

REVIEW

Open Access

Human detection in surveillance videos and its applications - a review

Manoranjan Paul*, Shah M E Haque and Subrata Chakraborty

Abstract: Detecting human beings accurately in a visual surveillance system is crucial for diverse application areas including abnormal event detection, human gait characterization, congestion analysis, person identification, gender classification and fall detection for elderly people. The first step of the detection process is to detect an object which is in motion. Object detection could be performed using background subtraction, optical flow and spatio-temporal filtering techniques. Once detected, a moving object could be classified as a human being using shape-based, texture-based or motion-based features. A comprehensive review with comparisons on available techniques for detecting human beings in surveillance videos is presented in this paper. The characteristics of few benchmark datasets as well as the future research directions on human detection have also been discussed.

1. Review

1.1 Introduction

Over the recent years, detecting human beings in a video scene of a surveillance system is attracting more attention due to its wide range of applications in abnormal event detection, human gait characterization, person counting in a dense crowd, person identification, gender classification, fall detection for elderly people, etc.

The scenes obtained from a surveillance video are usually with low resolution. Most of the scenes captured by a static camera are with minimal change of background. Objects in the outdoor surveillance are often detected in far field. Most existing digital video surveillance systems rely on human observers for detecting specific activities in a real-time video scene. However, there are limitations in the human capability to monitor simultaneous events in surveillance displays [1]. Hence, human motion analysis in automated video surveillance has become one of the most active and attractive research topics in the area of computer vision and pattern recognition.

An intelligent system detects and captures motion information of moving targets for accurate object classification. The classified object is being tracked for high-level analysis. In this study, we focus on detecting humans and do not consider recognition of their complex activities. Human detection is a difficult task from a machine vision perspective as it is influenced by a wide

range of possible appearance due to changing articulated pose, clothing, lighting and background, but prior knowledge on these limitations can improve the detection performance.

The detection process generally occurs in two steps: object detection and object classification. Object detection could be performed by background subtraction, optical flow and spatio-temporal filtering. Background subtraction is a popular method for object detection where it attempts to detect moving objects from the difference between the current frame and a background frame in a pixel-by-pixel or block-by-block fashion. There are few available approaches to perform background subtraction. The most common ones are adaptive Gaussian mixture [2-10], non-parametric background [11-17], temporal differencing [18-20], warping background [21] and hierarchical background [22] models. The optical flow-based object detection technique [18,23-26] uses characteristics of flow vectors of moving objects over time to detect moving regions in an image sequence. Apart from their vulnerability to image noise, colour and non-uniform lighting, most of the flow computation methods have large computational requirements and are sensitive to motion discontinuities. For motion detection based on the spatio-temporal filter methods, the motion is characterized via the entire three-dimensional (3D) spatio-temporal data volume spanned by the moving person in the image sequence [27-37]. Their advantages include low computational complexity and a simple implementation

* Correspondence: mpaul@csu.edu.au
Centre for Research in Complex Systems (CRiCS), School of Computing and Mathematics, Charles Sturt University, Bathurst, Australia

process. However, they are susceptible to noise and variations of the timings of movements.

The object classification methods could be divided into three categories: shape-based, motion-based and texture-based. Shape-based approaches first describe the shape information of moving regions such as points, boxes and blobs. Then, it is commonly considered as a standard template-matching issue [18,23,38-43]. However, the articulation of the human body and the differences in observed viewpoints lead to a large number of possible appearances of the body, making it difficult to accurately distinguish a moving human from other moving objects using the shape-based approach. This challenge could be overcome by applying part-based template matching [39]. Texture-based methods such as *histograms of oriented gradient* (HOG) [44] use high dimensional features based on edges and use *support vector machine* (SVM) to detect human regions.

A large number of studies described in this review use publicly available datasets that are specifically recorded for training and evaluation. *KTH human motion* dataset [45] contains six activities, whereas *Weizmann human action* dataset [46] and *INRIA XMAS multi-view* dataset [47] contains 10 and 11 actions, respectively. *Performance Evaluation of Tracking and Surveillance* (PETS) datasets [48-59] have a number of datasets for different purposes of vision-based research. Each year, PETS run an evaluation framework on specific datasets with specific objective. The Institute of Automation, Chinese Academy of Sciences (CASIA) provides the CASIA Gait Database [60] for gait recognition and related research.

The key purpose of this paper is to provide a comprehensive review on studies conducted in the area of human detection process of a visual surveillance system. A flow chart of the human detection process is illustrated in Figure 1. Various available techniques are reviewed in Section 1.2. Details of several benchmark databases are presented in Section 1.3. Several major applications are reviewed in Section 1.4. We present a review and analyses of recent developments and highlight future directions of research in the area of human detection in visual surveillance. Future directions are discussed in Section 1.5. The main contributions of this paper are as follows:

- Object detection and object classification are discussed in a clearly organized manner according to the general framework of visual surveillance. This, we believe, can help readers, especially newcomers to this area, to obtain an understanding of the state of the art in visual surveillance and the scope of its application in the real world.
- The pros and cons of a variety of different algorithms for motion detection and classification are discussed.

- We provide a discussion on future research directions in human detection in visual surveillance.

1.2 Techniques

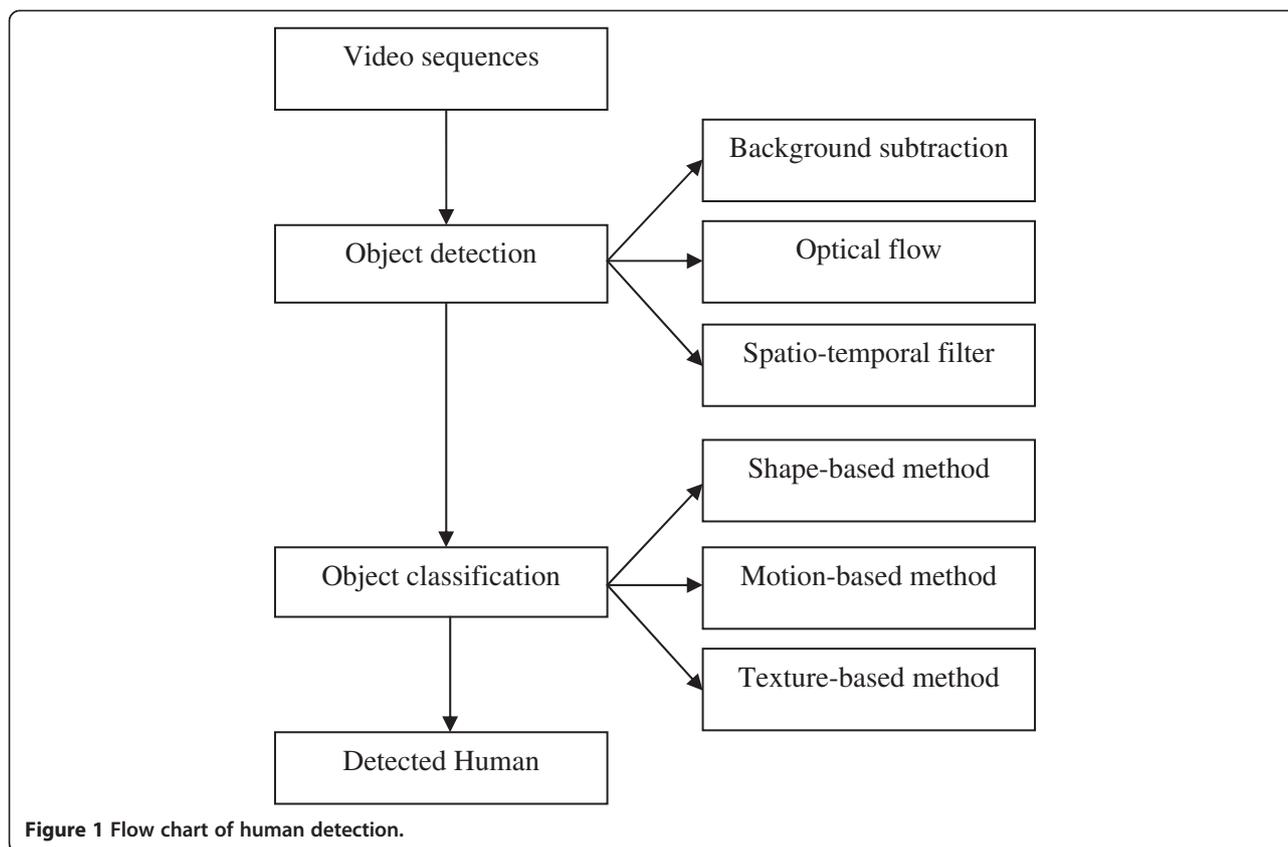
Human detection in a smart surveillance system aims at making distinctions among moving objects in a video sequence. The successful interpretations of higher level human motions greatly rely on the precision of human detection [61-63]. The detection process occurs in two steps: object detection and object classification.

1.2.1 Object detection

An object is generally detected by segmenting motion in a video image. Most conventional approaches for object detection are background subtraction, optical flow and spatio-temporal filtering method. They are outlined in the following subsections.

1.2.1.1 Background subtraction Background subtraction is a popular method to detect an object as a foreground by segmenting it from a scene of a surveillance camera. The camera could be fixed, pure translational or mobile in nature [63]. Background subtraction attempts to detect moving objects from the difference between the current frame and the reference frame in a pixel-by-pixel or block-by-block fashion. The reference frame is commonly known as 'background image', 'background model' or 'environment model'. A good background model needs to be adaptive to the changes in dynamic scenes. Updating the background information in regular intervals could do this [64], but this could also be done without updating background information [65]. Few available approaches have been discussed in this section:

- *Mixture of Gaussian model*. Stauffer and Grimson [2] introduced an *adaptive Gaussian mixture* model, which is sensitive to the changes in dynamic scenes derived from illumination changes, extraneous events, etc. Rather than modelling the values of all the pixels of an image as one particular type of distribution, they modelled the values of each pixel as a mixture of Gaussians. Over time, new pixel values update the mixture of Gaussian (MoG) using an online *K-means* approximation. In the literature, many approaches are proposed to improve the MoG [3-11]. In [4], an effective learning algorithm for MoG is proposed to overcome the requirement of the prior knowledge about the foreground and background ratio. In [5], authors presented an algorithm to control the number of Gaussians adaptively in order to improve the computational time without sacrificing the background modelling quality. In [6], each pixel is modelled by support vector regression. *Kalman filter* is used for adaptive



background estimation in [7]. In [8], a framework for *hidden Markov Model* (HMM) topology and parameter estimation is proposed. In [9], colour and edge information are fused to detect foreground regions. In [10], normalized coefficients of five kinds of orthogonal transform (*discrete cosine transformation*, *discrete Fourier transformation* (DFT), Haar transform, *single value decomposition* and Hadamard transform) are utilized to detect moving regions. In [11], each pixel is modelled as a group of adaptive local binary pattern histograms that are calculated over a circular region around the pixel.

- *Non-parametric background model.* Sometimes, optimization of parameters for a specific environment is a difficult task. Thus, a number of researchers introduced non-parametric background modelling techniques [12-17]. Non-parametric background models consider the statistical behaviour of image features to segment the foreground from the background. In [13], a non-parametric model is proposed for background modelling, where a kernel-based function is employed to represent the colour distribution of each background pixel. The kernel-based distribution is a generalization of MoG [4], which does not require parameter estimation. The computational requirement is high for this method.

Kim and Kim [12] proposed a non-parametric method, which was found effective for background subtraction in dynamic texture scenes (e.g. waving leaves, spouting fountain and rippling water). They proposed a clustering-based feature, called *fuzzy colour histogram* (FCH) to construct the background model by computing the similarity between local FCH features with an online update procedure. Although the processing time was high in comparison with the adaptive Gaussian mixture model [2], the false positive rate of detection is significantly low at high true positive rates.

- *Temporal differencing.* The temporal differencing approach [19] involves three important modules: block alarm module, background modelling module and object extraction module (see Figure 2). The block alarm module efficiently checked each block for the presence of either a moving object or background information. This was accomplished using temporal differencing pixels of the Laplacian distribution model and allowed the subsequent background modelling module to process only those blocks that were found to contain background pixels. Next, the background modelling module is employed in order to generate a high-quality adaptive background model using a unique two-stage training

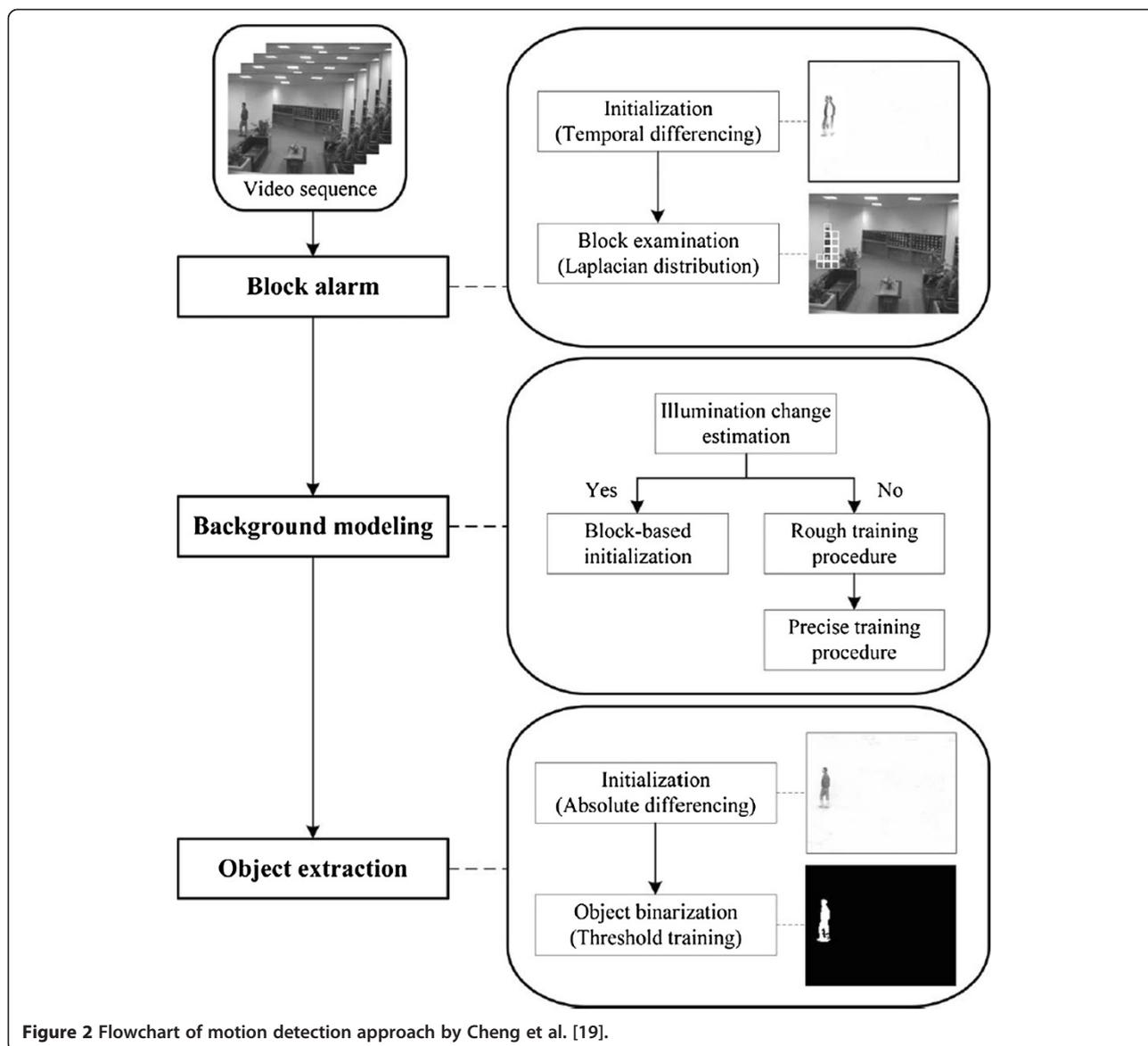


Figure 2 Flowchart of motion detection approach by Cheng et al. [19].

procedure and a mechanism for recognizing changes in illumination. As the final step of their process, the proposed object extraction module computes the binary object detection mask by applying suitable threshold values. This is accomplished using their proposed threshold training procedure.

The performance evaluation of their proposed method is accomplished by quantitative and qualitative processes. The overall results showed that their proposed method attained a substantially higher degree of efficacy.

- *Warping background.* Ko et al. [21] presented a background model that differentiates between background motion and foreground objects. Unlike most models that represent the variability of pixel

intensity at a particular location in the image, they modelled the underlying warping of pixel locations arising from background motion. The background is modelled as a set of warping layers where at any given time, different layers may be visible due to the motion of an occluding layer. Foreground regions are thus defined as those that cannot be modelled by some composition of some warping of these background layers.

- *Hierarchical background model.* Chen et al. [22] proposed a hierarchical background model, which is based on region segmentation and pixel descriptors to detect and track foreground. It first segments the background images into several regions by the *mean-shift* algorithm. Then, a hierarchical model, which consists of the region models and pixel

models, is created. The region model is one kind of approximate Gaussian mixture model extracted from the histogram of a specific region. The pixel model is based on the co-occurrence of image variations described by HOG of pixels in each region. Benefiting from the background segmentation, the region models and pixel models corresponding to different regions can be set to different parameters. The pixel descriptors are calculated only from neighbouring pixels belonging to the same object. The hierarchical models first detect the regions containing foreground and then locate the foreground only in these regions, thus avoid detection failure in other regions and reduce the time and cost. A similar two-stage hierarchical method has been introduced earlier by Chen [66] where the block-based stage provides a coarse foreground segmentation followed by the pixel-based stage for finer segmentation. The method showed promising results when compared with MoG. Recent application of this approach can be seen in the study of Quan [67] where the hierarchical background model (HBM) is combined with the codebook [68] technique.

1.2.1.2 Optical flow Optical flow is a vector-based approach [18,23,26] that estimates motion in video by matching points on objects over image frame(s). Under the assumption of brightness constancy and spatial smoothness, optical flow is used to describe coherent motion of points or features between image frames. Optical flow-based motion segmentation uses characteristics of flow vectors of moving objects over time to detect moving regions in an image sequence. One key benefit of using optical flow is that it is robust to multiple and simultaneous cameras and object motions, making it ideal for crowd analysis and conditions that contain dense motion. Optical flow-based methods can be used to detect independently moving objects even in the presence of camera motion. Apart from their vulnerability to image noise, colour and non-uniform lighting, most of flow computation methods have large computational requirements and are sensitive to motion discontinuities. A real-time implementation of optical flow will often require a specialized hardware due to the complexity of the algorithm and moderately high frame rate for accurate measurements [18].

1.2.1.3 Spatio-temporal filter For motion recognition based on spatio-temporal analysis, the action or motion is characterized via the entire 3D spatio-temporal data volume spanned by the moving person in the image sequence. These methods generally consider motion as a whole to characterize its spatio-temporal distributions [27,37]. Zhong et al. [27] processed a video sequence

using a spatial Gaussian and a derivative of Gaussian on the temporal axis. Due to the derivative operation on the temporal axis, the filter shows high responses at regions of motion. These responses were then used to generate thresholds to yield a binary motion mask, followed by aggregation into spatial histogram bins. Such a feature encodes motion and its corresponding spatial information compactly and is useful for far-field and medium-field surveillance videos. As these approaches are based on simple convolution operations, they are fast and easy to implement. They are quite useful in scenarios with low-resolution or poor-quality video where it is difficult to extract other features such as optical flow or silhouettes. Spatio-temporal motion-based methods are able to better capture both spatial and temporal information of gait motion. Their advantage is low computational complexity and a simple implementation. However, they are susceptible to noise and to variations of the timings of movements.

1.2.1.4 Performance comparisons of detection techniques

A generic comparison among object detection methods in terms of accuracy and computational time is presented in Table 1. The table shows accuracy and computational time of different object detection techniques in terms of three criteria, namely low, moderate and high. It is very difficult to generalize the accuracy and computational time of different techniques in each category by three simple attributes because there are several techniques in each category, and each technique has its own accuracy and computational time. We have provided the general trends of these techniques in each category based on various available comparative studies. The readers will have a general understanding about their performances using this table. This should act as a guide for the readers and practitioners to conduct further investigation to find the appropriate technique suitable for their specific contexts.

The MoG-based models compute at pixel level (or small block level) and provide moderate accuracy and relatively low computational time [2]. It has been applied widely, and several improved models are introduced based on MoG. The MoG models are widely used as base model for performance comparisons of new models. The general non-parametric techniques provide high accuracy in dynamic background scenarios but require lower computational time [13]. Temporal differencing technique attained between 10% and 25% more accuracy than some well-known techniques including MoG and has excellent capabilities to handle sudden illumination issues [19]. Warping background techniques provide significantly better results (between 10% and 40% for various datasets) for separating background motion from foreground motion using neighbouring pixel information compared

Table 1 Comparison of object detection methods in terms of accuracy and computational time

Methods	Accuracy	Computational time	Comments	
Background subtraction	Mixture of Gaussian model [2-10]	Moderate	Moderate	Simple implementation and good performance but not so well with dynamic background. It requires parameters to be defined by the practitioners. It can capture multi-modal scenarios
	Non-parametric background model [12-17]	Moderate to high	Low to moderate	In dynamic background scenarios, NP performs very well compared to MoG-based algorithm. It requires significant post-processing. In occlusion situation, it does not perform well compare to MoG
	Temporal differencing [19,20,47,69]	High	Low to moderate	Very good with sudden illumination changes in indoor environment
	Warping background [21]	High	Moderate to high	Good in outdoor environment with high background motion. It does not handle occlusion well. Some variations are computationally intensive
	Hierarchical background model [22,66-68]	High	Low to moderate	Make use of both block-based and pixel-based approaches. May be quicker than pixel-based approach, but quality could be compromised
Optical flow [18,23-26]		Moderate	High	Good with camera motion and crowd detection but highly computation intensive
Spatio-Temporal filter [27-37,70]		Moderate to high	Low to moderate	Works well for low-resolution scenarios but suffers from noise issues

to few classic methods including the non-parametric technique, and the implicit version claims to require less computational overhead [21]. The HBM method provides high accuracy (about 5% to 15% less error) compared to some classic methods including MoG and requires slightly less computational time compared to MoG-based methods as it uses hybrid techniques [22].

Optical flow methods have distinct advantages in moving object detection compared to background subtraction methods as they can handle camera motion and perform well in crowd detection; however, they require higher computational time and special hardware for real-time applications [18,23]. A comprehensive comparative study among several classic optical flow techniques can provide in-depth understanding to interested readers [24].

Spatio-temporal-based methods are better in accuracy where noise is less as they consider motion in a holistic way. These methods showed promising results in unusual event detection scenarios, and they are good in terms of computational time [27-37]. Recently, a new texture descriptor and hysteresis thresholding-based object detection technique has been introduced by Lai et al. [70] which shows better performance than traditional MoG in challenging conditions such as illumination, shadow- and motion-induced problems.

A modified MoG-based approach by replacing the mean pixel intensity value with the recent pixel intensity value in background frame generation performs better to detect object in a general situation [71] compared to other approaches. A number of video-coding techniques also used the MoG-based approach to generate a background frame and use an additional reference frame to encode uncovered/occluded regions of a frame for better coding efficiency [72-75]. Due to computational time, implementation issues, accuracy and memory requirement,

it is very difficult to incorporate other approaches into video-coding applications to encode uncovered/occluded regions.

1.2.1.5 A comparison study In order to demonstrate the comparison technique, we have conducted a comparison study using a readily available software tool *MFC BGS Library x86 1.3.0* [76]. The tool provides a wide array of background subtraction methods. In this comparison test, we have chosen the MoG [2], the NP-KDE [13], the temporal median [77] and the frame difference [78] methods. We have chosen these four methods due to their class leading reputations and applications by a large number of researchers.

For this study, we have used the Wallflower dataset [78,79]. A total of 248 frames were provided as input to the MFC BGS Library x86 1.3.0 tool which provided the detected foreground frames for each input frame. A hand-drawn ground truth has been provided for frame 248 with the Wallflower data. We have compared the foreground for frame 248 with the ground truth. The results are shown in Table 2. From the output detection pictures and the numeric results for *false positive* (FP) and *false negative* (FN), we can observe that the non-parametric (NP) one has been most successful in detecting the moving tree in the background from the foreground. The temporal median method has been most successful in identifying the foreground regions but was not as good as the NP in detecting the moving tree as background.

Although this is a simple and short study, it provides a general guidance to the readers regarding the process of such comparative studies. Software tools such as *MFC BGS Library x86 1.3.0* or self-implemented tools can be used for such comparative studies. Although we have chosen only four methods, they are the initial ones in their respective category. We would like to highlight

Table 2 Comparative experimental results

	Original frame	Ground truth		
				
	MoG	NP-KDE	Temporal median	Frame difference
Output				
FP	21.5%	2.3%	6.9%	21.5%
FN	12.9%	11.5%	8%	13.4%

the fact that a significant number of new methods have been proposed by researchers with modification to these methods, most of which require some post-processing work such as noise reduction. A comprehensive comparison with all the methods is time consuming and may not be very useful as all the methods may not be suitable for a particular application. Researchers and practitioners are thus recommended to research on comparative studies such as [78,80-84] to identify potential methods suitable for their intended applications. A comparative study can then be conducted to find the most suitable one among the potential methods.

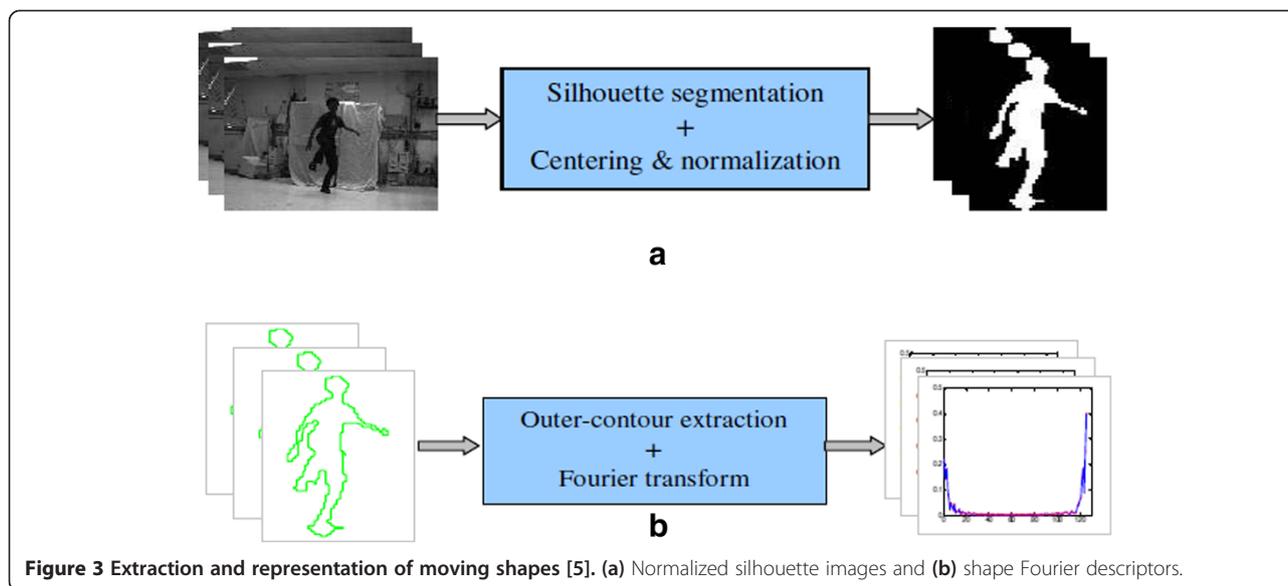
Object classification. An object in motion needs to be classified accurately for its recognition as a human being. The available classification methods could be divided into three main categories: shape-based method, motion-based method and texture-based method.

1.2.1.6 Shape-based method Shape-based approaches first describe the shape information of moving regions such as points, boxes and blobs. Then, it is commonly considered as a standard pattern recognition issue [18,23,38,43]. However, the articulation of the human body and differences in observed viewpoints lead to a large number of possible appearances of the body, making it difficult to accurately distinguish a moving human from other moving objects using the shape-based approach. Eishita et al. [43] proposed a simple but effective method for object tracking after full or partial occlusion using shape, colour and texture information even if the colour and textures are the same for the objects. Wang et al. [38] investigated how the deformations of human silhouettes (or shapes) during articulated motion could be used as discriminating features to implicitly capture motion

dynamics and exploited the applicability of *discrete wavelet transform* and DFT for the purpose of human motion characterization and recognition (see Figure 3).

Huang et al. [85] presented a performance evaluation of shape similarity metrics for 3D video sequences of people with unknown temporal correspondence. Lin and Davis [40] proposed a shape-based, hierarchical part-template-matching approach to simultaneous human detection and segmentation combining local part-based and global shape-template-based schemes. Their approach relied on the key idea of matching a part-template tree to images hierarchically to detect humans and estimate their poses. One major disadvantage of the shape-based method is that it cannot capture the internal motion of the object within the silhouette region. Even state-of-the-art background subtraction techniques do not always reliably recover precise silhouettes, especially in dynamic environments. This reduces the robustness of techniques in this method.

1.2.1.7 Motion-based method This classification method is based on the idea that object motion characteristics and patterns are unique enough to distinguish between objects. Motion-based approaches directly make use of the periodic property of the captured images to recognize human beings from other moving objects. Bobick and Davis [86] developed a view-based approach for the recognition of human movements by constructing a vector image template comprising two temporal projection operators: binary *motion-energy image* and *motion-history image*. Cutler et al. [87] presented a self-similarity-based time-frequency technology to detect and analyze periodic motion for human classification. Unfortunately, methods based on periodicity are restricted to periodic motion.



Efros et al. [26] characterized the human motion within a spatio-temporal volume by a descriptor, which was based on computing the optical flow, projecting the motion onto a number of motion channels and blurring with a Gaussian. Recognition was performed in a nearest-neighbour framework. By computing a spatio-temporal cross correlation with a stored database of previously labelled action fragments, the most similar to the motion descriptor of the query action fragment could be found.

1.2.1.8 Texture-based method *Local binary pattern* (LBP) is a texture-based method that quantifies intensity patterns in the neighbourhood of the pixel [88]. Zhang et al. [89] proposed the *multi-block local binary pattern* (MB-LBP) to encode intensities of the rectangular regions by LBP. HOG [44] introduced another texture-based method which uses high-dimensional features based on edges and then applies SVM to detect human body regions. This technique counts the occurrences of gradient orientation in localized portions of an image, is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. Zhu et al. [90] applied the HOG descriptors in combination with the cascade of *rejecters* algorithm and introduced blocks that vary in size, location and aspect ratio. In order to isolate the blocks best suited for human detection, they applied the *AdaBoost* algorithm to select those blocks to be included in the rejecter cascade. Moctezuma et al. [91] proposed *HOG with Gabor filter* and showed improved performances in both person counting and identification.

1.2.1.9 Detection of non-moving human We have focused on motion-based human detection in this study

due to the fact that some unique human motion features aid in better identification of human beings from other objects [92]. The method for human detection from static images has also a number of applications such as smart rooms and visual surveillance. A human detection scheme in a crowded scene from static images is described in [93]. The method models an individual human as an assembly of natural body parts using edgelet features, which are a new type of silhouette-oriented features. Local body part and global shape-based approach showed promising results [40]. Probability part detector has been used successfully for human detection [94]. A learning-based human detection framework was proposed earlier by Papageorgiou et al. [95]. Recently, motionless human detection based on sensor data has been proposed with particular application interests in the area of aged care support [96,97].

1.2.1.10 Comparisons of classification techniques A comparison among object classification methods in terms of accuracy and computational time is presented in Table 3. The table shows accuracy and computational time of different object classification techniques in terms of three criteria, namely low, moderate and high. As we have mentioned earlier, it is very difficult to conclude the accuracy and computational time of different techniques in each category by three simple attributes (e.g. low, moderate and high) because in each category, there are a number of techniques and each technique has its own accuracy and computational time. However, we have provided average or normal trend of the techniques in each category to give an overall understanding of a category.

The main criticism of the shape-based approach with templates for human detection is that local deformation

Table 3 Comparison of object classification methods in terms of accuracy and computational time

Methods	Accuracy	Computational time	Comments
Shape-based method [18,23,38-43,85]	Moderate	Low	Simple pattern-matching approach can be applied with appropriate templates. It does not work well in dynamic situations and is unable to determine internal movements well
Motion-based method [26,86,87]	Moderate	High	Does not require predefined pattern templates but struggles to identify a non-moving human
Texture-based method [44,88-91]	High	High	Provides improved quality with the expense of additional computation time

of body parts due to motion could not be captured properly thus provides less accurate performance compared to other methods. However, if the methods use fixed templates, they might provide a slightly better performance than SVM-based variations and process reasonably faster [40]. The motion-based approaches use predefined actions to recognize the human motions. As these approaches need to process motion and then categorize the object, they need more computational time. The texture-based approaches also work similar to motion-based approaches but with the help of texture pattern recognition. They provide better accuracy (around 10%) [91,98] but may require more time, which can be improved using some fast techniques [90].

1.3 Benchmark datasets for indoor and outdoor

In this section, a brief overview of few datasets for surveillance-based research has been presented.

1.3.1 KTH human motion dataset

KTH dataset [45] is the largest available and most standard dataset widely used for benchmarking results for human action classification. The dataset contains six activities

(boxing, hand waving, handclapping, running, jogging and walking) performed by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). There are $25 \times 6 \times 4 = 600$ video files for each combination of 25 subjects, six actions and four scenarios. All sequences were taken over homogeneous backgrounds with a static camera with 25 *frames per second* (fps) frame rate. The sequences were then down-sampled to the spatial resolution of 160×120 pixels and have a length of 4 s in average. Some sample sequences are shown in Figure 4.

1.3.2 Weizmann human action dataset

Weizmann human action dataset [46] contains a total of ten actions performed by nine people, to provide a total of 90 videos. Sample sequences are shown in Figure 5. The dataset contains videos with a static camera unlike that of the KTH dataset, where some of the videos had zooming and also have simple background. As this dataset contains ten activities, which is more comparative to the six activities of the KTH dataset, it provides a good test to the approach in the setting in which the number of activities are increased.



Figure 4 Sample sequences from KTH human motion dataset.



Figure 5 Example sequences from Weizmann dataset (jack, walk, wave1, skip, side, bend, p-jump, wave2, run and jump).

1.3.3 PETS dataset

PETS datasets [48] have a number of datasets for different purposes of vision-based research. Each year, PETS run an evaluation framework on specific datasets with specific objective. PETS'2000 [49] and PETS'2001 [50] datasets are designed for tracking outdoor people and vehicles. PETS'2000 used a single camera, while PETS'2001 used two synchronized views. The later datasets are significantly more challenging than the previous one in terms of significant lighting variation, occlusion, scene activity and use of multi-view data. Two sample images are shown in Figure 6. PETS'2002 [51] has indoor people tracking (and counting) and hand posture classification data. PETS-ICVS'2003 [52] has annotations of a smart meeting, which includes facial expressions, gaze and gesture/action. VS-PETS'2003 [53] has outdoor people tracking - football data from two synchronized camera views. PETS-ECCV'2004 [54] has a number of video clips recorded for the CAVIAR project. These include people walking alone, meeting with others, window shopping, fighting and passing out and, last but not least, leaving a package in a public place. All video clips were filmed for the CAVIAR project with a wide-angle camera lens in the entrance lobby of the INRIA Labs at

Grenoble, France. PETS'2006 [55] has surveillance data of public spaces and detection of left luggage events. PETS'2007 [56] considers both volume crime (theft) and a threat scenario (unattended luggage.) The datasets for PETS'2009 [57], PETS'2010 [58] and PETS'2012 [59] consider crowd image analysis and include crowd count and density estimation, tracking of individual(s) within a crowd and detection of separate flows and specific crowd events.

1.3.4 INRIA XMAS multi-view dataset

Weinland et al. [47] introduced the INRIA XMAS dataset that contains actions captured from five viewpoints. A total of 11 persons perform 14 actions (check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw over head and throw from bottom up). The actions are performed in an arbitrary direction with regard to the camera set-up. The camera views are fixed, with a static background and illumination settings. Silhouettes and volumetric voxel representations are part of the dataset.

1.3.5 Other datasets

The Institute of Automation, Chinese Academy of Sciences provides the CASIA Gait Database for gait recognition



Figure 6 Sample images from PETS'2001 dataset.

and related research. The database consists of three datasets: dataset A, dataset B (multi-view dataset) and dataset C (infrared dataset). The details of these databases are found in [60].

The Hollywood human action dataset [99] contains eight actions (answer phone, get out of car, handshake, hug, kiss, sit down, sit up and stand up), which are extracted from movies and performed by a variety of actors. A second version of the dataset includes four additional actions (drive car, eat, fight and run) and an increased number of samples for each class. One training set is automatically annotated using scripts of the movies; another is manually labelled. There is a huge variety of performance of the actions, both spatially and temporally. Occlusions, camera movements and dynamic backgrounds make this dataset challenging. Most of the samples are at the scale of the upper body, but some show the entire body or a close up of the face.

The UCF sports action dataset [100] contains 150 sequences of sport motions (diving, golf swinging, kicking, weightlifting, horseback riding, running, skating, swinging a baseball bat and walking). Bounding boxes of the human figure are provided with the dataset. For most action classes, there is considerable variation in action performance, human appearance, camera movement, viewpoint, illumination and background.

The Wallflower dataset [79] contains seven scenarios: one on them is outdoor, and six are indoor. The scenarios include moved object, time of day, light switch, waving tree, camouflage, bootstrapping and foreground aperture. In this dataset, for each scenario, training and test sequences are provided along with hand-drawn ground truth for one specific frame.

1.4 Applications

For an intelligent video surveillance system, the detection of a human being is important for abnormal event detection, human gait characterization, people counting, person identification and tracking, pedestrian detection, gender classification, fall detection of elderly people, etc.

1.4.1 Abnormal event detection

The most obvious application of detecting humans in surveillance video is to early detect an event that is not normal. Candamoo et al. [18] classified the abnormal events as single-person loitering, multiple-person interactions (e.g. fighting and personal attacks), person-vehicle interactions (e.g. vehicle vandalism), and person-facility/location interactions (e.g. object left behind and trespassing). Detecting sudden changes and motion variations in the points of interest and recognizing human action could be done by constructing a motion similarity matrix [26] or adopting a probabilistic method [101]. Methods based on probability statistics use the minimum change

of time and space measure to model the method of probability. The most representative probability chart model is HMM. In addition, also there is *conditional random field*, the *maximum entropy Markov model* and *dynamic Bayesian network*. More information on human action recognition techniques for abnormal event detection can be found in [102].

1.4.2 Human gait characterization

Ran et al. [103] detected humans in walking by extracting *double helical signatures* (DHS) from surveillance video sequences. They found that DHS is robust to size, viewing angles, camera motion and severe occlusion for simultaneous segmentation of humans in periodic motion and labelling of body parts in cluttered scenes. They used the change in DHS symmetry for detecting humans in normal walking, carrying an object with one hand, holding an object in both hands, attaching an object to the upper body and attaching an object to the legs. Although DHS is independent of silhouettes or landmark tracking, it is ineffective when the target walks toward the camera as the DHS degenerates into ribbon and no strong symmetry can be observed. Cutler et al. [87] used the area-based image similarity technique to address this issue and detected the motion of a person who was walking at approximately 25° offset the camera's image plane from a static camera. They segmented the motion and track objects in the foreground. Each object was then aligned along the temporal axis (using the object's tracking results), and the object's self-similarity was computed as it evolves in time. For periodic motions, the self-similarity metric is periodic, and they apply time-frequency analysis to detect and characterize the periodicity.

1.4.3 Person detection in dense crowds and people counting

Detecting and counting persons in a dense crowd is challenging due to occlusions. Eshel and Moses [104] used *multiple height homographies* for head top detection to overcome this problem. Yao and Odobez [105] proposed to take advantage of the stationary cameras to perform background subtraction and jointly learn the appearance and the foreground shape of people in videos. Sim et al. [106] proposed a representation called the colour bin image which is extracted from the initially detected windows, and they use it for training a classifier to improve the performance of the initial detector. The proposed system was applied for detecting individual heads in dense crowds of 30 to 40 people against cluttered backgrounds from a single video frame. However, the performance of their approach may be challenged by the colour intensities of the heads to be detected. Chen et al. [107] proposed an online people counting system for electronic advertising machines. A vision-based people counting model was

proposed by Chih-Wen et al. [108]. The cross camera people counting model proposed by Lin et al. [109] was composed of a pair of collaborative *Gaussian processes*, which were respectively designed to count people by taking the visible and occluded parts into account. Weng et al. [110] also presented an algorithm for accomplishing cross camera correspondence and proposed a counting model which was composed of a pair of collaborative regressors. A multi-camera people counting technique with occlusion handling is presented by Weng et al. [70]. Recently, Chen and Huang proposed two crowd behaviour detection models based on motion [111] and visual with graph and matching [112].

1.4.4 Person tracking and identification

A person in a visual surveillance system can be identified using face recognition [85,113-122] and gait recognition [123-131] techniques. The detection and tracking of multiple people in cluttered scenes at public places is difficult due to a partial or full occlusion problem for either a short or long period of time. Leibe et al. [132] tried to address this issue using trajectory estimation while Andriluka et al. [133] used a tracklet-based detector, which was capable of detecting several partially occluded people that cannot be detected in a single frame alone. Yilmaz et al. [134] made a comprehensive survey on tracking methods and categorized them on the basis of the object and motion representations used. The wider application of human detection is not only limited to analysis surveillance videos but also extended to player tracking and identification in sport videos. The system introduced by Lu et al. [135] identified players in broadcast sports videos using conditional random fields and achieved a player recognition accuracy up to 85% on unlabeled NBA basketball clips. Sun et al. [136] proposed an *individual level sports video indexing* scheme, where a principal axis-based contour descriptor is used to solve the jersey number recognition problem. Lu et al. [137] proposed a novel *linear programming* relaxation algorithm for predicting player identification in a video clip using weakly supervised learning with play-by-play texts, which greatly reduced the number of labelled training examples required.

1.4.5 Gender classification

Gender classification is another application of human detection in surveillance cameras. The classification could be carried out by fusion of similarity measures from multi-view gait sequences [138], exploiting separability of features from different views [139] and training a linear SVM classifier based on the averaged gait image [140]. Cao et al. [141] introduced a *part-based gender recognition* algorithm using patch features for modelling different body parts, which could recognize the gender from either

a single frontal or back view image with the accuracy of 75.0% and is robust to tolerate small misalignment errors. Recently, Hu et al. [142] integrated shape appearance and temporal dynamics of both genders into a sequential model called *mixed conditional random field* (MCRF). By fusion of shape descriptors and stance indexes, the MCRF is constructed in coordination with intra- and inter-gender temporary *Markov properties*. Their results showed the superior performance of the MCRF over HMMs and separately trained *conditional random field*. A new face-based gender recognition technique has been proposed by Chen and Hsieh which shows strong gender recognition capabilities [143].

1.4.6 Pedestrian detection

Pedestrian detection is another important application of human detection. Viola et al. [144] described a pedestrian detection system that integrates image intensity information with motion information. Their detector was built over two consecutive frames of a video sequence and was based on motion direction filters, motion shear filters, motion magnitude filters and appearance filters. Their system detected pedestrians from a variety of viewpoints with a low false positive rate using multiple classifiers with cascade architecture. A pedestrian could also be detected by extracting *regions of interest* (ROI) from an image and then sending it to a classification module for detection. However, ROIs must fulfil the pedestrian size constraints, i.e. the aspect ratio, size and position, to be considered to contain a pedestrian [145]. Chen [146] proposed the orientation filter-enhanced detection technique based on the combination of AdaBoost learning with a local histogram's features which shows better performance and robustness.

1.4.7 Fall detection for elderly people

Automatic detection of a fall for elderly people is one of the major applications of human detection in surveillance videos. Nasution and Emmanuel [147] used the *projection histograms* of segmented human body silhouette as the main feature vector posture classification and used the speed of fall to differentiate real fall incident and an event where a person is simply lying without falling. Thome and Miguet [148] proposed a multi-view (two-camera) approach to address occlusion and used a *layered* HMM for motion modelling where the hierarchical architecture decoupled the motion analysis into different temporal granularity levels, which made the algorithm able to detect very sudden changes.

1.5 Discussion

A significant amount of work has been done with a view to detect human beings in a surveillance video. However the low-resolution images from the surveillance cameras

always make this work challenging. Most of the object detection methods rely on known operation environments. The model adaptation speed based on observed scene statistics could be improved in the future for faster adaptation of changed background and better persistency. However, occlusion is a major problem for background segmentation technique. Optical flow and spatio-temporal filter techniques address this issue to some extent where the object of interest is occluded by a fixed object, but it is always difficult to detect an object in motion which is occluded by objects with similar shape and motion. One solution could be constructing a 3D image for a 3D system using volume information obtained from multiple cameras.

From the machine vision perspective, it is hard to distinguish an object as a human due to its large number of possible appearances [102]. Moreover, the human motion is not always periodic, but a combination of features could be useful in identifying humans. Interesting progress is being made using a local-based approach [149] for human detection. Future models based on LBP and HOGs might have several benefits over other descriptor methods as they work on localized parts of the image and hence are capable of addressing occlusion problems.

2 Conclusions

Detecting human beings accurately in a surveillance video is one of the major topics of vision research due to its wide range of applications. It is challenging to process the image obtained from a surveillance video as it has low resolution. A review of the available detection techniques is presented. The detection process occurs in two steps: object detection and object classification. In this paper, all available object detection techniques are categorized into background subtraction, optical flow and spatio-temporal filter methods. The object classification techniques are categorized into shape-based, motion-based and texture-based methods. The characteristics of the benchmark datasets are presented, and major applications of human detection in surveillance video are reviewed.

At the end of this paper, a discussion is made to point the future work needed to improve the human detection process in surveillance videos. These include exploiting a multi-view approach and adopting an improved model based on localized parts of the image.

Competing interests

The authors declare that they have no competing interests.

Received: 10 May 2013 Accepted: 31 October 2013

Published: 22 November 2013

References

1. N Sulman, T Sanocki, D Goldgof, R Kasturi, How effective is human video surveillance performance? in *19th International Conference on Pattern Recognition, (ICPR 2008)* (IEEE, Piscataway, 2008), pp. 1–3

2. C Stauffer, W Grimson, Adaptive background mixture models for real-time tracking, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 1999)* (IEEE, Piscataway, 1999), pp. 246–252
3. YL Tian, RS Feris, H Liu, A Hampapur, M-T Sun, Robust detection of abandoned and removed objects in complex surveillance videos. *Syst. Man Cybern. Part C Appl. Rev. IEEE Trans.* **41**(5), 565–576 (2011)
4. DS Lee, Effective Gaussian mixture learning for video background subtraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(5), 827–835 (2005)
5. A Shimada, D Arita, Dynamic control of adaptive mixture-of-Gaussians background model, in *IEEE International Conference on Video and Signal Based Surveillance (AVSS'06)* (IEEE, Piscataway, 2006), p. 5
6. J Wang, G Bebis, R Miller, Robust video-based surveillance by integrating target detection with tracking, in *IEEE Computer Vision and Pattern Recognition Workshop (CVPRW '06)* (IEEE, Piscataway, 2006), p. 137
7. C Ridder, O Munkelt, and H. Kirchner, Adaptive Background Estimation and Foreground Detection Using Kalman-Filtering, *Proc. Int'l Conf. Recent Advances in Mechatronics, ICRAM 95*, pp. 193–199 (1995)
8. B Stenger, V Ramesh, N Paragios, F Coetzee, JM Buhmann, Topology free hidden Markov models: application to background modeling, in *IEEE International Conference on Computer Vision (ICCV 2001)* (IEEE, Piscataway, 2001), pp. 294–301
9. S Jabri, Z Duric, H Wechsler, A Rosenfeld, Detection and location of people in video images using adaptive fusion of color and edge information, in *15th International Conference on Pattern Recognition (ICPR2000)* (IEEE, Piscataway, 2000), pp. 627–630
10. W Zhang, X Zhong, FY Xu, Detection of moving cast shadows using image orthogonal transform, in *18th International Conference on Pattern Recognition (ICPR'06)* (IEEE, Piscataway, 2006), pp. 626–629
11. M Heikkilä, M Pietäikinen, A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 657–662 (2006)
12. W Kim, C Kim, Background subtraction for dynamic texture scenes using fuzzy color histograms. *Signal Process. Lett. IEEE* **19**(3), 127–130 (2012)
13. A Elgammal, D Harwood, L Davis, Non-parametric model for background subtraction, in *6th European Conference on Computer Vision - Part II (ECCV '00)* (Springer, London, 2000), pp. 751–767
14. A Elgammal, R Duraiswami, L Davis, Efficient kernel density estimation using the fast Gauss transform with applications to color modeling and tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 1499 (2003)
15. B Han, D Comaniciu, L Davis, Sequential kernel density approximation through mode propagation: Applications to background modeling, in *Asian Conference on Computer Vision Jeju Island, Korea* (2004)
16. L Li, W Huang, Y-H Gu, Q Tian, Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. Image Process.* **13**, 1459–1472 (2004)
17. A Lanza, Background subtraction by non-parametric probabilistic clustering, in *8th IEEE International Conference on Advanced Video and Signal-Based Surveillance* (IEEE, Piscataway, 2011), pp. 243–248
18. J Candamo, M Shreve, DB Goldgof, DB Sapper, R Kasturi, Understanding transit scenes: A survey on human behavior-recognition algorithms. *IEEE Trans. Intell. Transp. Syst.* **11**(1), 206–224 (2010)
19. F-C Cheng, S-C Huang, S-J Ruan, Scene analysis for object detection in advanced surveillance systems using Laplacian distribution model. *Syst. Man Cybern. Part C Appl. Rev. IEEE Trans.* **41**(5), 589–598 (2011)
20. D-M Tsai, S-C Lai, Independent component analysis-based background subtraction for indoor surveillance. *IEEE Trans. Image Process.* **18**(1), 158–167 (2009)
21. T Ko, S Soatto, D Estrin, Warping background subtraction, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)* (IEEE, Piscataway, 2010), pp. 1331–1338
22. S Chen, J Zhang, Y Li, J Zhang, A hierarchical model incorporating segmented regions and pixel descriptors for video background subtraction. *IEEE Trans. Ind. Inform.* **8**(1), 118–127 (2012)
23. J Xiaofei, L Honghai, Advances in view-invariant human motion analysis: a review. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **40**(1), 13–24 (2010)
24. JL Barren, DJ Fleet, SS Beauchemin, TA Burkitt, Performance of optical flow techniques, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '92)* (IEEE, Piscataway, 1992), pp. 236–242
25. H Jeon, J Jeong, J Bang, C Hwang, The efficient features for tracking, in *20th IEEE International Conference on Tools with Artificial Intelligence, 2008 (ICTAI '08)* (IEEE, Piscataway, 2008), pp. 241–244
26. A Efros, A Berg, G Mori, J Malik, Recognizing action at a distance, in *Ninth IEEE International Conference on Computer Vision (ICCV 2003)* (IEEE, Piscataway, 2003), pp. 726–733

27. H Zhong, J Shi, M Visontai, Detecting unusual activity in video, in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)* (IEEE, Piscataway, 2004), pp. 819–826
28. I Laptev, On space-time interest points. *Int. J. Comput. Vis.* **64**(2–3), 107–123 (2005)
29. P Dollár, V Rabaud, G Cottrell, S Belongie, Behavior recognition via sparse spatio-temporal features, in *2nd IEEE Joint International Workshop Visual Surveillance and Performance Evaluation of Tracking Surveillance* (IEEE, Piscataway, 2005), pp. 65–72
30. SA Niyogi, EH Adelson, Analyzing and recognizing walking figures in XYT, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1994 (CVPR '94)* (IEEE, Piscataway, 1994), pp. 469–474
31. SA Niyogi, EH Adelson, Analyzing gait with spatio-temporal surface, in *1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects* (IEEE, Piscataway, 1994), pp. 64–69
32. C BenAbdelkader, R Cutler, H Nanda, L Davis, EigenGait: motion-based recognition of people using image self-similarity, in *Audio- and Video-Based Biometric Person Authentication*, ed. by J Bigun, F Smeraldi. Third International Conference, AVBPA 2001 Halmstad, Sweden, 6–8 June 2001. Lecture notes in Computer Science, vol. 2091 (Springer, Heidelberg, 2001), pp. 312–317
33. A Kale, A Rajagopalan, N Cuntoor, V Kruger, Gait-based recognition of humans using continuous HMMs, in *5th IEEE International Conference on Automatic Face and Gesture Recognition* (IEEE, Piscataway, 2002), pp. 336–341
34. C BenAbdelkader, R Cutler, L Davis, Motion-based recognition of people in eigengait space, in *5th IEEE International Conference on Automatic Face and Gesture Recognition* (IEEE, Piscataway, 2002), pp. 267–274
35. R Collins, R Gross, J Shi, Silhouette-based human identification from body shape and gait, in *5th IEEE International Conference on Automatic Face and Gesture Recognition* (IEEE, Piscataway, 2002), pp. 366–371
36. L Wang, HZ Ning, WM Hu, Gait recognition based on procrustes statistical shape analysis, in *2002 International Conference on Image Processing (ICIP2002)* (IEEE, Piscataway, 2002), pp. 433–436
37. R Piroddi, T Vlachos, A simple framework for spatio-temporal video segmentation and delayering using dense motion fields. *IEEE Signal Process. Lett.* **13**(7), 421 (2006)
38. L Wang, X Geng, C Leckie, R Kotagiri, Moving shape dynamics: a signal processing perspective, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)* (IEEE, Piscataway, 2008), pp. 1–8
39. M Singh, A Basu, MK Mandal, Human activity recognition based on silhouette directionality. *IEEE Trans. Circuits Syst. Video Technol.* **18**(9), 1280–1292 (2008)
40. Z Lin, LS Davis, Shape-based human detection and segmentation via hierarchical part-template matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(4), 604–618 (2010)
41. B Wu, R Nevatia, Detecting and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *Int. J. Comput. Vision (IJCV)* **75**(2), 247–266 (2007)
42. DM Gavriila, A Bayesian, exemplar-based approach to hierarchical shape matching. *PAMI* **29**(8), 1408–1421 (2007)
43. FZ Eishita, A Rahman, SA Azad, A Rahman, Occlusion handling in object detection. Multidisciplinary computational intelligence techniques: applications in business, engineering, and medicine. IGI Global. (2013). doi:10.4018/978-1-4666-1830-5.ch005
44. N Dalal, B Triggs, Histograms of oriented gradients for human detection, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)* (IEEE, Piscataway, 2005), pp. 886–893
45. C Schuld, I Laptev, B Caputo, Recognizing human actions: a local SVM approach, in *17th International Conference on Pattern Recognition (ICPR 2004)* (IEEE, Piscataway, 2004), pp. 32–36
46. M Blank, L Gorelick, E Shechtman, M Irani, R Basri, M Blank, L Gorelick, E Shechtman, M Irani, R Basri, Actions as space-time shapes, in *Tenth IEEE International Conference on Computer Vision (ICCV '05)* (IEEE, Piscataway, 2005), pp. 1395–1402
47. D Weinland, R Ronfard, E Boyer, Free viewpoint action recognition using motion history volumes. *Comput. Vision Image Understanding (CVIU)* **104**(2–3), 249–257 (2006)
48. PETS, Performance Evaluation of Tracking and Surveillance. <http://www.cvg.rdg.ac.uk/slides/pets.html>. Accessed 17 Nov 2013
49. PETS, (2000). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
50. PETS, (2001). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
51. PETS, (2002). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
52. PETS ICVS, (2013). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
53. VS PETS, (2013). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
54. R Fisher, CAVIAR test case scenarios. (2007). <http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/>. Accessed 17 Nov 2013
55. PETS, (2006). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
56. PETS, (2007). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
57. PETS, (2009). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
58. PETS, (2010). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
59. PETS, (2012). <http://ftp.pets.rdg.ac.uk>. Accessed 17 Nov 2013
60. CBSR, CASIA Gait Database, (2005). <http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp>. Accessed 17 Nov 2013
61. P Turaga, R Chellappa, VS Subramanian, O Udrea, Machine recognition of human activities: a survey. *IEEE Trans. Circuits Syst. Video Technol.* **18**(11), 1473–1488 (2008)
62. G Lavee, E Rivlin, M Rudzsky, Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Trans. Syst., Man, Cybern. C* **39**(5), 489–504 (2009)
63. W Hu, T Tan, L Wang, S Maybank, A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst., Man, Cybern. Part C, Appl. Rev.* **34**(3), 334–352 (2004)
64. H-H Lin, T-L Liu, J-H Chuang, Learning a scene background model via classification. *IEEE Trans. Signal Process.* **57**(5), 1641–1654 (2009)
65. T Du-Ming, L Shia-Chih, Independent component analysis-based background subtraction for indoor surveillance. *IEEE Trans. Image Process.* **18**(1), 158–167 (2009)
66. Y-T Chen, C Chu-Song, H Chun-Rong, H Yi-Ping, Efficient hierarchical method for background subtraction. *Pattern Recognit.* **40**, 2706–2715 (2007)
67. S Quan, T Zhixing, H Songchen, Hierarchical CodeBook for background subtraction in MRF. *Infrared Phys. Technol.* **61**, 259–264 (2013)
68. K Kim, TH Khalidabhongse, D Harwood, L Davis, Real-time foreground-background segmentation using codebook model. *Real-time Imaging* **11**(3), 172–185 (2005)
69. AJ Lipton, H Fujiyoshi, RS Patil, Moving target classification and tracking from real-time video, in *Fourth IEEE Workshop on Applications of Computer Vision (WACV'98)* (IEEE, Piscataway, 1998), pp. 8–14
70. H-E Lai, C-Y Lin, M-K Chen, L-W Kang, C-H Yeh, Moving objects detection based on hysteresis thresholding. *Adv. Intell. Syst. Appl.* **2**, 289–298 (2013)
71. M Haque, M Murshed, M Paul, A hybrid object detection technique from dynamic background using Gaussian mixture models, in *IEEE 10th Workshop on Multimedia Signal Processing* (IEEE, Piscataway, 2008), pp. 915–920
72. M Paul, C Evans, M Murshed, Disparity-adjusted 3D multi-view video coding with dynamic background modelling, in *IEEE International Conference on Image Processing (ICIP 2013)* (IEEE, Piscataway, 2013)
73. M Paul, M Murshed, Video coding focusing on block partitioning and occlusions. *IEEE Trans. Image Process.* **19**(3), 691–701 (2010)
74. M Paul, W Lin, CT Lau, BS Lee, Explore and model better I-frame for video coding. *IEEE Trans. Circuits Syst. Video Technol.* **21**, 1242–1254 (2011)
75. M Paul, W Lin, CT Lau, BS Lee, Video coding with dynamic background. *EURASIP J. Adv. Signal Process.* (2013). doi:10.1186/1687-6180-2013-11
76. A Sobral, BGSLibrary: an OpenCV C++ background subtraction library. (2010). <https://code.google.com/p/bgslibrary/>. Accessed 17 Nov 2013
77. R Cucchiara, C Grana, M Piccardi, A Prati, Detecting moving objects, ghosts, and shadows in video streams. *Pattern Anal. Mach. Intell. IEEE Trans.* **25**(10), 1337–1342 (2003)
78. K Toyama, J Krumm, B Brumitt, B Meyers, Wallflower: principles and practice of background maintenance, in *Seventh IEEE International Conference on Computer Vision (ICCV 1999)* (IEEE, Piscataway, 1999), pp. 255–261
79. J Krumm, Test images for Wallflower paper. (1999). <http://research.microsoft.com/en-us/um/people/jkrumm/wallflower/testimages.htm>. Accessed 17 Nov 2013
80. DH Parks, SS Fels, Evaluation of background subtraction algorithms with post-processing, in *IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance (AVSS '08)* (IEEE, Piscataway, 2008), pp. 192–199
81. S-CS Cheung, C Kamath, Robust techniques for background subtraction in urban traffic video. *SPIE04* **5308**, 881–892 (2004)
82. S-CS Cheung, C Kamath, Robust background subtraction with foreground validation for urban traffic video. *JASPO5* **14**, 2330–2340 (2005)
83. T Wang, S Gong, C Liu, Y Ji, An improved warping background subtraction model for moving object detection, in *2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)* (IEEE, Piscataway, 2011), pp. 668–672

84. TH Chalidabhongse, K Kim, D Harwood, L Davis, A perturbation method for evaluating background subtraction algorithms, in *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (IEEE, Piscataway, 2011), pp. 11–12
85. P Huang, A Hilton, J Starck, Shape similarity for 3D video sequences of people. *Int. J. Comput. Vision* **89**(2–3), 362–381 (2010)
86. AF Bobick, JW Davis, The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(3), 257–267 (2001)
87. R Cutler, LS Davis, Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 781–796 (2000)
88. T Ojala, M Pietikinen, T Maenpaa, Multi-resolution grayscale and rotation invariant texture classification with local binary patterns. *PAMI* **24**(7), 971–987 (2002)
89. L Zhang, SZ Li, X Yuan, S Xiang, Real-time object classification in video surveillance based on appearance learning, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2007 (CVPR 2007)* (IEEE, Piscataway, 2007), pp. 1–8
90. Q Zhu, S Avidan, M-C Yeh, K-T Cheng, Fast human detection using a cascade of histograms of oriented gradients, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2006 (CVPR '06)* (IEEE, Piscataway, 2006), pp. 1491–1498
91. D Moctezuma, C Conde, IM Diego, E Cabello, Person detection in surveillance environment with HoGG: Gabor filters and histogram of oriented gradient, in *ICCV Workshops* (IEEE, Piscataway, 2011), pp. 1793–1800
92. N Dalal, B Triggs, C Schmid, Human detection using oriented histograms of flow and appearance, in *Computer Vision—ECCV* (Springer, Heidelberg, 2006), pp. 428–441
93. B Wu, R Nevatia, Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors, in *IEEE International Conference on Computer Vision (ICCV 2005)* (IEEE, Piscataway, 2005), pp. 90–97
94. K Mikolajczyk, C Schmid, A Zisserman, Human detection based on a probabilistic assembly of robust part detectors, in *Computer Vision—ECCV 2004* (Springer, Heidelberg, 2004), pp. 69–82
95. CP Papageorgiou, M Oren, T Poggio, A general framework for object detection, in *IEEE Sixth International Conference on Computer Vision* (IEEE, Piscataway, 1998), pp. 555–562
96. S Zhang, P McCullagh, C Nugent, H Zheng, A theoretic algorithm for fall and motionless detection, in *IEEE Third International Conference on Pervasive Computing Technologies for Healthcare* (IEEE, Piscataway, 2009), pp. 1–6
97. D Curone, GM Bertolotti, A Cristiani, EL Secco, G Magenes, A real-time and self-calibrating algorithm based on triaxial accelerometer signals for the detection of human posture and activity. *IEEE Trans. Info. Technol. Biomed.* **14**, 1098–1105 (2010)
98. C Conde, D Moctezuma, I Martín De Diego, E Cabello, HoGG: Gabor and HoG-based human detection for surveillance in non-controlled environments. *Neurocomputing* **100**, 19–30 (2013)
99. M Marszalek, I Laptev, C Schmid, Actions in context, in *Conference on Computer Vision and Pattern Recognition (CVPR'09)* (IEEE, Piscataway, 2009), pp. 2929–2936
100. MD Rodriguez, J Ahmed, M Shah, Action MACH: a spatiotemporal maximum average correlation height filter for action recognition, in *Conference on Computer Vision and Pattern Recognition (CVPR'08)* (IEEE, Piscataway, 2008), pp. 1–8
101. AF Bobick, YA Ivanov, Action recognition using probabilistic parsing, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08)* (IEEE, Piscataway, 2008), pp. 196–202
102. R Poppe, A survey on vision-based human action recognition. *Image Vision Comput.* **28**, 976–990 (2010)
103. Y Ran, Q Zheng, R Chellappa, TM Strat, Applications of a simple characterization of human gait in surveillance. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **40**(4), 1009–1020 (2010)
104. R Eshel, Y Moses, Homography based multiple camera detection and tracking of people in a dense crowd, in *Conference on Computer Vision and Pattern Recognition (CVPR'08)* (IEEE, Piscataway, 2008), pp. 1–8
105. J Yao, JM Odobez, Fast human detection from joint appearance and foreground feature subset covariances. *Comput. Vision Image Understanding* **115**, 1414–1426 (2011)
106. C-H Sim, E Rajmadhan, S Ranganath, Detecting people in dense crowds. *Mach. Vision Appl.* **23**, 243–253 (2012)
107. D-Y Chen, C-W Su, Y-C Zeng, S-W Sun, W-R Lai, H-Y Mark Liao, An online people counting system for electronic advertising machines, in *IEEE International Conference on Multimedia and Expo (ICME 2009)* (IEEE, Piscataway, 2009), pp. 1262–1265
108. C-W Su, H-YM Liao, H-R Tyan, A vision-based people counting approach based on the isometry measure, in *IEEE International Symposium on Circuits and Systems (ISCAS 2009)* (IEEE, Piscataway, 2009), pp. 2617–2620
109. TY Lin, YY Lin, MF Weng, YCF Wang, YF Hsu, HYM Liao, Cross camera people counting with perspective estimation and occlusion handling, in *2011 IEEE International Workshop on Information Forensics and Security (WIFS 2011)* (IEEE, Piscataway, 2011), pp. 1–6
110. M-F Weng, Y-Y Lin, NC Tang, H-Y Mark Liao, Visual knowledge transfer among multiple cameras for people counting with occlusion handling, in *20th ACM international conference on Multimedia Pages* (ACM, New York, 2012), pp. 439–448
111. D-Y Chen, P-C Huang, Motion-based unusual event detection in human crowds. *J. Visual Commun. Image Represent.* **22**(2), 178–186 (2011)
112. C Duan-Yu, H Po-Chung, Visual-based human crowds behavior analysis based on graph modeling and matching. *Sens. J. IEEE* **13**(6), 2129–2138 (2013)
113. A Samal, PA Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognit.* **25**(1), 65–77 (1992)
114. R Chellappa, CL Wilson, S Sirohey, Human and machine recognition of faces: a survey. *Proc. IEEE* **83**, 705–741 (1995)
115. D Swets, J Weng, Discriminant analysis and eigenspace partition tree for face and object recognition from views, in *Second International Conference on Automatic Face and Gesture Recognition* (IEEE, Piscataway, 1996), pp. 182–187
116. B Moghaddam, W Wahid, A Pentland, Beyond eigenfaces: probabilistic matching for face recognition, in *Third IEEE International Conference on Automatic Face and Gesture Recognition* (IEEE, Piscataway, 1998), pp. 30–35
117. G Guo, S Li, K Chan, Face recognition by support vector machines, in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (IEEE, Piscataway, 2000), pp. 196–201
118. H Rowley, S Baluja, T Kanade, Neural network based face detection. *IEEE Trans. Pattern Anal. Machine Intell.* **20**, 23–38 (1998)
119. C Garcia, G Tziritas, Face detection using quantified skin color regions merging and wavelet packet analysis. *IEEE Trans. Multimedia* **1**, 264–277 (1999)
120. B Menser, M Wien, Segmentation and tracking of facial regions in color image sequences. *Proc. SPIE Visual Communications and Image Processing* **4067**, 731–740 (2000)
121. A Saber, AM Tekalp, Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognit. Lett.* **19**(8), 669–680 (1998)
122. G Xu, T Sugimoto, Rits Eye: a software-based system for real-time face detection and tracking using pan-tilt-zoom controllable camera, in *Fourteenth International Conference on Pattern Recognition (ICPR 1998)* (IEEE, Piscataway, 1998), pp. 1194–1197
123. D Cunado, MS Nixon, JN Carter, Using gait as a biometric: via phase-weighted magnitude spectra, in *First International Conference on Audio- and Video-Based Biometric Person Authentication* (Springer, London, 1997), pp. 95–102
124. D Cunado, MS Nixon, JN Carter, Extracting a human gait model for use as a biometric. *Proc. Inst. Elect. Eng. (IEE) Colloq. Computer Vision for Virtual Human Modelling* **11**, 1–4 (1998)
125. JM Nash, JN Carter, MS Nixon, Dynamic feature extraction via the velocity Hough transform. *Pattern Recognit. Lett.* **18**(10), 1035–1047 (1997)
126. CY Yam, MS Nixon, JN Carter, Extended model-based automatic gait recognition of walking and running, in *International Conference on Audio- and Video-Based Biometric Person Authentication* (Springer, Heidelberg, 2001), pp. 278–283
127. CY Yam, MS Nixon, JN Carter, Gait recognition by walking and running: a model-based approach, in *Fifth Asian Conference Computer Vision (ACCV2002)* (Melbourne, 2002)
128. D Cunado, J Nash, MS Nixon, JN Carter, Gait extraction and description by evidence gathering, in *Second International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA99)* (Washington, DC, 1999)
129. R Tanawongsuwan, A Bobick, Gait recognition from time-normalized joint-angle trajectories in the walking plane, in *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)* (IEEE, Piscataway, 2011), pp. 726–731
130. R Murase, Sakai, Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recognit. Lett.* **17**(2), 155–162 (1996)
131. PS Huang, CJ Harris, MS Nixon, Human gait recognition in canonical space using temporal templates. *Proc. Inst. Elect. Eng. (IEE) Vision Image and Signal Process.* **146**(2), 93–100 (1999)

132. B Leibe, E Seemann, B Schiele, Pedestrian detection in crowded scenes, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)* (IEEE, Piscataway, 2005), pp. 878–885
133. M Andriluka, S Roth, B Schiele, People-tracking-by-detection and people-detection-by-tracking, in *IEEE Conference on Computer Vision and Pattern Recognition, 2008* (IEEE, Piscataway, 2008), pp. 1–8
134. A Yilmaz, O Javed, M Shah, Object tracking: a survey. *ACM Comput. Surv.* **38**, 4 (2006)
135. W-L Lu, J-A Ting, KP Murphy, JJ Little, Identifying players in broadcast sports videos using conditional random fields, in *IEEE Conference on Computer Vision and Pattern Recognition, 2008* (IEEE, Piscataway, 2008), pp. 3249–3256
136. S-W Sun, W-H Cheng, Y-L Hung, I Fan, C Liu, J Hung, C-K Lin, H-Y Mark Liao, Who's who in a sports video? An individual level sports video indexing system, in *2012 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, Piscataway, 2012), pp. 937–942
137. W-L Lu, J-A Ting, JJ Little, KP Murphy, Learning to track and identify players from broadcast sports videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(7), 1704–1716 (2013)
138. G Huang, Y Wang, Gender classification based on fusion of multi-view gait sequences, in *8th Asian Conference on Computer Vision (ACCV'07)* (Springer, Heidelberg, 2007), pp. 462–471
139. D Zhang, Y Wang, Investigating the separability of features from different views for gait based gender classification, in *19th IEEE International Conference on Pattern Recognition* (IEEE, Piscataway, 2008), pp. 1–4
140. X Li, SJ Maybank, S Yan, Gait components and their application to gender recognition. *IEEE Trans. Syst. Man Cybern.* **38**(2), 145–154 (2008)
141. L Cao, M Dikmen, Y Fu, TS Huang, Gender recognition from body, in *16th ACM International Conference on Multimedia (MM '08)* (ACM, Yew York, 2008), pp. 725–728
142. M Hu, Y Wang, Z Zhang, D Zhang, Gait-based gender classification using mixed conditional random field. *Syst. Man Cybern Part B Cybern. IEEE Trans.* **41**(5), 1429–1439 (2011)
143. D-Y Chen, P-C Hsieh, Face-based gender recognition using compressive sensing, in *IEEE International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)* (IEEE, Piscataway, 2012), pp. 157–161
144. P Viola, M Jones, D Snow, Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vis.* **63**(2), 153–161 (2005)
145. D Gerónimo, A Sappa, A López, D Ponsa, Adaptive image sampling and windows classification for on-board pedestrian detection, in *Fifth International Conference on Computer Vision Systems* (Bielefeld University, Bielefeld, 2007)
146. D-Y Chen, Orientation filter enhanced pedestrian detection. *Electron. Lett.* **46**(20), 1377–1379 (2010)
147. AH Nasution, S Emmanuel, Intelligent video surveillance for monitoring elderly in home environments, in *IEEE 9th Workshop on Multimedia Signal Processing (MMSP 2007)* (IEEE, Piscataway, 2007), pp. 203–206
148. N Thome, S Miguët, A real-time, multiview fall detection system: a LHMM-based approach. *TCSVT* **18**(11), 1522–1532 (2008)
149. A Ta, C Wolf, G Lavoue, A Baskurt, J Jolion, Pairwise features for human action recognition, in *20th International Conference on Pattern Recognition (ICPR 2010)* (IEEE, Piscataway, 2010), pp. 3224–3227

doi:10.1186/1687-6180-2013-176

Cite this article as: Paul et al.: Human detection in surveillance videos and its applications - a review. *EURASIP Journal on Advances in Signal Processing* 2013 2013:176.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com
