTDP in OpenCV, it also runs around 1 second per frame).

The proposed approach has been implemented as a post-processing stage. However, it could also be applied as a preprocessing step. This would produce similar detection results and at the same time, a computational cost reduction of the subsequent people detector approach.

5.4 Results using decision-level fusion

5.4.1 Evaluation dataset A

According to the original results that have been already discussed in section 5.2.1, we have evaluated every possible minimum number of matches m ($m = 1, ..., L$) required in the fusion. Figure 10 shows the average results fusing the six detectors over the five experimental dataset complexity categories (C1-C5). Firstly, the effect of the minimum number of matches required in the fusion is clear. With low concurrence requirements $m = 1$ or high concurrence requirements $m = 6$ the final results are clearly worse. In the first case, it is because every detection is considered in the fusion, so every independent and isolated detection error is included in the final results. In the second case, there are missing detections due to the excessive detection concurrence requirements. The best results are obtained around $m = 3$. In relation to the fusion technique, the product method provides clearly the worst fused results: the product method is optimal only if all the detectors are totally independent. Although all the detectors are independently build, there is some kind of dependence since all of them are based on people appearance. The rest of fusion methods provide similar results, being slightly better the average.

In order to visualize the detection results per each experimental dataset complexity category, we have selected the best number of minimum matches required for each configuration ($m = 3$) and the best performance fusion method (average). All experimental results are available as additional material (http://www-vpu.eps.uam.es/publications/PeopleDetectionPostProcessing/).

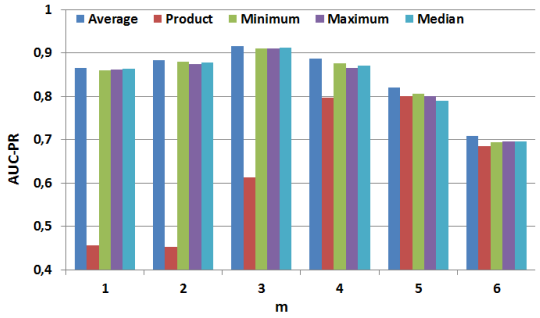


Figure 10. Total average fusion performance over dataset A, for each fusion technique [52] (average, product, minimum, maximum and median) and minimum number (m) of matches required in the fusion.

$m = 3$	Edge	HOG	ISM	TUD	DTDP	ACF	Total
A.C1	1.0	+2.0	+8.7	+5.3	+7.5	+4.2	+6.4	+5.6
A.C2	0.96	+3.2	+11.6	+5.5	+9.1	+4.3	+9.1	+7.1
A.C3	0.87	+2.4	+17.6	+8.8	+16.0	+7.4	+8.8	+10.1
A.C4	0.92	+3.4	+12.2	+9.5	+9.5	+7.0	+9.5	+8.5
A.C5	0.82	+17.1	+15.5	+15.5	+22.4	+10.8	+5.1	+14.4
Total	0.91	+5.6	+13.1	+8.9	+12.9	+6.7	+7.8	+9.2

Table 8
People detection performance fusing the six detectors using average fusion over dataset A. Percentage increase (%Δ) over the original individual performance (see section 5.2.1).

Table 8 shows the people detection performance fusing the six detectors per each experimental dataset complexity. The results clearly show that the proposed people detection fusion improves considerably the original people detection results. The average improvements obtained for each experimental dataset complexity are between 5.6 and 14.4%. Finally, the average improvements obtained are clearly higher in more complex scenarios (C3-C5) than in the simplest ones (C1-C2). It is logical because the range of possible improvement is greater and the advantage of combining detectors is more evident (allowing to reduce errors and increase the overall detection rate).

	$m = 3^{*}/2^{**}$	Edge	HOG	ISM	TUD	DTDP	ACF	Total
B.C5*	0.74	+25.4	+12.1	+7.2	+32.1	+8.8	+2.8	+14.8
B.C5**	0.76	-	+15.2	+10.1	-	+11.8	+5.6	+10.7

Table 9
People detection performance fusing the six* or four detectors (HOG, ISM, DTDP and ACF)** using average fusion over dataset B. Percentage increase (%Δ) over the original individual performance (see section 5.2.1).

According to the individual people detector results, the improvements on those detectors with worse original performance are logically greater than the improvements on those detectors with better original performance. On the one hand, the HOG approach provides the worst original performance results (see section 5.2.1) and the greatest improvement (average improvement 13.1%). On the other hand, the Edge detector provides the best original performance results (see section 5.2.1) and the lowest improvement (average improvement 5.6%).

5.4.2 Evaluation dataset B

As already commented, the detectors and fusion techniques can be replaced by others without great difficulty. The use of different modules will vary the overall performance of the system, but the combination of detectors and additional post-processing stages will always be useful for improving the system (except in the ideal case of perfect detection). In order to validate this statement, we have defined two different people detection fusion configurations with the available sources: the first one including the six detectors in the fusion and the second one including only the four best detectors over dataset B (HOG, ISM, DTDP and ACF).

As in the evaluation of dataset A, in order to visualize the detection results, we have selected the best number of minimum matches required for each configuration ($m = 3$ and $m = 2$ respectively) and we have selected only the best performance fusion method (average).

Table 9 shows the people detection performance of both configurations. In almost all the cases we improve the global detection results: we can see how the final results obtained fusing only the best four detectors (76%) are better than the ones obtained fusing the six detectors (74%).

In relation to the individual people detector results, on the one hand, the TUD approach provides the worst original performance results (see section 5.2.1) and the greatest improvement (average improvement between 32.1%). On the other hand, the ACF detector provides the best original performance results (see section 5.2.1) and the lowest improvement (average improvement between 2.8$\sim$5.6%).

5.4.3 Evaluation dataset B with motion

According to the original results and following the same evaluation scheme as in the evaluation of dataset B (see previous section 5.4.2), we have defined the same two people detection fusion configurations and the same evaluation parameters (average fusion method and minimum matches required for each configuration).

Table 10 shows, firstly, the people detection performance fusing six or four appearance based detectors respectively and the motion based detector performance. In this case, the results are quite similar to the ones obtained with the dataset B without motion (see previous section 5.4.2). And secondly, table 10 also shows the appearance and motion based detectors combinations. Again, in almost all the cases we improve the global detection results, we can see how the results obtained fusing only the best detectors (76%) are better than the ones obtained fusing the six detectors (74%).

According to the individual people detector results, on the one hand, the TUD+IMM and Edge+IMM approaches provide the worst original performance results (see section 5.2.1) and the greatest improvement (average improvement of 19.4%). On the other hand, the ACF+IMM detector provides the best original performance results (see section 5.2.1) and the lowest improvement (average improvement between 2.8$\sim$5.6%).

Finally, the results show how the use of motion and the proposed fusion obtains the best final results (AUC-PR final between 74 or 76%).

5.4.4 Computational cost

In this case, the proposed post-processing approach only includes two additional tasks, the matching and fusion between detectors.

According to the original computational cost (see section 5.2.2), the additional computational cost is almost insignificant (see section 4). For this reason the computational cost will be established by the chosen people detection approaches. Assuming that we run all people detectors in parallel, the final fusion approach computational cost will be established by the detection approach with the higher computational cost (in this case the TUD or TUD+IMM detector).

5.5 Results using both post-processing approaches

In this section, we evaluate the successive application of both post-processing approaches. In first place, we present the results over the PDds dataset (A, B and B with motion). And, in second place, we also present the results over an additional dataset designed for specific crowd analysis: PETS 2009 benchmark.

5.5.1 Evaluation dataset PDds

In order to present the final results combining both post-processing approaches, we use the post-processing configurations with the best independent results and the best successive application order, i.e., we first apply the segmentation post-processing and then the decision-level fusion. We make use of the DEBP-P segmentation mask. In the evaluation of dataset A, we combine the six

	$m = 3^{*}/2^{**}$	Edge	HOG	ISM	TUD	DTDP	ACF	Total	IMM
B.C5 motion*	0.72	+24.1	+9.1	+12.5	+28.6	+7.5	+2.9	+14.1	+20.0
B.C5 motion**	0.74	-	+12.1	+15.6	-	+10.4	+5.7	+11.0	-

	$m = 3^{*}/2^{**}$	Edge+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM	ACF+IMM	Total
B.C5 motion*	0.74	+19.4	+8.8	+10.4	+19.4	+5.7	+2.8	+11.1
B.C5 motion**	0.76	-	+11.8	+13.4	-	+8.6	+5.6	+9.8

Table 10
People detection performance fusing the six* or four (HOG, ISM, DTDP and ACF)** appearance and/or motion based detectors combinations using average fusion over dataset B with motion. Percentage increase (%Δ) over the original individual performance (see section 5.2.1).

detectors and in the case of dataset B (with or without motion), we fuse the four best detectors (HOG, ISM, DTDP and ACF). Additional experimental results are available as additional material (http://www-vpu.eps.uam.es/publications/PeopleDetectionPostProcessing/).

Tables 11 and 12 show the results of the successive application of both post-processing approaches in dataset A and B without/with motion respectively. The original people detection results have been already discussed in section 5.2.1 and both independent post-processed results in sections 5.3 and 5.4. In relation with the original results and as in the previous independent post-processed results, in all cases there is a significant improvement. However, the global results and the improvements are higher than the ones obtained with the individual post-processing approaches. In the case of dataset A, the average improvement obtained (11.2%) is higher than the ones obtained using only the fusion post-processing approach (9.2%, see Table 8). In the case of dataset B and B with motion, the average improvements obtained with the appearance based detectors (12.1 or 17.0% respectively) are higher than the ones obtained using only the fusion post-processing approach (10.7 or 11.1% respectively, see Tables 9 and 10). And, finally, the average improvement obtained with the appearance-motion based detectors (15.6%) is higher than the ones obtained using only the fusion post-processing approach (9.8%, see Table 10). The results show that the additional improvements obtained in dataset B (with or without motion) are higher than the ones obtained in dataset A (C1-C4). It is logical because those scenarios (B and A.C5) are more complex and therefore the range of possible improvement is greater.

Finally, in order to summarize all the different detection results, Table 13 shows only the average detection results of the four detection configurations (original approach, using people-background segmentation, using decision-level fusion and using both post-processing approaches) on the experimental dataset PDds.

5.5.2 PETS2009 dataset

In this section, we also evaluate the successive application of both post-processing approaches over an additional challenging dataset: PETS 2009 benchmark. As in the previous section, we use the post-processing

	$m = 3$	Edge	HOG	ISM	TUD	DTDP	ACF	Total
A.C1	1.0	+2.0	+8.7	+5.3	+7.5	+4.2	+6.4	+5.7
A.C2	0.97	+4.3	+12.8	+6.6	+10.2	+5.4	+10.2	+8.3
A.C3	0.89	+4.7	+20.3	+11.3	+18.7	+9.9	+11.3	+12.7
A.C4	0.94	+5.6	+14.6	+11.9	+11.9	+9.3	+11.9	+10.9
A.C5	0.85	+21.4	+19.7	+19.7	+26.9	+14.9	+9.0	+18.6
Total	0.93	+7.6	+15.2	+10.9	+15.0	+8.7	+9.7	+11.2

Table 11
People detection performance using the DEBP-P segmentation mask, fusing the six detectors and using average fusion over dataset A. Percentage increase (%Δ) over the original individual performance (see section 5.2.1).

configurations with the best independent results (HOG, ISM, DTDP and ACF) and the best successive application order, i.e., we first apply the segmentation post-processing and then the decision-level fusion. Additional experimental results are available as additional material (http://www-vpu.eps.uam.es/publications/PeopleDetectionPostProcessing/).

The results show clearly how the performance decreasesfrom the simplest sequences to the medium and high complexity sequences (see Table 2). The HOG detector provides the worse results in all the sequences. The main reason for this behavior is that the HOG detector is based on a holistic person model and presents difficulties dealing with occlusions. On the other hand, the DTDP detector is a part-based version of the HOG, and provides the best results in all the sequences independently of the complexity. In general, both ACF and ISM approaches have acceptable results over low complexity categories. However, they presents more difficulties dealing with partial occlusions than the DTDP over more complex sequences.

Table 15 shows the best combination obtained using both proposed post-processing approaches, i.e., the DEBP-P and only the two best detectors (DTDP and ACF). Again the results show clearly how the use of both post-processing approaches improve the final detection results in different complexity sequences (average improvement of 9.6%), being this improvement more significant in more complex scenarios (6.9∼18.4%) than the simplest ones (3.7∼7.0%).

	$m=2$	HOG	ISM	DTDP	ACF	Total
B.C5	0.77	+16.7	+11.6	+13.2	+6.9	+12.1
B.C5 motion	0.78	+18.2	+21.9	+16.4	+11.4	+17.0
	$m=2$	HOG+IMM	ISM+IMM	DTDP+IMM	ACF+IMM	Total
B.C5 motion	0.80	+17.6	+19.4	+14.3	+11.1	+15.6

Table 12
People detection performance using the DEBP-P segmentation mask, fusing the four detectors (HOG, ISM, DTDP and ACF) and using average fusion over dataset B without/with motion. Percentage increase (%Δ) over the original individual performance (see section 5.2.1).

			Dataset B motion	
	Dataset A	Dataset B	Appearance	Appearance+IMM
Original	0.84	0.65	0.64	0.67
People-background segmentation	0.87	0.67	0.66	0.69
Decision-level fusion	0.91	0.76	0.74	0.76
Both post-processing approaches	0.93	0.77	0.78	0.80

Table 13
Total people detection performance average results: original detection results (see section 5.2.1), using people-background segmentation confidence (see section 5.3), using decision-level fusion (see section 5.4) and using both post-processing approaches on the experimental dataset (dataset A, B, B with motion and B with motion combining appearance and the motion information of the IMM detector).

	HOG	ISM	DTDP	ACF
PETS2009-S2L1	0.60	0.78	0.93	0.85
PETS2009-S3L1	0.68	0.82	0.93	0.94
PETS2009-S1L1-1	0.40	0.45	0.63	0.63
PETS2009-S1L1-2	0.41	0.49	0.73	0.68
PETS2009-S2L2	0.50	0.55	0.66	0.58
PETS2009-S1L2-1	0.28	0.30	0.48	0.44
PETS2009-S1L2-2	0.34	0.36	0.50	0.51
PETS2009-S2L3	0.31	0.34	0.55	0.47

Table 14
Original results over dataset PETS 2009 in terms of AUC-PR.

	$m=1$	DTDP	ACF	Total
PETS2009-S2L1	0.95	+2.2	+11.8	+7.0
PETS2009-S3L1	0.97	+4.3	+3.2	+3.7
PETS2009-S1L1-1	0.68	+7.9	+7.9	+7.9
PETS2009-S1L1-2	0.75	+2.7	+10.3	+6.5
PETS2009-S2L2	0.71	+7.6	+22.4	+15.0
PETS2009-S1L2-1	0.51	+6.3	+15.9	+11.1
PETS2009-S1L2-2	0.54	+8.0	+5.9	+6.9
PETS2009-S2L3	0.60	+9.1	+27.7	+18.4
Total	0.71	+6.0	+13.1	+9.6

Table 15
People detection performance using the DEBP-P segmentation mask, fusing the best two detectors (DTDP and ACF) and using average fusion over dataset PETS2009. Percentage increase (%Δ) over the original individual performance.

6 Conclusions

Firstly, we have presented a new subtask for people detection filtering.This subtask enhances people detection results making use of the information about where there are not people obtained with people-background segmentation. The experimental results show the performance of our proposal over the proposed evaluation dataset PDds. There is a global detection improvement in almost every category and original people detection approach, being this improvement more clear in those scenarios with medium or high background complexity. It is logical because those scenarios are more likely to generate false detections. The results also show how the use of motion in addition to our approach obtains the best final results.

Secondly, we have evaluated the combination or fusion of six independent appearance based people detectors at decision-level. We have also evaluated their combination with a motion based people detector. In order to fuse the different detectors, we have evaluated a multiple matching criteria and the application of traditional fusion techniques: average, product, minimum, maximum and median. The experimental results show the performance of our proposed approach with the mentioned fusion techniques. The product method shows clearly worse results, whilst the average method provides slightly better results than the other three methods. There is a global detection improvement in every category and original people detection approach. This improvement is more clear in those scenarios with higher complexity, since those scenarios are more likely to generate false detections and missing detections. Again, the results show how the use of motion in addition to the

proposed fusion obtains the best final results.

And, finally, we have also evaluated the successive application of both post-processing approaches over both chosen evaluation datasets from the state of the art: PDds and PETS2009. The results show the additional improvements obtained in all the cases thanks to the combination of both post-processing stages.

As future work, we will try to improve the segmentation confidence using its evolution over time or its combination with another more traditional segmentation strategy: color based, motion based, etc. After showing that this processing allows improving detection results, we will study the use of the people-background segmentation as a preprocessing state in order to maintain/reduce computation cost. In addition, other combinations of detection and segmentation confidences may be explored. In relation to the fusion post-processing approach, we will explore other more complex fusion possibilities, not only fixed fusion rules but also trainable fusion rules or adaptive weights based on online quality estimation, and not only parallel fusion schemes but also cascade, hierarchical or hybrid. It is clear that "independently built" detectors exhibit positive correlation, and this is attributed to the fact that difficult parts of the decision space are difficult for all detectors. So we also propose to explore other independent detectors (e.g., based on motion) or other fusion techniques robust to decision correlations. Finally, we also propose a further evaluation including other different evaluation setups (point of view, occupation, etc), i.e., other complexity categories over the evaluation datasets.

7 Acknowledgments

This work has been partially supported by the Spanish Government (TEC2011-25995 EventVideo).

References

[1] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22(8), pp. 781–796, 2000.

[2] J. Giebel, D. M. Gavrila, and C. Schnorr, "A bayesian framework for multi-cue 3d object tracking," in *Proc. of ECCV*, 2004, pp. 241–252.

[3] F. Xu and K. Fujimura, "Human detection using depth and gray images," in *Proc. of AVSS*, 2003, pp. 115–121.

[4] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(9), pp. 1208–1221, 2004.

[5] J. Zhou and J. Hoang, "Real time robust human detection and tracking system," in *Proc. of CVPR*, 2005, pp. 149–156.

[6] M. Hussein, W. Abd-Almageed, Y. Ran, and L. Davis, "Real-time human detection, tracking, and verification in uncontrolled camera motion environments," in *Proc. of ICVS*, 2006, pp. 41–47.

[7] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *International Journal of Computer Vision*, vol. 73(1), pp. 41–59, 2007.

[8] N. Koenig, "Toward real-time human detection and tracking in diverse environments," in *Proc. of ICDL*, 2007, pp. 94–98.

[9] V. Fernández-Carbajales, M. A. García, and J. M. Martínez, "Robust people detection by fusion of evidence from multiple methods," in *Proc. of WIAMIS*, 2008, pp. 55–58.

[10] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Computer Vision and Image Understanding*, vol. 110(1), pp. 43–59, 2008.

[11] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22(8), pp. 809–830, 2000.

[12] N. Sprague and J. Luo, "Clothed people detection in still images," in *Proc. of ICPR*, 2002, pp. 585–589.

[13] S. Harasse, L. Bonnaud, and M. Desvignes, "Human model for people detection in dynamic scenes," in *Proc. of CVPR*, 2006, pp. 335–354.

[14] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. of ICCV*, 2003, pp. 734–741.

[15] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proc. of ECCV*, 2004, pp. 28–39.

[16] H. Sidenbladh, "Detecting human motion with support vector machines," in *Proc. of ICPR*, 2004, pp. 188–191.

[17] N. Dalal and B. Triggs, "Human detection using oriented histograms of flow and appearance," in *Proc. of ECCV*, 2006, pp. 428–441.

[18] S. Avidan, "Ensemble tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261–271, 2007.

[19] X. Cui, Y. Liu, S. Shan, X. Chen, and W. Gao, "3d haar-like features for pedestrian detection," in *Proc. of ICME*, 2007, pp. 1263–1266.

[20] B. Leibe, K. Schindler, and L. V. Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *Proc. of ICCV*, 2007, pp. 1–8.

[21] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[22] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. of CVPR*, 2008, pp. 1–8.

[23] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1728–1740, 2008.

[24] X. Ren, "Finding people in archive films through tracking," in *Proc. of CVPR*, 2008, pp. 1–8.

[25] A. Ess, B. Leibe, K. Schindler, and L. V. Gool, "Robust multiperson tracking from a mobile platform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(10), pp. 1831–1846, 2009.

[26] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2010.

[27] S. Stalder, H. Grabner, and L. V. Gool, "Cascaded confidence filtering for improved tracking-by-detection," in *Proc. of ECCV*, 2010, pp. 369–382.

[28] A. Garcia-Martin, A. Hauptmann, and J. M. Martinez, "People detection based on appearance and motion models," in *Proc. of AVSS*, 2011, pp. 256–260.

[29] J. Yu, D. Farin, and B. Schiele, "Multi-target tracking in crowded scenes," in *Proc. of DAGM*, 2011, pp. 406–415.

[30] A. Garcia-Martin and J. M. Martinez, "On collaborative people detection and tracking in complex scenarios," *Image and Vision Computing*, vol. 30(4), pp. 345–354, 2012.

[31] B. Leibe and B. Schiele, "Scale invariant object categorization using a scale-adaptive mean-shift search," in *Proc. of DAGM*, 2004, pp. 145–153.

[32] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57(2), pp. 137–154, 2004.

[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of CVPR*, 2005, pp. 886–893.

[34] E. Seemann and B. Schiele, "Cross-articulation learning for robust detection of pedestrians," in *Proc. of DAGM*, 2006, pp. 242–252.

[35] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. of CVPR*, 2006, pp. 1491–1498.

[36] W. Zhang, G. Zelinsky, and D. Samaras, "Real-time accurate object detection using multiple resolutions," in *Proc. of ICCV*, 2007, pp. 1–8.

[37] 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.

[38] C. Wojek, G. Dorkó, A. Schulz, and B. Schiele, "Sliding-windows for rapid object class localization: A parallel technique," in *Proc. of DAGM*, 2008, pp. 71–81.

[39] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proc. of ICCV*, 2005, pp. 90–97.

[40] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. of CVPR*, 2009, pp. 1014–1021.

[41] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(9), pp. 1627–1645, 2010.

[42] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Proc. of ECCV*, no. 645-659, 2012.

[43] I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. R. de Toro, J. Nuevo, M. Ocaña, and M. A. G. Garrido, "Combination of feature extraction methods for svm pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8(2), pp. 292–307, 2007.

[44] A. Garcia-Martin and J. M. Martinez, "Robust real time moving people detection in surveillance scenarios," in *Proc. of AVSS*, 2010, pp. 241–247.

[45] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool, "Coupled object detection and tracking from static cameras and moving vehicles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30(10), pp. 1683–1698, 2008.

[46] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(7), pp. 1239–1258, 2010.

[47] V. Eiselein, T. Senst, I. Keller, and T. Sikora, "A motion-enhanced hybrid probability hypothesis density filter for real-time multi-human tracking in video surveillance scenarios," in *Proc. of PETS*, 2013, pp. 6–13.

[48] D. Hall and J. Llinas, *Handbook of multisensor data fusion*, ser. Electrical Engineering & Applied Signal Processing Series. CRC Press Inc, June 2001.

[49] G. Gan and J. Cheng, "Pedestrian detection based on hog-lbp feature," in *Proc. of CIS*, 2011, pp. 1184–1187.

[50] T. Ojala and M. Pietikainen, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(7), pp. 971–988, 2002.

[51] A. Garcia-Martin, A. Cavallaro, and J. M. Martinez, "People-background segmentation with unequal error cost," in *Proc. of ICIP*, 2012, pp. 157–160.

[52] L. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(2), pp. 281–286, 2002.

[53] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. of CVPR*, 2005, pp. 878–885.

[54] A. Garcia-Martin, J. M. Martinez, and J. Bescos, "A corpus for benchmarking of people detection algorithms," *Pattern Recognition Letters*, vol. 33(2), pp. 152–156, 2012.

[55] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

[56] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34(4), pp. 743–761, 2012.

[57] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(12), pp. 2179–2195, 2009.

[58] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28(11), pp. 1863–1868, 2006.

[59] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. of CVPR*, 2009, pp. 794–801.

[60] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 4," http://people.cs.uchicago.edu/ rbg/latent-release4/.