Robust Real Time Moving People Detection in Surveillance Scenarios

Álvaro García-Martín, José M. Martínez
Video Processing and Understanding Lab
Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain
{alvaro.garcia,josem.martinez}@uam.es

Abstract

In this paper an improved real time algorithm for detecting pedestrians in surveillance video is proposed. The algorithm is based on people appearance and defines a person model as the union of four models of body parts. Firstly, motion segmentation is performed to detect moving pixels. Then, moving regions are extracted and tracked. Finally, the detected moving objects are classified as human or non-human objects. In order to test and validate the algorithm, we have developed a dataset containing annotated surveillance sequences of different complexity levels focused on the pedestrians detection. Experimental results over this dataset show that our approach performs considerably well at real time and even better than other real and non-real time approaches from the state of art.

1. Introduction

In the last years signal processing has been in constant evolution. In particular, it has been making big efforts and progress in digital image and video processing because of their utility in the information society that we are living in. Considering the huge demand existing in the area of security systems, one of the biggest research lines is video surveillance. The need for providing security to people and their properties in today's world explains the huge development and expansion of video surveillance systems nowadays. Within the digital image and video processing research area, there exists a rich variety of algorithms for motion detection, object detection, event detection, etc, which are being used in security [19, 24]. Automatic people detection in video sequences [12] is one group of them. It is actually a complex problem with multiple applications, not only in video surveillance, but also in different areas like intelligent systems (robotic), video games, etc.

The complexity of the people detection problem is mainly based on the difficulty of modeling persons because of their huge variability in physical appearances, poses, movements, points of views and interactions between different people and objects. Currently, many different systems exist which try to solve this problem. The state of the art in people detection and tracking includes several successful solutions working in specific and constrained scenarios. Most of them obtain good detection results but do not operate in real time. In contrast, the systems operating in real time usually get worse results. The work presented in this paper is inspired by a well-established non-real time solution in the field [28], on which we introduce some useful modifications to operate in real time and add robustness to the detection.

This paper is structured as follows: section 2 gives a brief introduction of the literature related to the work presented in this paper, section 3 overviews our complete system, whilst section 4 describes the proposed people detection algorithm. Section 5 describes the developed dataset, before showing experimental results in section 6. In section 7, the main conclusions are summarized and future work is described.

2. State of the art

The people detection task consists mostly of, firstly, the design and training of a person model based on characteristic parameters (motion, dimensions, silhouette, etc), and 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/classified as person, whilst all the others won't be detected/classified as person.

Most of the existing approaches are only based on appearance information [3, 7, 9, 13, 16, 17, 28, 30, 32] although some of them add robustness to the detection incorporating motion information through tracking algorithms. There are few approaches based only on motion information [6, 22] which main advantages are that they are independent of appearance variability and usually have low complexity. However, they usually have poorer results and they do not support partial occlusions.

The methods based on appearance can be classified according to the person model used. The methods based on simplified person models (only a region or shape) [7, 9, 16, 17, 30, 32] usually have low complexity but they do not support partial occlusions neither pose variations. However, the methods based on more complex person models [3, 13, 28], which usually have higher complexity than the previous ones, support partial occlusions and pose variations. Another advantage is that they made the final decision by combining multiple evidences, so they are usually more reliable than methods based on simpler human models.

Focusing on the idea of a real video surveillance system, people detection algorithms can be classified into two main families depending on whether they work in real time or not. This ramification splits the problem, and even the approach used in each case, in two systems clearly differentiated. On the one hand, systems that operate in real time usually get initial candidates location using image segmentation. Some approaches employ background subtraction [31, 32], whilst other approaches use stereo vision or 3D information [1, 11]. Besides, due to computational constraints, these approaches usually employ simplified person models (ellipse, human shape templates, etc). On the other side, the systems that do not operate in real time [3, 7, 17, 21, 27, 28, 29] get these initial candidates location scanning the complete image at various scales and rotations; in this case, person models must be complex to classify correctly many negative examples. The scanning and use of more complex models improve the detection rate but the computational costs are too high to allow for real time processing. Some approaches [33], try to speed these methods while maintaining a similar accuracy level, obtaining near real time human detection.

3. System overview

The processing stages of a ”canonical” automated video analysis system for people detection include: background/foreground extraction, object extraction, object classification, object tracking, and event or action recognition [14, 24]. As we can see in Figure 1 our system includes these four stages:

Background/Foreground extraction. Background/Foreground extraction is a commonly used technique for motion detection and segmentation. Motion detection aims at segmenting regions corresponding to moving objects from the rest of the image. The consecutive stages depend on the background accuracy obtained, that is, the rest of stages have a strong dependency with the results obtained in this process: a bad background model could cause false object detections, missing objects or partial object detections. In our system, foreground extraction is based on [5].

Object extraction. After background/foreground segmentation, morphological operations are typically applied to reduce the noise of the resulting image mask and improve the object extraction [24]. In our system, after object extraction, a connected component analysis is applied [8]. Only objects extracted in this stage are analyzed in following stages. Each object is defined with a blob (localization and dimensions).

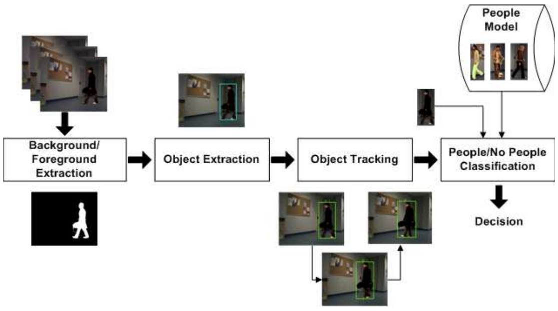

People Model
Background/ Foreground Extraction
Object Extraction
Object Tracking
People/No People Classification
Decision

Figure 1. Overall System Architecture

Object tracking. After motion detection and object extraction, surveillance systems generally track moving objects. The aim of an object tracker is to generate the trajectory of an object over time by locating its position in every frame of the video sequence where it appears. In our system, a simple tracking algorithm based on the Kalman filter [4] is used and generates the trajectories of the blobs between consecutive frames using color information, blob’s dimensions (width and height) and blob’s position (centroid). The color information is the Hue-channel color histogram calculated on all points inside of the object‘s mask.

Object classification. Object classification can be considered as a standard pattern recognition issue. This process compares previously trained object models and generated object models from an image or sequence and makes a final decision based on their similarity. The details of this module in our system are described with more detail in section 4.

4. Object classification

Our people detector is based on the algorithm proposed in [28] but proposing modifications in order to achieve real time performance in video surveillance scenarios.

4.1. Base algorithm

[28] proposes a method for human detection in crowded scenes, but working only with static images (frames). An individual human is modeled as an assembly of natural body parts. The main idea consists of identifying characteristic edges of each body part and generating four edge models of body parts (body, head, torso and legs). The image is scanned with four independent edge feature detectors previously trained. The training phase is performed using the Real Adaboost algorithm [10] and a nested cascade structure [15]. Responses of each part detectors are combined to obtain a joint likelihood model that includes cases of multiple, and possibly inter-occluded, humans. This algorithm

also supports changes in pose or camera point of view.

4.2. Proposed algorithm

The base algorithm is targeted to static images and scans the complete image; for these reasons person models must be complex in order to be able to classify correctly many negative examples. In addition, as computation time is not a main objective, the training phase is focused on reducing false positive rate (complex person models) what greatly increases the processing time. In order to get a faster algorithm, we propose not to scan the complete image and to simplify the person model. Firstly, instead of scanning the complete image, we only process moving objects detected in previous stages (see Figure 1). Secondly, the model of each body part is simplified and, consequently, the final person model what reduces the time needed during the detection process. The proposed simplifications are the following: we use a ranking of the best edges of each body part and the training phase is not focused on reducing false positive but also on getting good precision results.

4.2.1 Edge shapes

In this work, according to the size of the images (58x24 pixels) and the base method [28], the possible length (k) of one single edge is from 4 pixels to 12 pixels. The edge features we use consist of single shapes, including lines, 1/8 circles, 1/4 circles and 1/2 circles. We use 36 types of lines (four orientations: 0º, 45º, 90º and 135º; and 9 dimensions: 4-12 pixels; $4\, orientations \times 9\, dimensions = 36$). We generate arcs from 4 pixels to 12 pixels such that the perimeter of their circumference (P) follows 1. Finally, we have a total of 775 edges (36 lines and 739 arcs). For example, when the size of the body image is 58×24, the overall number of possible edge features is 1078800 ($58 \times 24 \times 775 = 1078800$).

$$\left(\frac{1}{8}, \frac{1}{4}\, or\, \frac{1}{2}\right) \times P \geqslant k,\ k = 4, \ldots, 12,\ P \in \mathbb{N} \qquad (1)$$

4.2.2 Learning part detectors

For each edge feature, one weak classifier [28] is built. Then the AdaBoost algorithm [10] is used to learn strong classifiers. The AdaBoost algorithm has many variations such as Discrete Adaboost, Real Adaboost and Gentle Adaboost. Instead of using the Real AdaBoost variation, Gentle Adaboost is chosen because it outperforms other variations as reported by [18]. In order to reduce computational cost and to identify the most characteristic edges of each body part, we make a top-100 edge ranking. We iteratively train in a bootstrap way the best classifier of each edge and select the best 100 associated edges. Finally, instead of using a complex nested structure focused on reducing the false positive rate, the cascade Adaboost algorithm [26] (Gentle variation) is used to learn each detector. This training phase is not only focused on reducing the false positive rate but also on getting good precision results.

4.2.3 People detection

Only objects detected after the three previous stages (see Figure 1) are classified. Each blob's image is normalized and then the four models of body parts (cascade classifiers) are generated.

The classification process consists of evaluating the four models of body parts, providing four independent evidences. The final evidence about the analyzed blob being a person is obtained by averaging the evidences provided by the four body parts detectors.

The final people detection is less complex (by using a simplified person model and a smaller number of classifiers: those which belong to the top-100 edge ranking) and the completed system is faster (by not scanning the entire image), whilst maintaining good precision results.

5. Dataset

5.1. Image dataset

The proposed algorithm consists of four models of edge body parts. Each model has to be trained with an image collection with people and non-people examples and therefore we need a complete image dataset with positive and negative examples.

Negative images have been chosen from the LabelMe dataset [20]. Each image has been cropped in small pieces in order to obtain a huge number of different negative images. Positive images have been chosen from the INRIA dataset [7]. Person body blobs have been extracted, normalized (58x24 pixels and gray scaled) and segmented in body parts (see Figure 2) according to the base algorithm [28]. Finally our image dataset stores 3.542 positive images (already extracted, normalized and segmented) and more than 40.000 negative images.

5.2. Video dataset

In order to evaluate the performance of the proposed approach, we introduce a video dataset containing 20 surveillance annotated sequences of varying difficulty. We have grouped all the test sequences into different complexity categories depending on two aspects:

- **People classification complexity**, defined as the difficulty to classify moving and temporally stationary people in a scenario. It is related with the number of pedestrians, their velocity, partial occlusions and pose variations.

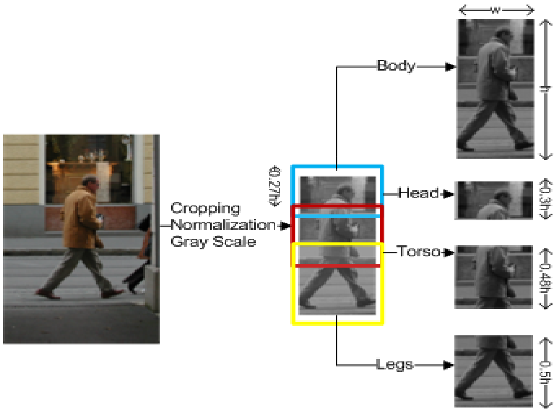


Figure 2. Body Part Segmentation

- **Background complexity**, defined as the difficulty to extract the foreground due to the presence of edges, multiple textures, lighting changes, reflections, shadows and objects belonging to the background.

A description of complexity levels of the associated content is shown in Table 1, whilst Figure 3 shows two examples of each category. The videos have been collected from several public datasets related with the people detection/object classification task [23, 25] and PETS 2006 dataset (available at http://pets2006.net/). The image dataset, the dataset of selected videos and their annotations are freely available for academic purposes (**http://www-vpu.ii.uam.es/PDds/**).

	Complexity	
Category	**Classification**	**Background**
C1	Low	Low
C2	Medium	Low
C3	Medium	Medium
C4	High	Low
C5	High	High

Table 1. Sequence categorization

6. Experimental results

In this section, we describe the experiments carried out for testing the proposed people system over our video dataset and we compare the results of our approach *Edge*, with three other people detectors approaches from the state of the art: two non-real time approaches, *HOG* and *TUD* detectors [3, 7], and one real time approach, *Fusion* [9]. Our approach *Edge*, is based on [28], the authors themselves show in [29] similar results than *HOG* in terms of classification accuracy. On the other hand *TUD* detector outperforms *HOG* detector, previous authors' detector partISM [2] and ISM variations (4D-ISM [21] and standard ISM [17]), in terms of classification accuracy.

Experimental results include an evaluation of people detection rates and computational cost. The system has been implemented in C++, using the OpenCV image processing library (http://sourceforge.net/projects/opencv/). The tests have been performed on a Pentium IV with a CPU frequency of 2.4 GHz and 3GB RAM.

6.1. People detection results

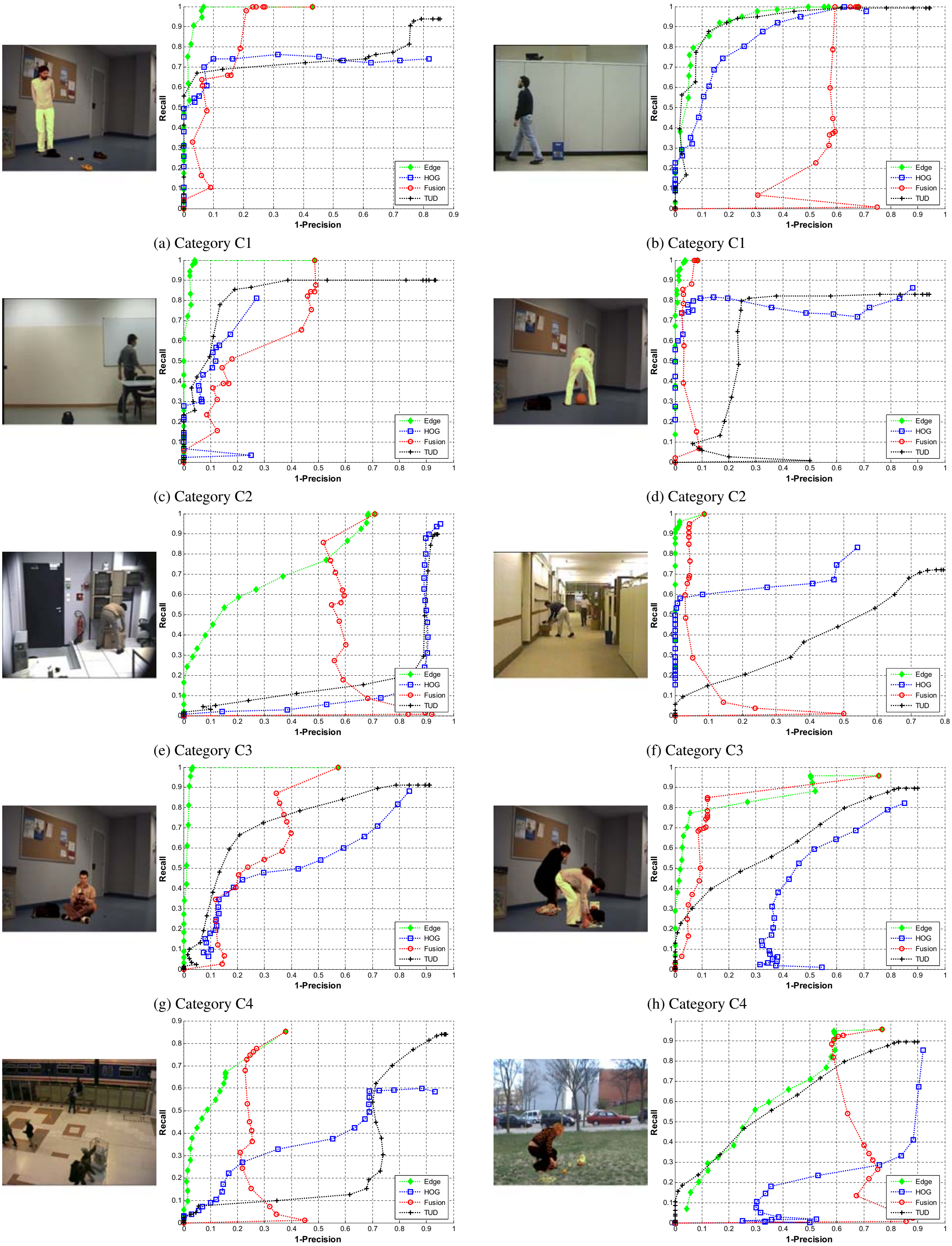
Despite the fact that all algorithms performance depends on the hit rate, or confidence level of decision, we only classify objects detected in previous stages (see Figure 1) as person or non-person. Consequently, the maximum/minimum recall and precision will be limited by previous stages. One of the previously mentioned approaches, *Fusion*, is based on the same scheme and also performs in real time. Moreover, the non-real time approaches, *HOG* and *TUD*, are limited by the image scanning.

Figure 3 shows the detection performance on some examples of different complexity categories included within the used video dataset. At low levels of classification and background complexity C1 -Fig.3 (a) and (b)-, our method, *Edge*, outperforms or obtains the same results than the other approaches.

At intermediate complexity categories C2, C3 and C4, our proposal is clearly superior to non-real time systems. In particular, we can see how in sequences with many partial occlusions and pose variations -Fig.3 (e), (f), (g) and (h)-, the non-real time systems performance is significantly reduced. While in our system even if some individual parts detectors may have poor results (partial occlusion and pose variations), the combined detector maintains high detection rates.

At high levels of classification and background complexity C5, the global performance of our system is reduced, mainly, due to the high background complexity. The first example -Fig.3 (i)- shows a scene with lighting changes, reflections and shadows; while the second example -Fig.3 (j)- shows a scene with lighting changes, shadows and a multi-modal background (moving branches). However, the results are still better, or slightly better, than the best non-real time performance (*TUD*).

Non-real time approaches are robust due to the full person search carried out and their complex person models. However, in some cases they show an unreliable performance because of the high number of false positive examples that appear during exhaustive search. The previously explained problem affects to both algorithms at the same time in Fig.3 (e) and (i), and, individually, for *TUD* in Fig.3 (f) and for *HOG* in Fig.3 (j). The real time approach, *Fusion*, also shows an unreliable performance. Its usage of a highly simplified person model achieves fast people detection but not quite robust, as a result, this approach presents an irregular behavior in all categories.



(a) Category C1

(b) Category C1

(c) Category C2

(d) Category C2

(e) Category C3

(f) Category C3

(g) Category C4

(h) Category C4

(i) Category C5

(j) Category C5

Figure 3. Examples of the sequences categories and people detection results

In our system, even though we use a simplified person model in order to work in real time, our performance is, in general, equal or superior to other approaches in all categories.

6.2. Computational cost

In this subsection, the computational cost, measured as the output processing rate in frames per second (fps), generated by our approach will be compared with two different approaches. In first place, the previously mentioned real time approach, and, secondly, the non-real time approach called *HOG*. *TUD* algorithm was also previously mentioned, however, due to its high computational cost it will not be considered in the comparison[1]. The above described dataset includes different video resolutions; the results obtained with 352x288 images are summarized in Table 2.

The computational cost of real time approaches depend a lot on each sequence. It does not only depend on people detection and background complexity, but also on many other factors: object's dimensions, number of tracked objects, etc. For this reason, we show a summary with the worst, best and average results obtained over our proposed dataset (20 sequences).

The results show clearly how our proposed detector, *Edge*, works in real time, and even faster than the previously mentioned real time approach, *Fusion*. Both real time approaches computational costs depend on the different sequences. Nevertheless, the non-real time approach remains almost invariant to different sequences because of the exhaustive search carried out.

Average fps	Edge	Fusion	HOG	TUD
Minimum	64.5	14.5	11.4	NC[1]
Average	71.6	32.6	11.5	NC[1]
Maximum	80.8	62.8	11.6	NC[1]

Table 2. Computational cost

7. Conclusions

In this paper, an improved approach for real time and robust people detection is presented. A complete surveillance video system has been implemented to evaluate the proposed detection approach. Besides, in order to provide a good performance evaluation of the proposed framework, the VPULab Person Detection dataset composed of several annotated surveillance sequences of different levels of complexity has been developed.

Experimental results over the proposed dataset show that the proposed system performs considerably well at real time and even better than other non-real time approaches from the state of the art and that it is significantly more efficient and stable than others approaches from the state of the art.

As our framework is composed by different functional modules, there are several proposals for improving it with future work. We propose the study of techniques for multimodal background modeling, noise removal, shadows detection, etc, in order to refine the background extraction. We also propose to study tracking algorithms which are more robust to occlusions and allow to track multiple targets. As currently we do not make use of motion information, we will study how to add robustness to the detection by incorporating some type of motion information.

8. Acknowledgments

This work has partially supported by the Cátedra UAM-Infoglobal ("Nuevas tecnologías de vídeo aplicadas a sistemas de video-seguridad") and by the Spanish Government (TEC2007-65400 SemanticVideo).

References

[1] 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 Trans. on Intelligent Transportation Systems*, 8(2):292–307, 2007. 2

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. of CVPR*, pages 1–8, 2008. 4

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. of CVPR*, pages 1014–1021, 2009. 1, 2, 4

[4] T. J. Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(2):90–99, 1986. 2

[5] A. Cavallaro and T. Ebrahimi. Video object extraction based on adaptive background and statistical change detection. In *Proc. of SPIE*, pages 465–475, 2001. 2

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

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, volume 1, pages 886–893, 2005. 1, 2, 3, 4

[8] M. B. Dillencourt, H. Samet, and M. Tamminen. A general approach to connected-component labeling for arbitrary image representations. *Journal of the ACM*, 39(2):253–280, 1992. 2

[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*, pages 55–58, 2008. 1, 2, 4

[10] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computer and System Sciences*, 55(1):119–139, 1997. 2, 3

[1] NC: Not considered in the comparison due to its high computational cost.

[11] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007. 2

[12] D. M. Gavrila and S. Munder. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, 2009. 1

[13] I. Haritaoglu, D. Harwood, and L. S. Davis. Ghost: a human body part labeling system using silhouettes. In *Proc. of ICPR*, volume 1, pages 77–82, 1998. 1, 2

[14] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(3):334–352, 2004. 2

[15] C. Huang, H. Al, B. Wu, and S. Lao. Boosting nested cascade detector for multi-view face detection. In *Proc. of ICPR*, volume 2, pages 415–418, 2004. 2

[16] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos. Estimating pedestrian counts in groups. *Computer Vision and Image Understarding*, 110(1):43–59, 2008. 1, 2

[17] B. Leibe and B. Schiele. Scale invariant object categorization using a scale-adaptive mean-shift search. In *Proc. of DAGM*, volume 3175, pages 145–153, 2004. 1, 2, 4

[18] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Lecture Notes in Computer Science*, pages 297–304, 2003. 3

[19] K. N. Platanioitis and C. S. Regazzoni. Visual-centric surveillance networks and services. *IEEE Signal Processing Magazine*, 22(2):12–15, 2005. 1

[20] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1):157–173, 2008. 3

[21] E. Seemann and B. Schiele. Cross-articulation learning for robust detection of pedestrians. In *Proc. of DAGM*, 2006. 2, 4

[22] H. Sidenbladh. Detecting human motion with support vector machines. In *Proc. of ICPR*, pages 188–191, 2004. 1

[23] F. Tiburzi, M. Escudero, J. Bescos, and J. M. Martinez. A ground truth for motion-based video-object segmentation. In *Proc. of ICIP*, pages 17–20, 2008. 4

[24] M. Valera and S. A. Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings on Visual Image Signal Processing*, 152(2):192–204, 2005. 1, 2

[25] R. Vezzani and R. Cucchiara. Annotation collection and online performance evaluation for video surveillance: The visor project. In *Proc. of AVSS*, pages 227–234, 2008. 4

[26] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR*, volume 1, pages 511–518, 2001. 3

[27] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. of ICCV*, volume 2, pages 734–741, 2003. 2

[28] 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*, pages 90–97, 2005. 1, 2, 3, 4

[29] 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*, 75(2):247–266, 2007. 2, 4

[30] F. Xu and K. Fujimura. Human detection using depth and gray images. In *Proc. of AVSS*, pages 115–121, 2003. 1, 2

[31] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, 2004. 2

[32] J. Zhou and J. Hoang. Real time robust human detection and tracking system. In *Proc. of CVPR*, page 149, 2005. 1, 2

[33] 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. 2