IET Computer Vision

Review Article

IET Journals
The Institution of Engineering and Technology

People detection in surveillance: classification and evaluation

ISSN 1751-9632
Received on 18th June 2014
Revised on 23rd February 2015
Accepted on 14th March 2015
doi: 10.1049/iet-cvi.2014.0148
www.ietdl.org

Álvaro García-Martín ✉, *José María Martínez*

Video Processing and Understanding Lab, Departamento de Tecnología Electrónica y de las Comunicaciones, Universidad Autónoma Madrid, Escuela Politécnica Superior, Madrid 28049, Spain
✉ *E-mail: alvaro.garcia@uam.es*

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. The main objective of this study is to give a comprehensive and extensive evaluation of the state of the art of people detection regardless of the final surveillance application. For this reason, first, the different processing tasks involved in the automatic people detection in video sequences have been defined, then a proper classification of the state of the art of people detection has been made according to the two most critical tasks, object detection and person model, that are needed in every detection approach. Finally, experiments have been performed on an extensive dataset with different approaches that completely cover the proposed classification and support the conclusions drawn from the state of the art.

1 Introduction

Computer vision has been an evolving field for the past years with multiple lines of research and different application domains. Video surveillance has been one of the most developed domains for the past 10 years [1–4]. The need for providing security to people and their properties in the entire world explains the huge development and expansion of video surveillance systems nowadays. Video surveillance systems try to automatically extract information from the video sequence and to generate a scene description useful for human interactions with the system: alarms, logs, statistics, indexing and retrieval etc.

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 surveillance systems. Automatic people detection in video sequences [5–8] is one of the most challenging problems in computer vision. The complexity of the people detection problem is mainly based on the difficulty of modelling people 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 typical real-world surveillance scenarios such as airports, malls etc., which often include multiple people, multiple occlusions and background variability.

There are a large number of people detection surveys in the literature, some of them partially cover only the state of the art or are clearly focused on some particular video surveillance application. Enzweiler and Gavrila [5] present a survey of people detection and also the integration of the detectors into onboard full systems. It decomposes people detection approaches into three processing tasks: generation of initial object hypotheses or regions of interest (ROIs) selection, verification (classification) and temporal integration (tracking). Gerónimo *et al.* [6] also presents a survey of people detection but with a clear focus on driver assistance systems and defines a processing pipe line: pre-processing, foreground segmentation, object classification, verification or refinement, tracking and application. Simonnet *et al.* [8] presents an overview of people detection algorithms focused only on exhaustive search approaches, whereas Dollár *et al.* [7] present an overview focused only on sliding-window approaches.

In this paper, we present a state-of-the-art classification not focused on a particular video surveillance application. We decompose the people detection in subtasks, identify the critical tasks and classify the state of the art according to these critical tasks. In this way, we are able to analyse the strengths and weaknesses of each approach independently and for each critical task. Any other additional subtask is considered as a specific video surveillance application pre-processing or post-processing and they are not part of the scope of this review.

The main contribution presented in this paper is an overview and extensive evaluation of people detection state of the art, in general, video surveillance applications. Therefore, first, the different processing tasks that imply the automatic people detection in video sequences have been defined. Then, the critical tasks have been identified and proper classifications of the people detection approaches from the state of the art have been made according to those critical tasks. Each classification includes a brief discussion about advantages and disadvantages of different approaches to solve the people detection problem in video sequences. Finally, experiments are performed over an extensive dataset with different complexity categories and including different approaches that cover every people detection issue identified from the state of the art.

The remainder of this paper is structured as follows: Section 2 presents a brief review of the state of the art, Section 3 describes the basic architecture of every people detector surveillance system, Section 4 presents the proposed classification of people detection state of the art, Sections 5 and 6 describe the performance evaluation methodology and experimental results, and, finally, the main conclusions are summarised in Section 7.

2 Architecture of people detection systems

As defined for canonical surveillance systems [2, 9], every people detection approach consists mostly of, first, the design and training (if training is required) of a person model based on characteristic parameters (motion, dimensions, silhouette etc.) and, second, 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, whereas all the others will not be detected or classified as person.

2.1 Input

There are many different possible input formats, which determine the type of input information available to the detector. In relation to computer vision, the basic processing input unit is the image or the frame in the case of video processing. Input images can be of multiple resolutions, two-dimensional (2D) or 3D, colour or grey scale, visible or infrared spectrum etc. Input videos can be from static or mobile cameras, mono or stereo-vision etc.

2.2 Object detection

Object detection consists in the generation or extraction from the scene of the initial object hypotheses, that is, candidates to be a person. This is a critical task for people detection. The chosen approach (e.g. background subtraction, sliding-window) will be very determinant for some global detection performance factors: processing speed, detection results, robustness to scene variations etc.

2.3 Person model

The person model defines the characteristics and rules that the objects must meet in the scene in order to be considered as people. Such as the previous step, this is also a critical task for people detection. The chosen approach (e.g. holistic, part-based) will be very determinant in some global detection performance factors: processing speed, robustness to pose variations, partial occlusions etc.

2.4 Verification or classification

The verification or classification task can be considered as a standard pattern recognition issue. This process compares previously trained object models and the generated object model from an image or sequence.

2.5 Decision

According to the comparison or similarity calculated in the previous stage, a final decision must be taken. Depending on the subsequent application, the decision may be binary (person or no person) or fuzzy (a confidence value or probability of being a person).

3 Proposed classification of state-of-the-art people detection

Many criteria can be used to classify people detection algorithms; for example, the techniques used (e.g. background or foreground extraction, movement estimation or compensation), the type of models used (e.g. stick figure-based, statistical, movement), the use of 2D or 3D information, the sensor modality (e.g. visible light, infrared), the sensor multiplicity (monocular, stereo or multicamera), the sensor placement (centralised against distributed), the sensor mobility (stationary against moving) etc.

As already mentioned in the previous section, the two main critical tasks of people detection (object detection and person model) determine the global detection performance; therefore it has been decided to propose a classification of the state-of-the-art algorithms according to these tasks [Any classification system could be perfectly debated because it depends on the discriminative aspects on which its hierarchy is based.]. In the remainder of this section, we describe the classification of different algorithms from the state of the art. First, we classify the people detection algorithms according to the approach used to generate or extract the initial candidate objects to be a person, whereas the second classification is based on the chosen person model (see Table 1).

Table 1 State-of-the-art people detection classification according to the two main critical tasks of people detection: object detection and person model

Object detection	Person model		
	Motion	Appearance	
		Holistic	Part-based
segmentation	[10, 11]	[11–19]	[20–24]
exhaustive search	[25–40]	[25, 26, 28–31, 34–49]	[23, 24, 32, 33, 36, 50–52]

3.1 Object detection approach or initial object hypotheses

There are two main conventional object detection approaches (see Fig. 1): those based on some kind of segmentation of the scene in foreground (objects) and background [10–22] and those based on an exhaustive scanning approach [25–52]. There are also some approaches that try to combine both approaches together [23, 24]. In any case, the result of this stage is the location and dimension (bounding box or blob [In the literature, both terms have been used without any distinction, for the rest of this paper we also use both without any distinction.]) of the different objects in the scene candidates to be a person. Table 2 summarises the different approaches from the state of the art according to the used object detection approach.

3.1.1 Segmentation: Currently, there are many approaches from the state of the art that use some kind of segmentation as a preliminary step in the people detection task. In particular, the use of background subtraction is very popular in surveillance applications [10, 13–15, 18–20, 24]. They try to detect moving objects from the difference between the current frame and a reference frame (background model) and threshold the results to generate the objects of interest. There are some approaches that use colour segmentation [21, 22], owing to the fact that the skin colour facilitates the people segmentation and detection process. There are multiple approaches that use some kind of 3D information to facilitate the segmentation by stereo-vision [11, 16, 23] or directly with 3D cameras [12, 17].

In relation to people detection, the use of segmentation directly generates the objects candidates to be a person and easily rejects irrelevant areas of the image, that is, without objects of interest. For this reason, the subsequent classification task is clearly simplified and, therefore, the person model is usually simpler and has lower computational cost. However, as there is a strong dependence on the segmentation, all the segmentation problems are inherited (under and over segmentations). These problems can affect the global detection performance, mainly limiting the

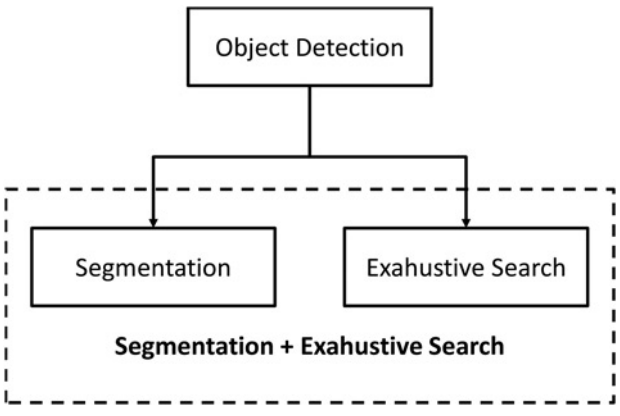

Fig. 1 *People detection classification according to the object detection approach*

IET Comput. Vis., 2015, Vol. 9, Iss. 5, pp. 779–788

Table 2 State-of-the-art people detection classification according to the object detection approach

Approach	Segmentation	Exhaustive search
[10, 13–15, 18–20]	background subtraction	—
[24]	background subtraction	sliding-window
[21, 22]	colour information	—
[11, 12, 16, 17]	3D information	—
[23]	3D information	bounded sliding-window
[26–30, 32, 34, 35, 37–39, 42, 43, 45, 46, 48–50, 52]	—	sliding-window
[36, 40]	—	sliding-window or feature-based
[31, 33, 41, 44, 47, 51]	—	feature-based

maximum detection rate (undetected objects) but also increasing the number of false detections (partial object detections or overlapping objects). Furthermore, these problems are magnified in complex scenarios where it is quite difficult to obtain a reliable segmentation.

3.1.2 Exhaustive search: The other technique to obtain initial object location hypotheses is the exhaustive search. Usually, it consists of scanning the full image looking for similarities with the chosen person model at multiple scales and locations. Through this mechanism, a dense detection confidence map or volume (scale and location) is obtained; in order to arrive at individual detections, these approaches must search for local maxima in the density volume and then apply some form of non-maximum suppression.

There are many people detection approaches from the state of the art that use this technique; in fact, this technique is currently the most widely used. Within this technique, two different approaches can be used as stated in [37]. On the one hand, there are some approaches that obtain this density volume implicitly sampling in a discrete 3D grid (location and scale) by evaluating different detection windows with a classifier; this is the case of using sliding-window-based detectors such as [23–30, 32, 34–40, 42, 43, 45, 46, 48, 49, 50, 52]. On the other hand, there are some approaches that create this density volume explicitly in a bottom-up fashion through probabilistic votes cast by local features matching; this is the case of using feature-based detectors such as [31, 33, 36, 40, 41, 44, 47, 51].

Generally, those detectors that use these kinds of approaches are more robust to scale and pose variations and, therefore, more reliable in complex environments than those based on segmentation. However, unlike in the previous case, the classification task is not simplified; it is even more complex because the person model must be able to classify a great number of negative examples correctly (potential false positive detections). In addition to the increased person model complexity, the exhaustive search process itself usually requires a higher computational cost, which makes it difficult to fulfil real-time requirements. Although some proposals have studied this problem [45, 46, 48], many irrelevant candidates are still passed to the next step, which increase the potential number of false positives.

3.1.3 Segmentation and exhaustive search: Another approach is the combination of both techniques trying to leverage their strengths and address its drawbacks. In [23], an initial selection of candidates is performed using segmentation with 3D information and then a second selection is performed using exhaustive search but because of computational efficiency only around the centre of those pre-selected candidates, that is, bounded sliding-window. In [24], the initial objects candidates to be person are extracted using background subtraction and then those selected candidates are processed with an exhaustive search, in this case with a full exhaustive search over the selected candidates.

3.1.4 Conclusions: Both approaches aim at the generation or extraction of the initial object hypotheses (candidates to be a person) in the scene. Therefore they extract ROIs from the image to be sent to the next processing module, avoiding as many background regions as possible. These techniques are of remarkable importance to reduce the number of candidates to be processed in the following stages, however, always keeping a balance between the number of candidates and the number of missing people. Otherwise, the number of false positive detections could be drastically increased or the subsequent modules will not be able to detect these missing people.

The segmentation approach greatly facilitates the subsequent classification task but it is affected by the inherited problems of the segmentation. In contrast, the exhaustive search approach provides a more robust candidate extraction, at the cost of increasing the subsequent classification task complexity and the global computational cost. The combination of both techniques can be a solution to merge their strengths and reduce their weaknesses.

3.2 Person models

As we have already commented, the verification or classification process applies a previously defined or trained person model to the objects candidates to be a person from an image or sequence and takes a final decision based on their similarity. Therefore the definition of a proper person model is a critical task for the verification or classification process. There are two main discriminative information sources to characterise the people model: appearance and motion (see Fig. 2). In any case, the model should be able to discriminate between people and any other object in the scene. Table 3 summarises the different approaches from the state of the art according to the used person model information.

3.2.1 Based on 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 because of 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 and to perform the detection using only motion information [10, 27, 40].

Within this classification, Cutler and Davis [10] propose an object classification system based on periodic motion analysis. The algorithm segments the motion, tracks objects in the foreground, aligns each object along time and finally computes the

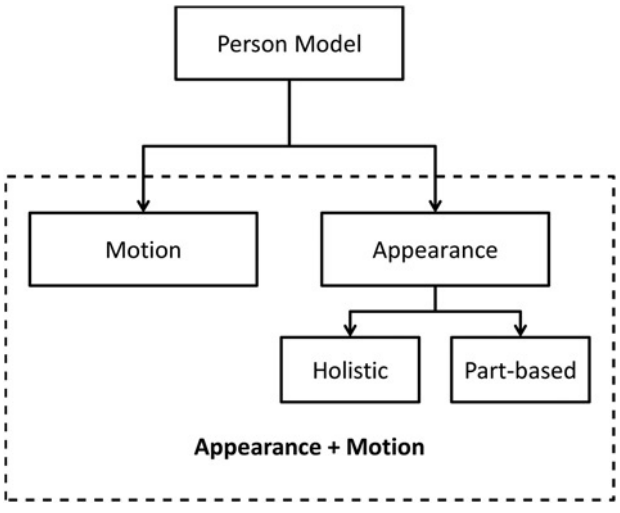

Fig. 2 *People detection classification according to the person model approach*

Table 3 State-of-the-art people detection classification according to the person model approach

Approach	Motion	Appearance	
		Holistic	Part-based
[10]	periodic motion	—	—
[27]	flow patterns	—	—
[12–19]	—	silhouette	—
[42]	—	Haar-like features	—
[43, 45, 46, 48]	——	HOG	—
[41, 44, 47]	—	ISM	—
[20]	—	—	silhouette
[21, 22]	—	—	colour distribution
[23]	—	—	Canny/Haar/…/ features
[52]	—	—	HOG
[49]	—	—	HOG/gradient/LUV
[24, 50]	—	—	Edgelets
[51]	—	—	ISM
[25, 30]	Haar-like features multi-frame	Haar-like features	—
[28]	HOG multi-frame	HOG	—
[11]	tracking	silhouette	—
[26, 34, 35]	tracking	Haar-like features	—
[29, 37–39]	tracking	HOG	—
[31]	tracking	ISM	—
[40]	IMM and tracking	HOG or ISM	—
[36]	tracking	HOG or ISM	HOG
[32]	tracking	—	Edgelets
[33]	tracking	—	ISM

self-similarity between objects and how it evolves in time. Sidenbladh [27] proposes a people detection system based on detecting people motion patterns. For each object present in two consecutive images, size normalisation is performed and its flow pattern is calculated that consists of dense optical horizontal and vertical flows. Another approach based on motion information [40] proposes a detection algorithm, the implicit motion model (IMM), based on the characteristic movements of people using the implicit shape model (ISM) framework and the MoSIFT interest points detector and descriptor.

In relation to people detection, methods based on motion usually obtain worst results than methods based on appearance but they are independent of appearance variability. They do not support partial occlusions because in this case we could not extract motion patterns correctly. For these reasons they can only be considered either complementary information or essential in specific scenarios where methods based on appearance do not work (e.g. bad illumination, small objects).

3.2.2 Based on appearance: 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, that is, holistic models [11–19, 25, 26, 28–31, 34–49] and more complex models that define the person as combination of multiple regions or shapes, that is, part-based models [20–24, 32, 33, 36, 50–52].

Within this classification (see Table 3), there are different chosen characteristics to define the people appearance, both holistic and part-based models. There are some approaches that extract the object silhouette and classify the object according to their similarity with reference people silhouettes or certain standards that the silhouette must meet. Some approaches make use of the colour distribution in a person (where the skin colour is essential) to determine if the object is a person or not. However, the most popular approaches are those that define the people appearance according to their characteristic edge information using some kind of shape descriptor: Haar-like features [23, 25, 26, 30, 34, 35, 42], histogram of oriented gradients (HOGs) [28, 29, 36–40, 43, 45, 46, 48, 52], Edgelets [24, 32, 50], ISM [31, 33, 36, 40, 41, 44, 47, 51] or combination of multiple features, aggregate channel features: HOG, gradient and colour (ACF) [49].

Generally, those detectors based on a simplified or holistic person model have lower complexity but do not support partial occlusions or pose variations. If you cannot see the whole region or shape, the model does not work properly. On the other hand, those detectors based on a more complex or part-based person model usually have higher complexity but they support partial occlusions and pose variations.

3.2.3 Based on appearance and motion: 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 [25, 28, 30]; 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 [11, 26, 29, 31–40]. Most approaches from the state of the art that combine detection and tracking are designed mainly with the aim of improving tracking results (tracking-by-detection). However, there are some approaches that try to improve or update explicitly the detection using the tracking history (detection-by-tracking). 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, Garcia-Martin and Martinez [40] 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.

3.2.4 Conclusions: As we have already commented, there are few approaches based only on motion information. Their main advantages are that they are independent of appearance variability and usually have low complexity. However, they usually have worst results and they do not support partial occlusions.

The methods based on holistic person models (only a region or shape) usually have lower complexity but they neither support partial occlusions nor pose variations. However, the methods based on part-based people models usually have higher complexity but they support partial occlusions and pose variations. Another advantage is that they make the final decision by combining multiple evidences, so they are usually more reliable than methods based on holistic human models. For these reasons, they usually have better results. Whichever holistic or part-based people models, the combination of multiple features is another option to make the final decision by combining multiple evidences, so they usually have better results than those based on only one feature.

Motion information can add robustness to the appearance model without adding too much complexity to the detection or even can be essential in specific scenarios where methods based on appearance do not work (e.g. tracking information could be very discriminant in complex scenarios which usually include multiple people, multiple occlusions and background variability).

Table 4 Sequences categorisation evaluation datasets

Category	#Sequences		Complexity	
	Dataset A	Dataset B	Classification	Background
C1	6	0	low	low
C2	6	0	medium	low
C3	4	0	medium	medium
C4	5	0	high	low
C5	8	61	high	high

IET Comput. Vis., 2015, Vol. 9, Iss. 5, pp. 779–788

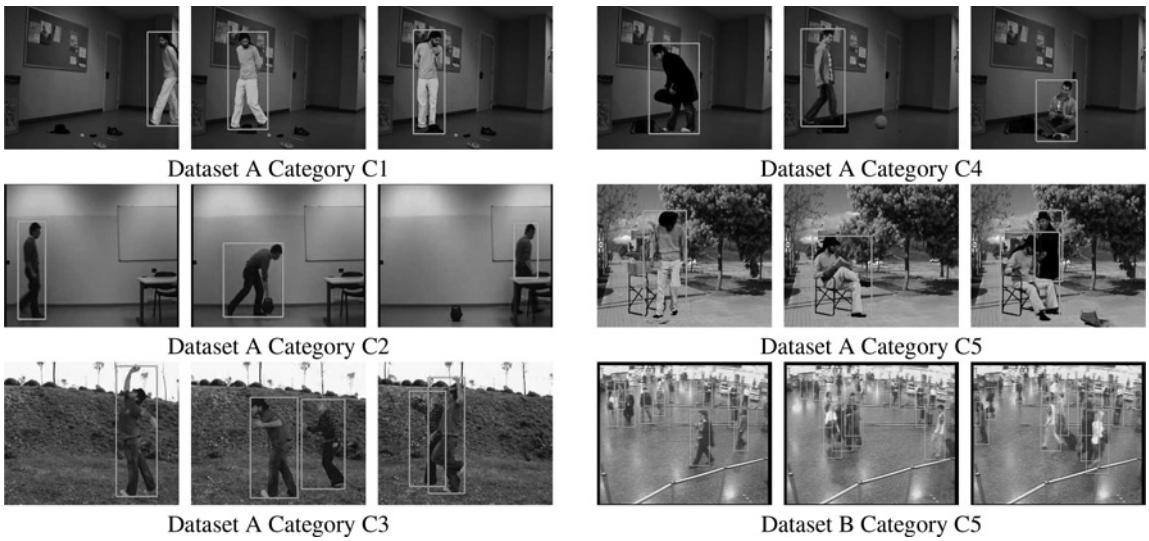

Fig. 3 *Experimental dataset examples*
Every example shows three random frames from a sequence

4 Performance evaluation methodology

This section describes the experimental setup or evaluation methodology. To define a proper evaluation methodology, it is necessary to define the chosen evaluation video corpus (or dataset) and the chosen evaluation metrics.

4.1 Evaluation dataset

The chosen experimental corpus, person detection datasets [53], mainly excels other datasets in the amount of sequences (90 videos, 28 358 frames) and variability of sequences. It includes sequences of different lengths from just seconds to a few minutes. It has been divided into two evaluation datasets. The first dataset, named A, has been selected to evaluate the different approaches at every complexity level; it includes the first 29 sequences from the experimental corpus. These sequences include five different complexity categories depending on the defined people detection critical factors. The experimental dataset includes both non-rigid and rigid people and objects differing in size, motion and textural appearance. These people and objects are involved in a number of interactions and in different contexts, such as typical everyday situations or surveillance video scenarios. Regarding the backgrounds, it includes indoor and outdoor scenarios with different background complexities.

The second dataset, named B, has been selected to evaluate more thoroughly the category with the highest complexity, that is, category C5. It includes the following 61 sequences from the experimental corpus. The sequences have been extracted from the TRECVID 2008 dataset [54], namely, the ones for the surveillance TRECVID event detection task recorded at London Gatwick International Airport. This dataset contains highly crowded scenes, severely cluttered background, people at different scales and people completely static along the whole sequences. Owing to the small size of the objects at the top of the image, during the annotation of sequences, the top 15% of the images has been discarded.

A summary of the complexity levels of both evaluation datasets is shown in Table 4. In addition, Fig. 3 shows some example frames from several sequences of the experimental datasets A and B, including annotated ground truth.

4.2 Evaluation metrics

To evaluate different people detection approaches, we need to quantify the different 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 [7, 43] or receiver operating characteristics (ROCs) [5, 55] curves. Global sequence performance is usually measured in terms of precision–recall (PR) curves [33, 47, 56]. The first level gives us information about the classification stage, whereas the second one provides overall system performance information. 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.

ROC curves represent the fraction of true positives out of the positives (true positive rate, recall or sensitivity) against the fraction of false positives out of the negatives (false positive rate or 1-Specificity). We aim to evaluate and compare the overall performance of different detection systems, so we have chosen the second evaluation method. For each value of the detection confidence, PR curves compute precision and recall as (1) and (2) (see (1) and (2))

To evaluate not only the yes/no detection decision but also the precise people locations and extents, we use three evaluation criteria, defined by Leibe *et al.* [57], that allow to compare hypotheses at different scales: relative distance, cover and overlap.

$$\text{precision} = \frac{\#\text{true positive people detections}}{\#\text{true positive people detections} + \#\text{false positive people detections}} \quad (1)$$

$$\text{recall} = \frac{\#\text{true positive people detections}}{\#\text{true positive people detections} + \#\text{false negative people detections}} \quad (2)$$

The relative distance (dr) measures the distance between the bounding box centres in relation to the size of the annotated bounding box. Cover and overlap measure how much of the annotated bounding box is covered by the detection hypothesis and vice versa (see Fig. 4). A detection is considered true if d$r \leq$ 0.5 (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%. Only one hypothesis per object is accepted as correct, so any additional hypothesis on the same object is considered as a false positive.

The integrated average precision is generally used to summarise the overall performance, represented geometrically as the area under the PR curve (AUC-PR); in order to express the results more clearly, we have chosen the representation recall against 1-precision (see Fig. 5). To approximate correctly the area, we use the approximation described by [58].

5 Experimental results

In this section, we describe the experiments performed over the experimental dataset and including different approaches that cover all the people detection issues identified from the state of the art. We have selected eight diverse people detection approaches: Edge [24], Fusion [18], HOG [43], ISM [57], TUD [51], DTDP [52], ACF [49] and IMM [40]. According to the chosen object detection approach, Edge combines segmentation and exhaustive search, Fusion is based only on segmentation 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 (Fusion, HOG, ISM and ACF) or part-based (Edge, TUD and DTDP). An overview of the selected people detection approaches is shown in Table 5.

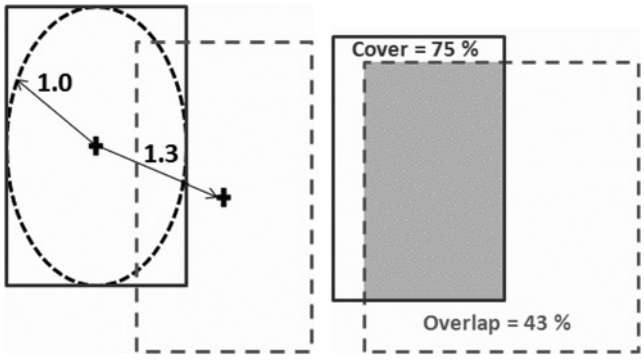

Fig. 4 *Evaluation criterion for comparing bounding boxes [57]: (left) relative distance and (right) cover and overlap*

Table 5 People detection approaches selected for the experimental evaluation and classification according to the two main critical tasks of people detection: object detection and person model

Object detection	Person model		
	Motion	Appearance	
		Holistic	Part-based
segmentation		Fusion [18]	Edge [24]
exhaustive search		HOG [43]	Edge [24]
		ISM [57]	TUD [51]
	IMM [40]	ACF [49]	DTDP [52]
		IMM [40]	IMM [40]

The Edge, Fusion and IMM results have been obtained with the original code, the HOG results have been obtained using the available binaries [http://www.pascal.inrialpes.fr/soft/olt/], the ISM results have been obtained using the available code and binaries [http://www.vision.ee.ethz.ch/bleibe/index.html], the TUD results have been obtained using the available code [http://www.d2.mpi-inf.mpg.de/andriluka_cvpr09], the DTDP results have been obtained using the available code [http://www.cs.brown.edu/pff/latent/] and the ACF results have been obtained using the available code [http://www.vision.ucsd.edu/pdollar/toolbox/doc/index.html].

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 and Fusion are mainly limited by the segmentation step. Moreover, HOG, ISM, TUD, DTDP, ACF and IMM are limited by the image scanning.

5.1 Evaluation dataset A

First, we evaluate and compare the appearance-based approaches at every complexity level using the evaluation dataset A. Fig. 6 shows the averaged detection performance in terms of recall against (1-precision) curves and Table 6 shows the results in terms of AUC-PR; in both cases, the results are for each complexity category included within the used video dataset A.

The results clearly show that all algorithms perform worst at higher complexity categories (from C1 to C5). However, it is

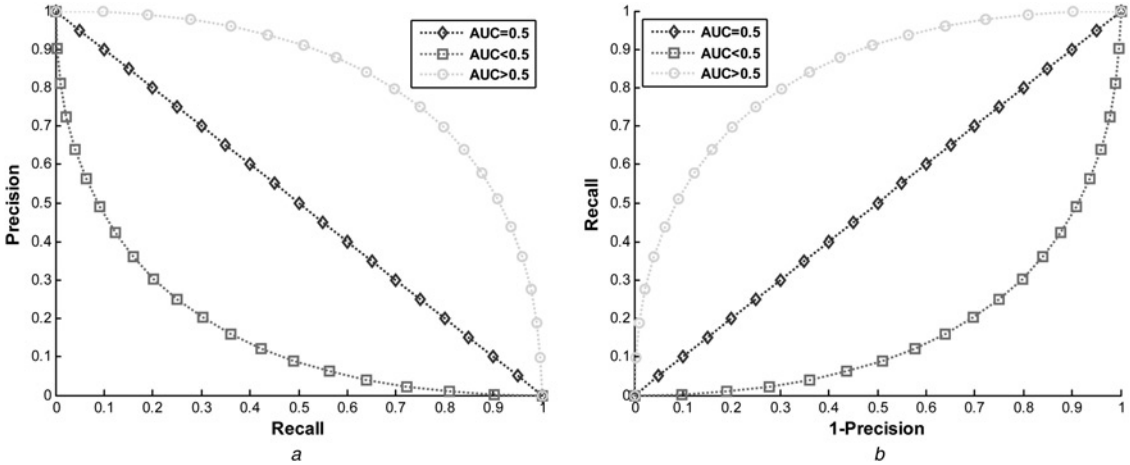

Fig. 5 *PR curves and area under the curve*

Equivalent representations
a Precision against recall representation
b Recall against 1-precision representation

IET Comput. Vis., 2015, Vol. 9, Iss. 5, pp. 779–788

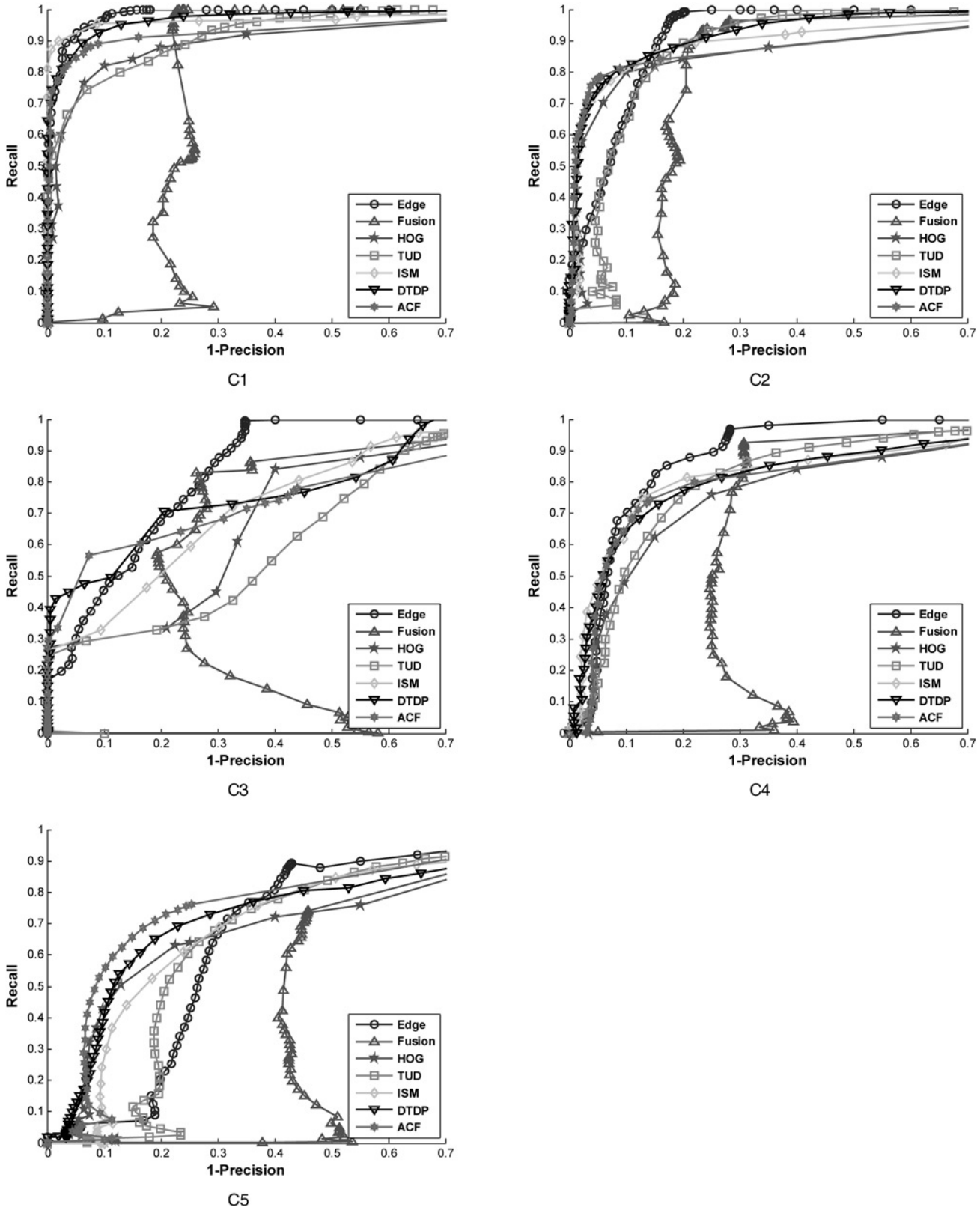

Fig. 6 *PR results per complexity category of evaluation dataset A*

observed that all approaches obtain generally worst results at category C3 than at category C4, because of the great influence of the background complexity in category C3 and, thus, the generation or extraction of the initial object hypotheses or candidates to be a person in the scene is more difficult. On the other hand, the complexity of the category C4 lies on the classification of those initial candidates.

The Fusion approach obtains the worst results. The use of segmentation makes the classification stage easier, allowing the approach to reach high recall results but the use of such a simplified person model and all the segmentation problems (under and over segmentations) reduce the global precision rate. The Edge approach obtains good results in all complexity categories and similar to the other approaches not based on segmentation. It

Table 6 AUC-PR average for each complexity category of evaluation dataset A

	Edge	Fusion	HOG	ISM	TUD	DTDP	ACF
C1	**0.98**	0.78 (−26)	0.92 (−7)	0.95 (−3)	0.93 (−5)	0.96 (−2)	0.94 (−4)
C2	**0.93**	0.81 (−15)	0.86 (−8)	0.91 (−2)	0.88 (−6)	0.92 (−1)	0.88 (−6)
C3	**0.85**	0.6(−41)	0.74 (−15)	0.8(−6)	0.75 (−13)	0.81 (−5)	0.8 (−6)
C4	**0.89**	0.69 (−28)	0.82 (−9)	0.84 (−6)	0.84 (−6)	0.86 (−3)	0.84 (−6)
C5	0.7 (−11)	0.48 (−63)	0.71 (−10)	0.71 (−10)	0.67 (−16)	0.74 (−5)	**0.78**

Percentage increase (%Δ) calculated with respect to the best result for each complexity category (in bold)

is because of the use of a more complex person model and the combination of segmentation and exhaustive search. Despite the fact that the combination of segmentation and exhaustive search reduces the segmentation problems, these problems are magnified in complex background scenarios (C3–C5), where it is quite difficult to obtain a reliable segmentation.

The exhaustive search approaches are more robust to scale and pose variations and, therefore, more reliable in complex environments than those based on segmentation. Even so, the background complexity still has a negative impact in the results (C3). Moreover, unlike in the previous case, the classification stage is not simplified; it is even more complex because the approach must deal with a great number of negative examples (potential false positive detections), reducing the recall rate in order to maintain the precision rate. The HOG and TUD approaches show similar results in all complexity categories but the ISM, DTDP and ACF obtain better results. The ISM is a holistic approach but with a flexible person model based on spatial feature probability distribution, the ACF is also a holistic approach but based on the combination of multiple features and the DTDP is a body part-based variation of the HOG approach.

5.2 Evaluation dataset B

In this section, we evaluate the highest complexity category (C5) more thoroughly using dataset B. Table 7 shows the results in terms of AUC-PR of dataset B. Owing to the greater complexity of the sequences extracted from TRECVID (the content set contains challenging scenarios, crowds and a wide range of scales), the results are worst than those obtained in dataset A.

In this case, because of the higher sequences complexity, all the approaches get worst results than with dataset A. Both approaches based on segmentation, the Edge and Fusion, obtain worst results than the other approaches from the state of the art. As already commented, the main problem of these approaches is the difficulty of making a reliable segmentation (foreground or background) in complex scenarios. However, the sequences extracted from TRECVID present an additional difficulty to both approaches: the sequences include completely static people along the whole sequences. Both approaches extract the objects candidates to be a person using motion information (background subtraction), being able to extract static objects, which reduces the recall rate and, therefore, the overall performance.

The results also show that the approaches based on exhaustive search also get worst results than with dataset A. However, except the TUD approach, they are more stable in more complex scenarios because they are more robust to scale and pose variations and more robust to the background complexity.

Table 7 AUC-PR average of evaluation dataset B

	Edge	Fusion	HOG	ISM	TUD	DTDP	ACF
C5	0.59 (−22)	0.44 (−64)	0.66 (−9)	0.69 (−4)	0.56 (−29)	0.68 (−6)	**0.72**

Percentage increase (%Δ) calculated with respect to the best result (in bold)

5.3 Evaluation dataset B with motion

In this section, we evaluate dataset B including the people detector based on motion IMM and all the appearance and motion combinations (Edge + IMM, Fusion + IMM, HOG + IMM, ISM + IMM, TUD + IMM, DTDP + IMM and ACF + IMM). To train the people motion model, the evaluation dataset B has been divided in training and test. To be homogeneous, the appearance-based detectors approaches have also been evaluated on the same video sequences, the test dataset. As in the experiments in [40], the training dataset is composed of 25 sequences and the test dataset is composed of the other 36 sequences. Table 8 shows the results in terms of AUC-PR of test dataset.

The results show that the IMM approach obtains good results in complex or realistic scenarios and comparable with the other approaches from the state of the art. The IMM is based only on motion, so it is only able to detect moving people. For this reason, the IMM approach, in general, is able to obtain high precision rates but low recall rates. Even so, in environments as complex as these ones, the use of motion information obtains results close to the use of appearance information. The combination of appearance and motion information (Edge + IMM, Fusion + IMM, HOG + IMM, ISM + IMM, TUD + IMM, DTDP + IMM and ACF + IMM) improves the global results in all the cases. Thus, it is clear that human motion provides useful information for people detection and independent from appearance information.

5.4 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, Fusion and ACF), near real (HOG and DTDP) or no real-time (ISM, IMM and TUD). The tests have been performed on a Pentium IV with a central processing unit frequency of 2.4 GHz and 3 GB random access memory.

The Edge detector [24] 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 [50]. The Edge approach [24] is implemented in C++ (OpenCV) and the computational cost is about 0.02 s per frame with 352 × 288 images.

The Fusion detector [18] is a real-time detection approach based on segmentation and a holistic person model. The initial objects candidates to be person are extracted using background subtraction and the holistic person model is the combination or fusion at decision level of three simple person models: ellipse fitting [12], ghost [59] and aspect ratio. The Fusion approach is implemented in C++ (OpenCV) and the computational cost is about 0.02 s per frame with 352 × 288 images.

The ACF detector proposes a very fast exhaustive search and a holistic person model using ACFs. The ACF approach [49] is implemented in MATLAB and the computational cost is about 0.02 s per frame with 352 × 288 images.

The HOG detector [43] is based on exhaustive search and a holistic person model using the HOGs. It consists of 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 [43] is implemented in C ++ and the computational cost is about 1 s per frame with 352 × 288 images (there is a faster implementation in OpenCV that runs about 0.1 s per frame).

The DTDP detector [52] is based on exhaustive search and a part-based person model. The DTDP approach [52] is implemented with MATLAB and the computational cost is about

IET Comput. Vis., 2015, Vol. 9, Iss. 5, pp. 779–788

Table 8 AUC-PR average of evaluation dataset B without and with motion information

	Edge	Fusion	HOG	ISM	TUD	DTDP	ACF	IMM
C5 (%Δ^1)	0.58(−21)	0.46(−52)	0.66(−6)	0.64(−9)	0.56(−25)	0.67(−4)	**0.7**	0.6(−17)
	Edge + IMM	Fusion + IMM	HOG + IMM	ISM + IMM	TUD + IMM	DTDP + IMM	ACF + IMM	
C5 (%Δ^2)	0.62(+7)	0.49(+7)	0.68(+3)	0.67(+5)	0.62(+11)	0.7(+4)	**0.72**(+3)	

Percentage increase (%Δ^1) calculated with respect to the best result or percentage increase (%Δ^2) calculated with respect to single appearance versions (in bold)

2 s per frame with 352 × 288 images (there is a faster implementation in OpenCV that runs about 1 s per frame).

The ISM people detector [57] is based on exhaustive search and a holistic person model. It consists of 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 [40] is a variation of the ISM detector where the chosen person model is based on the characteristic movements of people using the MoSIFT features. Both approaches have been implemented with C++ and have similar computational cost between 4 and 7 s per frame with 352 × 288 images.

The TUD people detector [51] is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original ISM detector [57] using pictorial structures. The TUD approach [51] is implemented with MATLAB subroutines and C++, the computational cost is several orders of magnitude greater than the other approaches.

A summary of the selected people detection approaches is shown in Table 9.

6 Summary and conclusions

In this paper, extensive classification and evaluation of automatic people detection in video sequences have been presented. First, the different processing tasks used for automatic people detection have been analysed. Then, a complete classification of the people detection approaches from the state of the art has been made regardless of their subsequent video surveillance application. Finally, experiments have been performed over an extensive dataset with different complexity categories and dealing with every people detection issue identified from the state of the art. This section sums up some conclusions extracted from our work.

As already explained in Section 2, the people detection task consists mostly of, first, the design and training of a person model based on characteristic parameters (motion, dimensions, silhouette etc.) and, second, the adjustment of this model to the candidate objects in the scene. Thus, the critical tasks in any people detection algorithm are the generation or extraction of the initial object hypotheses to be people from the scene and the person model used to classify those initial object hypotheses.

The object detection approach has a great influence on the final people detection results. First, every object not extracted during this stage cannot be classified as person. Moreover secondly, a poor initial object extraction makes it more difficult the later classification. Segmentation is a simple and powerful object extraction technique but with all their difficulties and limitations in complex environments. In contrast, the exhaustive search is more robust to rotation, scale and pose changes even in complex environments but has the complexity of adding many false examples to the classification task, in addition to a higher computational cost.

The chosen person model to classify initial objects candidates to be person determines the robustness of the algorithm to person variations and occlusions. Simple models based only on motion or holistic appearance models are less robust to people variations and occlusions, whereas more complex part-based models add complexity to the algorithm but they are much more robust to people variations and occlusions. Finally, the adequate combination of appearance and motion can improve the detection results.

The experimental results over the evaluation dataset show the people detection problems in video sequences. According to the chosen object detection approach, the use of segmentation makes easier the classification stage. However, they must deal with all the segmentation problems (under and over segmentations). The combination of segmentation and exhaustive search reduces these problems but they are still a drawback especially in complex scenarios where these problems are magnified. The exhaustive search approaches are more reliable in complex environments than those based on segmentation. However, unlike in the previous case, the classification task is not simplified, it is even more complex because the approach must deal with a great number of negative examples (potential false positive detections), reducing the recall rate in order to maintain the precision rate. According to the chosen person model, in general, the use of simplified person models gets worst results mainly in terms of precision than those more complex person models. Moreover finally, the motion information is less discriminant than the appearance of the people but the combination of motion and appearance shows to be useful even in complex scenarios.

In general, within simple or controlled scenarios all algorithms, including those working on real-time, achieve acceptable results. However, in more complex scenarios, the algorithms that usually have better results are based on exhaustive search methods and of course a person model based on appearance. According to the results, the ACF detector obtains the best results in realistic scenarios. However, in presence of many partial occlusions such

Table 9 Option1: selected people detection approaches summary according to the chosen object detection approach, the person model or discriminative information source and the computational cost

Approach	Object detection		Person model			Computational cost
	Segmentation	Exhaustive search	Motion	Appearance		
				Holistic	Part-based	
Edge [24]	background subtraction	sliding-window	—	—	Edgelets	real-time
Fusion [18]	background subtraction	—	—	silhouette	—	real-time
HOG [43]	—	sliding-window	—	HOG	—	near real-time
ISM [57]	—	feature-based	—	ISM	—	no real-time
TUD [51]	—	feature-based	—	—	ISM	no real-time
DTDP [52]	—	sliding-window	—	—	HOG	near real-time
ACF [49]	—	sliding-window	—	HOG	—	real-time
IMM [40]	—	feature-based	IMM	HOG or ISM		no real-time

as groups of people, the DTDP detector or a part-based variation of the ACF detector will obtain the best results.

In the future, people detection must evolve into systems that allow to add robustness to the detection by the use of any additional information. For example, the use of multi cameras, 2.5D or 3D systems in order to deal with occlusions. Following the scheme of the ACF detector, the combination of multiple features or 'channels' improves the final detection. Finally, we have already discussed that the motion information can be very useful. Therefore the combination of several sources such as appearance, motion, tracking and multi-view could be the solution to uncontrolled and complex scenarios.

7 Acknowledgment

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

8 References

1 Platanioitis, K.N., Regazzoni, C.S.: 'Visual-centric surveillance networks and services', *IEEE Signal Process. Mag.*, 2005, **22**, (2), pp. 12–15
2 Valera, M., Velastin, S.A.: 'Intelligent distributed surveillance systems: a review', *IEE Proc. Vis. Image Signal Process.*, 2005, **152**, (2), pp. 192–204
3 Haering, N., Venetianer, P.L., Lipton, A.: 'The evolution of video surveillance: an overview', *Mach. Vis. Appl.*, 2008, **19**, (5–6), pp. 279–290
4 Regazzoni, C.S., Cavallaro, A., Wu, Y., Konrad, J., Hampapur, A.: 'Video analytics for surveillance: theory and practice', *IEEE Signal Process. Mag.*, 2010, **27**, (5), pp. 16–17
5 Enzweiler, M., Gavrila, D.M.: 'Monocular pedestrian detection: survey and experiments', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **31**, (12), pp. 2179–2195
6 Gerónimo, D., López, A.M., Sappa, A.D., Graf, T.: 'Survey of pedestrian detection for advanced driver assistance systems', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (7), pp. 1239–1258
7 Dollár, P., Wojek, C., Schiele, B., Perona, P.: 'Pedestrian detection: an evaluation of the state of the art', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (4), pp. 743–761
8 Simonnet, D., Velastin, S., Turkbeyler, E., Orwell, J.: 'Backgroundless detection of pedestrians in cluttered conditions based on monocular images: a review', *IET Comput. Vis.*, 2012, **6**, (6), pp. 540–550
9 Hu, W., Tan, T., Wang, L., Maybank, S.: 'A survey on visual surveillance of object motion and behaviors', *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, 2004, **34**, (3), pp. 334–352
10 Cutler, R., Davis, L.S.: 'Robust real-time periodic motion detection, analysis, and applications', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (8), pp. 781–796
11 Giebel, J., Gavrila, D.M., Schnorr, C.: 'A Bayesian framework for multi-cue 3D object tracking'. Proc. of ECCV, 2004, pp. 241–252
12 Xu, F., Fujimura, K.: 'Human detection using depth and gray images'. Proc. of AVSS, 2003, pp. 115–121
13 Zhao, T., Nevatia, R.: 'Tracking multiple humans in complex situations', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004, **26**, (9), pp. 1208–1221
14 Zhou, J., Hoang, J.: 'Real time robust human detection and tracking system'. Proc. of CVPR, 2005, pp. 149–156
15 Hussein, M., Abd-Almageed, W., Ran, Y., Davis, L.: 'Real-time human detection, tracking, and verification in uncontrolled camera motion environments'. Proc. of ICVS, 2006, pp. 41–47
16 Gavrila, D.M., Munder, S.: 'Multi-cue pedestrian detection and tracking from a moving vehicle', *Int. J. Comput. Vis.*, 2007, **73**, (1), pp. 41–59
17 Koenig, N.: 'Toward real-time human detection and tracking in diverse environments'. Proc. of ICDL, 2007, pp. 94–98
18 Fernández-Carbajales, V., García, M.A., Martínez, J.M.: 'Robust people detection by fusion of evidence from multiple methods'. Proc. of WIAMIS, 2008, pp. 55–58
19 Kilambi, P., Ribnick, E., Joshi, A.J., Masoud, O., Papanikolopoulos, N.: 'Estimating pedestrian counts in groups', *Comput. Vis. Image Underst.*, 2008, **110**, (1), pp. 43–59
20 Haritaoglu, I., Harwood, D., Davis, L.S.: 'W4: real-time surveillance of people and their activities', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (8), pp. 809–830
21 Sprague, N., Luo, J.: 'Clothed people detection in still images'. Proc. of ICPR, 2002, pp. 585–589
22 Harasse, S., Bonnaud, L., Desvignes, M.: 'Human model for people detection in dynamic scenes'. Proc. of CVPR, 2006, pp. 335–354
23 Alonso, I.P., Llorca, D.F., Sotelo, M.A., *et al.*: 'Combination of feature extraction methods for SVM pedestrian detection', *IEEE Trans. Intell. Transp. Syst.*, 2007, **8**, (2), pp. 292–307
24 Garcia-Martin, A., Martinez, J.M.: 'Robust real time moving people detection in surveillance scenarios'. Proc. of AVSS, 2010, pp. 241–247
25 Viola, P., Jones, M.J., Snow, D.: 'Detecting pedestrians using patterns of motion and appearance'. Proc. of ICCV, 2003, pp. 734–741
26 Okuma, K., Taleghani, A., Freitas, N.D., Little, J.J., Lowe, D.G.: 'A boosted particle filter: multitarget detection and tracking'. Proc. of ECCV, 2004, pp. 28–39
27 Sidenbladh, H.: 'Detecting human motion with support vector machines'. Proc. of ICPR, 2004, pp. 188–191
28 Dalal, N., Triggs, B.: 'Human detection using oriented histograms of flow and appearance'. Proc. of ECCV, 2006, pp. 428–441
29 Avidan, S.: 'Ensemble tracking', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (2), pp. 261–271
30 Cui, X., Liu, Y., Shan, S., Chen, X., Gao, W.: '3D Haar-like features for pedestrian detection'. Proc. of ICME, 2007, pp. 1263–1266
31 Leibe, B., Schindler, K., Gool, L.V.: 'Coupled detection and trajectory estimation for multi-object tracking'. Proc. of ICCV, 2007, pp. 1–8
32 Wu, B., Nevatia, R.: 'Detection and tracking of multiple, partially occluded humans by Bayesian combination of Edgelet based part detectors', *Int. J. Comput. Vis.*, 2007, **75**, (2), pp. 247–266
33 Andriluka, M., Roth, S., Schiele, B.: 'People-tracking-by-detection and people-detection-by-tracking'. Proc. of CVPR, 2008, pp. 1–8
34 Li, Y., Ai, H., Yamashita, T., Lao, S., Kawade, M.: 'Tracking in low frame rate video: a cascade particle filter with discriminative observers of different life spans', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008, **30**, (10), pp. 1728–1740
35 Ren, X.: 'Finding people in archive films through tracking'. Proc. of CVPR, 2008, pp. 1–8
36 Ess, A., Leibe, B., Schindler, K., Gool, L.V.: 'Robust multiperson tracking from a mobile platform', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **31**, (10), pp. 1831–1846
37 Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: 'Online multi-person tracking-by-detection from a single, uncalibrated camera', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **33**, (9), pp. 1820–1833
38 Stalder, S., Grabner, H., Gool, L.V.: 'Cascaded confidence filtering for improved tracking-by-detection'. Proc. of ECCV, 2010, pp. 369–382
39 Yu, J., Farin, D., Schiele, B.: 'Multi-target tracking in crowded scenes'. Proc. of DAGM, 2011, pp. 406–415
40 Garcia-Martin, A., Martinez, J.M.: 'On collaborative people detection and tracking in complex scenarios', *Image Vis. Comput.*, 2012, **30**, (4), pp. 345–354
41 Leibe, B., Schiele, B.: 'Scale invariant object categorization using a scale-adaptive mean-shift search'. Proc. of DAGM, 2004, pp. 145–153
42 Viola, P., Jones, M.: 'Robust real-time face detection', *Int. J. Comput. Vis.*, 2004, **57**, (2), pp. 137–154
43 Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. Proc. of CVPR, 2005, pp. 886–893
44 Seemann, E., Schiele, B.: 'Cross-articulation learning for robust detection of pedestrians'. Proc. of DAGM, 2006, pp. 242–252
45 Zhu, Q., Yeh, M.C., Cheng, K.T., Avidan, S.: 'Fast human detection using a cascade of histograms of oriented gradients'. Proc. of CVPR, 2006, pp. 1491–1498
46 Zhang, W., Zelinsky, G., Samaras, D.: 'Real-time accurate object detection using multiple resolutions'. Proc. of ICCV, 2007, pp. 1–8
47 Leibe, B., Leonardis, A., Schiele, B.: 'Robust object detection with interleaved categorization and segmentation', *Int. J. Comput. Vis.*, 2008, **77**, (1–3), pp. 259–289
48 Wojek, C., Dorkó, G., Schulz, A., Schiele, B.: 'Sliding-windows for rapid object class localization: a parallel technique'. Proc. of DAGM, 2008, pp. 71–81
49 Dollár, P., Appel, R., Kienzle, W.: 'Crosstalk cascades for frame-rate pedestrian detection'. Proc. of ECCV, 2012, no. 645–659
50 Wu, B., Nevatia, R.: 'Detection of multiple, partially occluded humans in a single image by Bayesian combination of Edgelet part detectors'. Proc. of ICCV, 2005, pp. 90–97
51 Andriluka, M., Roth, S., Schiele, B.: 'Pictorial structures revisited: people detection and articulated pose estimation'. Proc. of CVPR, 2009, pp. 1014–1021
52 Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: 'Object detection with discriminatively trained part-based models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (9), pp. 1627–1645
53 Garcia-Martin, A., Martinez, J.M., Bescos, J.: 'A corpus for benchmarking of people detection algorithms', *Pattern Recognit. Lett.*, 2012, **33**, (2), pp. 152–156
54 TRECVID: 'Trecvid 2008 evaluation for surveillance event detection'. Available at http://www-nlpir.nist.gov/projects/trecvid/
55 Munder, S., Gavrila, D.M.: 'An experimental study on pedestrian classification', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (11), pp. 1863–1868
56 Wojek, C., Walk, S., Schiele, B.: 'Multi-cue onboard pedestrian detection'. Proc. of CVPR, 2009, pp. 794–801
57 Leibe, B., Seemann, E., Schiele, B.: 'Pedestrian detection in crowded scenes'. Proc. of CVPR, 2005, pp. 878–885
58 Davis, J., Goadrich, M.: 'The relationship between precision–recall and roc curves'. Proc. of ICML, 2006, pp. 233–240
59 Haritaoglu, I., Harwood, D., Davis, L.S.: 'Ghost: a human body part labeling system using silhouettes'. Proc. of ICPR, 1998, pp. 77–82