

A Literature Review on Video Analytics of Crowded Scenes

Myo Thida, Yoke Leng Yong, Pau Climent-Pérez, How-lung Eng,
and Paolo Remagnino

Abstract This chapter presents a review and systematic comparison of the state of the art on crowd video analysis. The rationale of our review is justified by a recent increase in intelligent video surveillance algorithms capable of analysing automatically visual streams of very crowded and cluttered scenes, such as those of airport concourses, railway stations, shopping malls and the like. Since the safety and security of potentially very crowded public spaces have become a priority, computer vision researchers have focused their research on intelligent solutions. The aim of this chapter is to propose a critical review of existing literature pertaining to the automatic analysis of complex and crowded scenes. The literature is divided into two broad categories: the macroscopic and the microscopic modelling approach. The effort is meant to provide a reference point for all computer vision practitioners currently working on crowd analysis. We discuss the merits and weaknesses of various approaches for each topic and provide a recommendation on how existing methods can be improved.

M. Thida (✉) · H.-l. Eng
ZWECC Analytics, Singapore, Singapore
e-mail: mthida81@gmail.com

H.-l. Eng
e-mail: howlungeng@zwecc.com

Y.L. Yong · P. Climent-Pérez · P. Remagnino
Kingston University, London, UK

Y.L. Yong
e-mail: J.Yong@kingston.ac.uk

P. Climent-Pérez
e-mail: P.Climent@kingston.ac.uk

P. Remagnino
e-mail: P.Remagnino@kingston.ac.uk

1 Introduction

Automated video content analysis of a crowded scene has been an active research area in the field of computer vision in the last few years. This strong interest is driven by the increased demand for public safety at crowded spaces such as airports, train stations, malls, stadiums, etc. In such scenes, conventional computer vision techniques for video surveillance cannot be directly applied in the crowded scene due to large variations of crowd densities, complex crowd dynamics and severe occlusions in the scene.

Algorithms for people detection, tracking and activity analysis which consider an individual in isolation (i.e., individual object segmentation and tracking) often face difficult situations such as the overlapping of pedestrians, complex events due to interactions among pedestrians in a crowd. For this reason, many papers consider the crowd as a single entity and analyse its dynamics. The status of crowd is updated as normal or abnormal based on the dynamics of the whole crowd. However, a crowded condition can also be unstructured where pedestrians are relatively free to move in many directions as opposed to a structured crowd where each individual moves coherently in one common direction. In an unstructured crowded scene, considering the crowd as one entity will fail to identify abnormal events which arise due to an inappropriate action of an individual in a crowd. For instance, a running person in a crowd can indicate an abnormal event if the rest of crowd are walking. Thus, considering the crowd as one entity can cause false detections.

Many paper works on modelling crowded scenes to identify different crowd events and/or to detect abnormal events. However, the definition of abnormal event or event of interest has been causing much confusion in the literature due to its subjective nature. Some researchers consider a rare and outstanding event as abnormal while some consider events that have not been observed are abnormal. The problem becomes more challenging as the density of people increases. As a result, more computer vision algorithms are being explored recently.

Despite the great interest and a large number of methods developed, there is a lack of a comprehensive review on crowd video analysis. As shown in Table 1, most current surveys focus on general human motion analysis [1, 5, 24, 75] of single or a small group of people, rather than addressing a crowded scenario. The survey paper by Zhan et al. [83], to the best of our knowledge, is the only one focusing on crowd video analysis. Zhan et al. reviewed some crowd density estimators and crowd modelling techniques, focusing on pedestrian detections, and tracking in a cluttered scene. However, they did not discuss the topic of crowd behaviour understanding and abnormality detection which is covered in this survey. We also present some advances on crowd motion modelling and multi-target tracking in a crowded scene which are not covered in the previous survey.

The goal of this survey is to review and organise the state-of-the-art methods in the domain of crowd video analysis such that their main focus becomes apparent. To achieve this, we have divided the research on crowd video analysis into three broad categories: macroscopic modelling, microscopic modelling and crowd event detection. The methods related to each task are further divided into sub-categories and a comprehensive description of representative methods is provided. In addi-

Table 1 A comparison of this chapter and previous surveys on human motion analysis and crowd video analysis

Year	Authors	Focus	Scenarios
1999	Aggarwal and Cai [1]	Motion analysis, tracking and recognising human activities	Individual or group
2003	Wang et al. [75]	Human detection, tracking and behaviour understanding	Individual or group
2004	Hu et al. [24]	Motion detection, tracking and behaviour understanding	Individual or group
2006	Yilmaz et al. [5]	Object tracking	Individual or group
2008	Zhan et al. [83]	Crowd information extraction and crowd modelling	Crowd
2013	This chapter	Macroscopic modelling, micro-scopic modelling, crowd event detection	Crowd

tion, we identify challenges and future directions for analysing a crowded scene. We believe this will help readers, especially newcomers to this area, to understand the major tasks of a crowded scene analysis system and hope to motivate for the development of new methods.

2 Macroscopic Modelling

In order to learn the typical motion patterns in a crowded scene, macroscopic observation-based methods utilise holistic properties of the scene such as motions in local spatio-temporal cuboid or instantaneous motion are utilised. It is also the preferred method in tracking and analysing the behaviour of both sparse and dense crowd using the following properties such as: density, velocity and flow [31]. Figure 1 depicts detailed various features available for use in macroscopic modelling and the techniques initialising those features.

2.1 Optical Flow Feature

Optical flow is a dense field of instantaneous velocities computed between two consecutive frames commonly used in extracting motion features [23]. Given a video of a crowded scene, the first step is to segment the input video into smaller video clips and compute pixel-wise optical flow between consecutive frames of each clip using the techniques in [11, 23, 49]. The extracted flow vectors may contain noise and redundant information. In order to reduce the computational cost and remove noise, researchers utilise unsupervised (Andrade et al. [6, 7] and Yang et al. [81]) or supervised (Hu and Shah [26, 27]) dimensional reduction techniques. Subsequently, the next step is to find the representative motion patterns of the scene by merging flow vectors from all video frames. Referring back to Fig. 1, it can be seen that

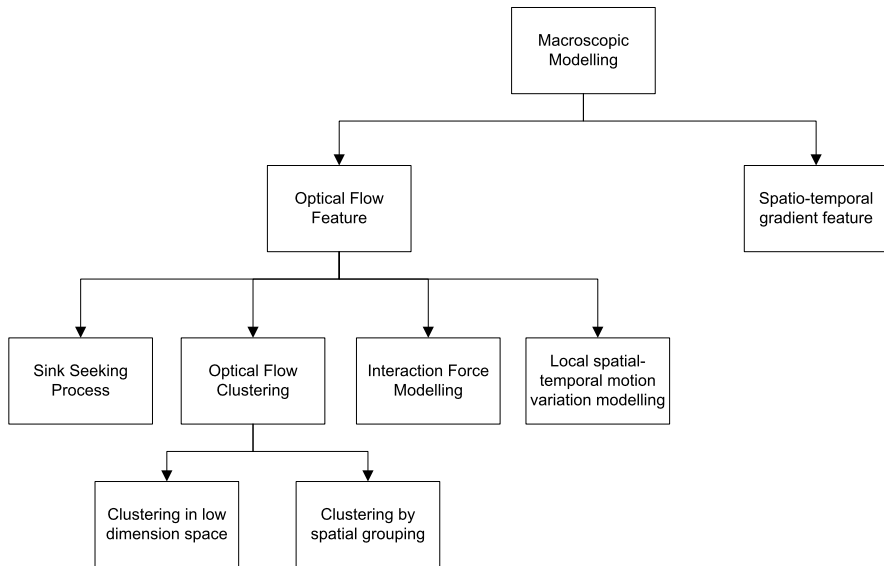


Fig. 1 A schematic illustration of the topics involved in macroscopic crowd video analysis

the motion features extracted from the optical flow can be utilised for motion pattern extraction such as: Sink Seeking Process, Optical Flow Clustering, Interaction Force Modelling, Local Spatio-temporal Motion Variation Modelling, and Spatio-temporal Gradient feature whereby the methods can be used separately or integrated with one another to obtain the desired crowd analysis.

2.1.1 Sink Seeking Process

In the sink seeking process, a grid of particles is overlaid on the first frame of the video clip and advected using a numerical scheme. The path taken by a particle to its final position is called a sink path and thus, the process of finding sinks (exits) and sink paths is called a sink seeking process. Hu and Shah [26, 27] carry out sink seeking process for each particle and thus generate one sink path per particle. These sinks and sink paths are later clustered to extract the dominant motion paths of the scene using an iterative clustering algorithm. On the other hand, Ali and Shah [3] generate a static floor field where each particle holds a value that represents the minimum distance to the nearest sink from its current location. Ali and Shah impose the static floor field together with dynamic and boundary floor field as constraints for tracking algorithm [4].

2.1.2 Optical Flow Clustering

Andrade et al. [6, 7] model the principal components of the optical flow vectors in each video clip using Hidden Markov Models. Then, video segments which have

similar motion pattern are grouped together using the spectral clustering method. The resulting clustered video segments are modelled using a chain of HMMs to represent the typical motion pattern of the scene. The emergency events in the monitored scene are detected by finding deviations from the obtained model.

Instead of the spatial segmentation of each video frame, the other approach is to cluster optical flow vectors by spatial grouping as in [64]. Imran et al. [64] proposed to cluster optical flow vectors in each video clip into N Gaussian mixture components. Then, these Gaussian components are linked over time using a fully connected graph. The connected component analysis of the graph is performed to discover different motion patterns. However, their method still faces the problem of having to determine how many components should there be in the mixture.

2.1.3 Interaction Force Modelling

In addition to learning dominant motion patterns, the optical flow vectors obtained can also be used to model interaction forces of a crowd, and then use the model to detect the stability of the crowd. For example, Mehran et al. [53] employ the optical flow vectors to model pedestrian motion dynamics using a social force model. Social force models [22] have been used in many studies in computer graphic fields for creating animations of the crowd [54]. In this model, the motions of pedestrians are modelled with two forces: a personal desire force and an interaction force. The interaction force is defined as an attractive and repulsive force between pedestrians. In [53], an interaction force between pedestrians is estimated based on optical flow computed over a grid of particles. The normal pattern of this force is later used to model the dynamics of a crowded scene and detect abnormal behaviours in crowds.

2.1.4 Local Spatio-Temporal Motion Variation Modelling

Optical flow data can also be used in modelling the variations of motions in local spatio-temporal volumes to describe the typical motion patterns of the scene [40–42, 50, 52, 79, 81]. In these approaches, an image space is usually divided into cells of a specific size (e.g., 10×10 in [81]) or cuboids (e.g., $30 \times 30 \times 20$ in [42]). Then, optical flow computed in each cell is quantised into different directions. For instance, Yang et al. [81], considered each quantised direction of a given location as a word and cluster these video words into different clusters using a diffusion embedding method. Each node in the graph corresponds to a word and the clusters extracted in the embedded space represent the typical motion patterns of the scenes. Kim and Grauman [40] used a space-time Markov Random Field (MRF) graph to detect abnormal activities in video. Each node in the graph corresponds to a local region in the video frames where the local motion is modelled using a mixture of probabilistic principle component analysis. Wu et al. [79] used Lagrangian framework to extract particle trajectories. These particle trajectories are later used for the modelling of regular crowd motion. The deviations of new motion from the learnt model indicates an abnormal event.

2.2 Spatio-Temporal Gradient Feature

In addition to optical flow information, other features such as spatio-temporal gradient are also used to model the regular movement of a crowd [42, 50]. In [42], the coupled HMM is trained based on the distribution of spatio-temporal motions to detect localised abnormalities in densely crowded scenes. Vijay et al. [52] combined motion information and appearance features to represent the local properties of a scene. The normality of a crowded scene is learned using a mixture of dynamic textures. Then, temporal and spatial abnormalities are separately detected by finding deviations from the normal pattern. Their method has been proved to achieve the better performance than the state-of-the-art methods, at a high computational cost. To address this limitation, Reddy et al. [61] proposed a simpler method using a set of similar features including shape, size and texture extracted from foreground pixels. The computational cost is reduced by removing background noise and considering each feature type individually. Compared to [52], the method proposed by Reddy et al. [61] achieved considerably better results.

2.3 Summary

To conclude the discussion on the macroscopic modelling, a summarisation of the strength and weaknesses of the various state-of-the-art implementation are provided in Table 2.

3 Microscopic Modelling

Microscopic analysis and modelling depends on the analysis of video trajectories of moving entities. This approach, in general, contains the following steps:

1. detection of the moving targets present in the scene;
2. tracking of the detected targets; and
3. analysis of the trajectories to detect dominant flows, and to model typical motion patterns.

Researchers have used different detection and tracking algorithms to generate reliable trajectories. Tracking people in crowds can be either used as a means to improve crowd dynamics analysis, using the tracks and mining trends out of these (*bottom-up* approach to crowd analysis); or, conversely, tracking methods can use cues obtained from the analysis of crowd dynamics, in order to improve accuracy (*top-down* approach). The complexity of tracking algorithms depends on the context and environment in which the tracking is performed. In the context of crowd video analysis, the problem of tracking individuals within a crowd introduces additional

Table 2 Summarisation of the macroscopic modelling techniques

Papers	Advantages	Disadvantages	Data-set and results
[26, 27]	• Instead of using the long term trajectories of the moving objects in learning the typical motion pattern, global motion flow field is used. • Not affected by density of objects within the image. • Does not require complete trajectory, therefore overcoming the problem of occlusion.	• More attention should be placed on learning motion patterns in crowds with less reliable tracking.	• Crowded scene. • Aerial vide from DARPA’s VIVID dataset. • Hong Kong street scene.
[3]	• Lagrangian Coherent Structures (LCS) reveals underlying flow structures that are generally not evident from the raw velocity field of the object.	• If the changes in dynamics is not big enough to be detected, it will not be segmented out within the image.	• Videos from stock footage web sites. • National Geographic documentary footage from “Inside Mecca”.
[4]	• Takes into consideration the crowd flow and scene layout for tracking. • Provides the shortest distance to a sink for each location.	• Tracking errors produced by the Static Floor field (SFF) is inconsistent and dependant on the trajectory of the tracked object and only works well when the object is moving in a straight path. • The Dynamic Floor Field produced error when there’s noise and interference of other objects in the view.	• Marathon sequence from different perspective – Overhead cameras. – View from high rise building.
[53]	• Does not depend on tracked object in the analysis of crowd behaviour. • The incorporation of particle advection assist in capturing the crowd flow.	• Requires a set of goal destination for the scene.	• University of Minnesota dataset.
[40]	• Anomalies could be detected on both the global and local context.	• Posterior probabilities of all previous descriptors are not recalculated, and assumed not to change.	• Surveillance videos from subway station – Entrance Gate. – Exit Gate.

Table 2 (Continued)

Papers	Advantages	Disadvantages	Data-set and results
	• With the information from the global and local context available, anomalies are easily detected. • Ease of implementation.		
[52]	• Works extremely well in detecting anomalies in video scenes compared to other tracking based method such as optical flow.	• System requires training. • Long computation time making it not feasible for real-time system.	• Crowded Scenes.
[81]	• The weight of the Diffusion Map is unique to the application. • By manipulating the diffusion time, Diffusion Map can also be used for multi-scale analysis of the scene.	• Might be more suitable for scene understanding as low level feature have a better manifold structure.	• INGSIM dataset. • Far-field traffic scene.
[61]	• Low-level computation required. • Separate modelling and analysis of motion, size and texture feature make it commutation efficient.	• Requires training. • The density information obtained changes drastically due to the handling of the training data.	• UCSD anomaly detection dataset.

complexity due to the interactions and occlusions between people in the crowd. A number of tracking methods has been proposed to overcome the challenges encountered in a crowded scene. In this section, some popular human tracking methods in the context of crowd video analysis are discussed. The reader is referred to the survey by Yilmaz et al. [5] for a comprehensive review of various trackers. Figure 2 shows the different topics covered by this section.

3.1 The Particle Filter (PF) Framework

The most popular approach for tracking is the *Particle Filter*-based framework. Particle filtering framework was first introduced for visual tracking by Isard and Blake in [29]. Initially, particle filter approaches were only based on colour cues, and could only track one single target.

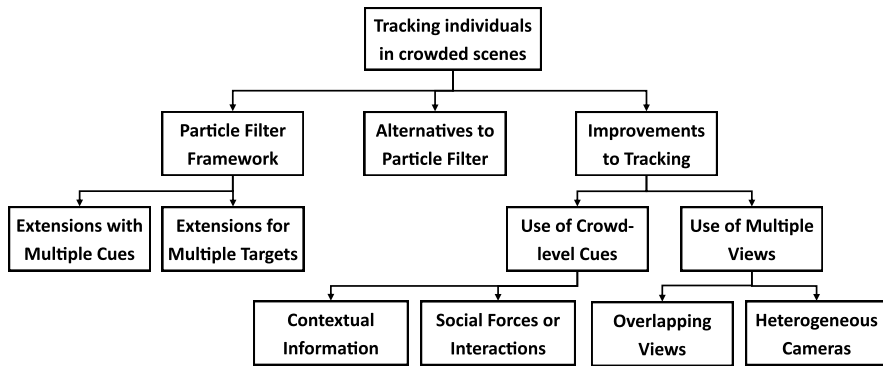


Fig. 2 The topics for microscopic approach, in which a focus is put on individual tracking in crowds

3.1.1 Additional Cues for Improved PF

The particle filter implementation based on appearance using colour information only does not perform well tracking more than one individual, specially when those wear similar clothing. In public demonstrations, sports matches and celebrations, it is normal that people's appearance is similar. Thus, a series of papers present alternatives to the plain 'colour-only' Particle Filter. Combinations include colour and contours, Harris, SIFT features [47, 59, 67, 78]; also Histograms of Oriented Gradients (HOGs) are used along with colour information in [69]; or Mean Shift and Joint Probabilities [10].

A completely different approach to improve tracking using particle filters is presented in [84]. The method proposed by the authors mines the interdependencies between particles in order to improve the results. Also different is the method proposed in [28], in which a new tracker is proposed which employs a particle filter tracking framework, where the state transition model is estimated by an optical-flow algorithm. That is, instead of using a pre-defined dynamic transition model.

There are also authors whose interest is in extending the particle filter to multiple cameras; in that case, particles are "shared" and "fused" among the views [57].

Others propose blob-based segmentation and tracking when no occlusions are present, and limit the use of Particle Filters as an occlusion resolution technique [70, 86]. The limitation of this techniques seems clear: blobs are needed and used as the main cue, which is not the case in most crowded scenes, although these techniques can be useful in sparse crowds.

Silhouettes or contours can be a useful cue for action recognition, or people counting in crowds; obviously, in the case of densely crowded scenes, only partial contours can be extracted, although those can be quite useful (e.g., as in ' Ω shape'-based methods). Since particle filter approaches work regardless of segmentation, reconstructing contours *a posteriori* to obtain shape cues might be of interest. Ma et al. [51] present this idea: Graph Cuts are applied to a particle filter method to obtain the silhouettes of tracked objects.

3.1.2 Alternative Cues for Tracking: Self-similarity

Schechtman and Irani [66] introduced the concept of self-similarity as a visual feature. Among the applications, they name object detection and recognition, action recognition, texture classification, data retrieval, tracking and image alignment and so on. BenAbdelkader et al. [12] seem to be the first to use image self-similarity plots (ISSPs) for gait recognition; according to the authors, some works state the ISSP of a moving person/object is a projection of its planar dynamics, and as such, these should encode much of gait information. Junejo et al. [34, 35] use a very similar descriptor as a means for action recognition, by using self-similarity matrices (SSMs) as descriptors of the action class. Dexter et al. [17] extend the SSM concept in order to apply it to the synchronisation of actions taken from multiple views. Rani and Arumugam [60] use it as a biometric signature in gait recognition as in [12]. Also, Walk et al. [74] introduced the self-similarity as a feature for pedestrian detection; and Cai et al. [14] have used it for person re-identification among different cameras or moments; the authors create a colour codebook and obtain the spatial occurrence distributions of colour self-similarities. To the best of our knowledge, as of today, no works seem to use self-similarities as a feature for tracking, although Gu et al. state it could be used as an alternative to other local descriptors such as SIFT or SURF.

3.1.3 Multiple Target Tracking Using PF

This framework has been extended in a series of papers [2, 15, 20, 39, 58] for tracking multiple targets. For example, Okuma et al. [58] extend a particle framework by incorporating a cascaded AdaBoost algorithm for the detection and tracking of multiple hockey players in a video. The AdaBoost algorithm is used to generate detection hypotheses of hockey players. Once the detection hypotheses are available, each hockey player is modelled with an individual particle filter that forms a component of a mixture particle filter. Similarly, Ali and Dailey [2] combine an ‘AdaBoost cascade classifier’-based head detection algorithm and the particle filtering method for tracking multiple persons in high density crowds. The performance is further improved by a confirmation-by-classification method to estimate confidence in a tracked trajectory.

To conclude this subsection, a summarisation of the presented methods is shown in Table 3. Both single and multiple view methods are presented, as well as single and multiple target ones.

3.2 Handling Occlusions

Occlusions are one of the most important problems trackers need to face, since generalised models for them are not straightforward [44]. According to the survey in [82], occlusion can be classified into three categories: self-occlusion, which occurs while tracking articulated objects; inter-object occlusion (or dynamic occlu-

Table 3 Summarisation of the presented techniques

Main tracking	Additional information or method	Multi-view/-target	Works/papers
Particle Filter	Contour information	No	[47]
Particle Filter	SIFT, Harris-SIFT	No	[59, 67, 78]
Particle Filter	Histogram of Oriented Gradients (HOG)	No	[69]
Particle Filter	Mean Shift/Joint Probabilities	No	[10]
Particle Filter	None: Changes in the transition model	No	[28, 84]
Particle Filter	None: Particles are fused among views	Multi-view	[57]
Blob tracking	Particle filter for occlusion handling	Multi-target	[70, 86]
Particle Filter	None: GCs ^a used to recover contours	No	[51]
Particle Filter	AdaBoost, Cascaded AdaBoost	Multi-target	[2, 58]
Particle Filter	MRF ^b , MCMC ^c	Multi-target	[39]
Particle Filter	NN Data Association, Mean-shift	Multi-target	[15]
Mean Shift/Kalman	Viterbi-style <i>tracklet</i> merge	Multi-target	[20]

^aGraph Cuts^bMarkov Random Field^cMarkov Chain Monte Carlo

sion [72]), which arises when two tracked objects occlude each other; and occlusion by the background (or scene occlusion [72]), which occurs when structures in the scene (e.g., tree branches, pillars, etc.) occlude the object/s being tracked. Some approaches have already been presented in Sect. 3.1.1 [70, 86]. Yilmaz et al. [82] deal with occlusion handling from the lens of the tracking technique in use. A series of different tracker families are presented (point, ‘geometric model’-based and silhouette); each tracking technique is then classified according to whether or not it can handle occlusions, and in the case it does, whether these can be full or only partial. Following this idea, trackers that respond well when occlusions are present, can be used for occlusion handling. In [85], the Kanade-Lucas-Tomasi (KLT) tracker is employed to resolve occlusions, while a particle filter is used as the main tracker. Similarly, a technique based on Mean-shift is used in [16].

Apart of exploiting the features of “occlusion-friendly” trackers, a series of occlusion handling techniques have also been devised, which can be found throughout the literature. Wang et al. [77], present a good historical review of such methods, which rely on the object’s motion model, and keep predicting the object’s location until it reappears. The authors state that serious long-term occlusions cannot be dealt with by this kind of techniques, since observations cannot be obtained while the object is occluded for a long period of time. Vezzani et al. [72] propose what they call the *non-visible regions model*, which deals with partial and full occlusions, whether these are inter-object or due to the scene. The object model is updated differently in a pixel-wise fashion: the appearance is updated only for the visible pixels; the probabilities associated to those are reinforced, while they remain unchanged for invisible pixels. Furthermore, in pixels with no correspondence due to changes in

the shape of the object (called appearance occlusions) probabilities are smoothed. Wang et al. [77], on the other hand, propose a means of modelling the occluder; once modelled, when objects disappear due to occlusion, a search is performed around the occluder in order to find the occluded object as it reappears.

In [37], the authors present a series of monocular approaches to occlusion handling, although this is only to conclude that single-view systems are intrinsically unable to handle occlusions correctly. The authors in [21, 37], use multiple oblique-view cameras to handle occlusions appropriately, and devise a common plane reconstruction, using communication among cameras. Approaches based on multiple views are designed to reduce the amount of hidden regions. Unfortunately, in the case of existing static camera networks, this is not always possible due to the restrictions of their infrastructures, which were not initially devised for automated surveillance. Another approach to occlusion handling is avoiding them in the first place. Occlusions can be reduced by placing the camera appropriately, as suggested by [82] (e.g., by placing a bird-eye view camera, no occlusions occur between the objects on the ground), but the problem of existing infrastructures persists.

Nevertheless, when dealing with occlusions under heavily crowded scenarios, full-body tracking is infeasible due to the continuous existence of partial occlusions, specially from side views [13]. Since the existing cameras tend to be placed above the heads of the people and tilted to face downwards looking at the scene, some authors suggest a good assumption is that heads and shoulders (often referred to as Omega-shape [46]) will be always visible, and that occlusions among subjects' heads is lower as compared to the rest of the body parts.

3.3 Improving Tracking Using Crowd-Level Cues

As stated in the introduction to this section (Sect. 3), tracking methods can use cues obtained from the analysis of crowd dynamics, in order to improve their accuracy, in a *top-down* approach. These higher-level cues can be either contextual or coming from the social interactions among the people in the crowd.

3.3.1 Higher-Level Contextual Information

The utility of high-level contextual information has demonstrated that exploiting contextual information improves the performance of human tracking significantly. Antonini et al. [9] use a discrete choice model (DCM) as motion priors to predict human motion patterns and then, fuse this model in a human tracker for improved performance. Similarly, Ali et al. [4] propose to exploit contextual information for tracking multiple people in a structured crowded scene. Assuming that all participants of the crowd are moving in one direction, Ali et al. learn the direction of motion as a prior information based on floor fields. The authors have demonstrated that a higher-level constraint greatly increases the performance of the tracker. However,

floor fields can be learned only when the scene has one dominant motion. As a result, the method proposed in [4] cannot be applied for unstructured crowded scenes where the motion of a crowd appears to be random with different participants moving in different directions over time. Some examples of unstructured crowded scenes include crowds at exhibitions, sporting events and railway stations. This shortcoming is addressed by Mikel et al. [63] where the authors employ a correlated topic model for modelling random motions in an unstructured crowded scene. Similarly, L. Kratz and K. Nishino [43] employ the normal motion pattern to predict tracking individuals in a crowd scene where the normal motion pattern is learnt based on local motion at fixed-size cells.

3.3.2 Social Interactions

Another interesting direction of tracking multiple targets is to integrate social interaction of targets in the tracking algorithm. This idea is motivated by the behaviour of targets in a crowd. In crowded scenarios, the behaviour of each individual target is influenced by the proximity and behaviour of other targets in the crowd. Several methods [8, 19, 39, 48, 80] have proposed to integrate the social interactions among targets in the tracking algorithms. This direction has shown promising performance to track multiple targets in crowded scenes. An early example which models the social interaction of targets is *Markov Chain Monte Carlo*-based (MCMC) particle filter [39]. Their method models social interactions of targets using Markov Random Field and adds motion prior in a joint particle filter. The traditional importance sampling step in the particle filter is replaced by a MCMC sampling step. French et al. [19] extended the method in [39] by adding social information to compute the velocity of particles. In [80], the authors formulated the tracking problem as a problem of minimising an energy function. The energy function is defined based on both the social information and physical constraint in the environment. Their preliminary results indicate that social information provides an important cue for tracking multiple targets in a complex scene. An overview of tracking algorithms that incorporate different high-level contextual information is illustrated in Fig. 3.

3.4 Tracking in Crowds from Multiple Views

Researchers have also explored the use of multiple cameras for tracking people under severe occlusion in a complex environment. Multiple camera tracking methods intend to expand the monitored area and provide complete information about interesting persons by gathering evidences from different camera views. Lee et al. [45] propose a multiple people tracking method for wide-area monitoring. An automated calibration method is introduced to find correspondences between distributed cameras. In their method, all camera views are calibrated to a global ground-plane view

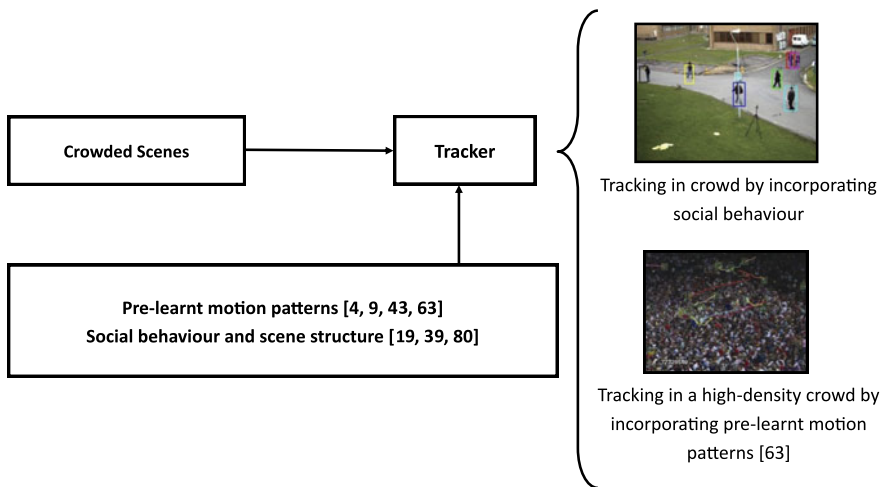


Fig. 3 An overview of different tracking algorithms that incorporate high-level contextual information

based on geometric constraints and tracking trajectories from each view. Another example in a similar context can be found in the papers by Khan and Shah [36, 38]. A planar homographic occupancy constraint that combines foreground likelihood information from different views is proposed for detection and occlusion resolution.

Another use of multiple cameras is to track people in an environment covered by multiple cameras with overlapping views. Mittal and Davis [55] use pairs of stereo cameras and combine evidences gathered from multiple cameras for tracking people in a cluttered scene. Foreground regions from different camera views are projected back into a 3D space so that the endpoints of the matched regions yield 3D points belonging to people. Dockastader and Tekalp [18] employ a Bayesian network for fusing 2D position information acquired from different camera views to estimate the 3D coordinate position of the interested person. Finally, a layer of Kalman filtering is used to update the position of people. A combination of static and pan-tilt-zoom (PTZ) cameras for multiple camera tracking is introduced in [65]. The static cameras are used to provide a global view of the interested persons when the PTZ cameras are used for face recognition of people.

The brief overview of the research literature indicates that multiple camera tracking methods provide an interesting mechanism to handle severe occlusion and to monitor large areas at public spaces (as seen in Sect. 3.2). However, advantages of the multiple cameras come together with additional issues such as camera calibration, matching information across the camera views, automated camera switching and data fusion. These challenges are still yet to be solved. On the other hand, integrating the social interaction among targets in the tracking algorithms has shown promising performance to track individual targets in a crowd.

4 Event Detection in Crowds

A series of surveys and reviews in this field [30, 32, 56, 62, 68, 73, 83] show there is a great interest in this area. Detecting anomalies or outstanding events in crowds has moved a lot of research efforts. Automatic systems would allow reducing the burden of manual video supervision, which makes is infeasible in most cases, given the enormous amounts of data, as compared to the manpower to process it [68, 73, 83]. Detection of anomalies in crowded scenes can be seen as a classification problem where only two classes are defined (i.e., “normal” versus “anomalous”) [68].

The survey by Sodemann et al. [68], analyses the works in the literature across five aspects:

1. the target/s of interest (a person, a crowd);
2. the definitions of what is anomalous, and the assumptions taken;
3. the types of sensors involved, and the features used;
4. the learning methods; and
5. the modelling algorithms.

According to the authors, their survey is focused on the broader problem formulation and assumptions, rather than providing a review on specific pattern classification methods. In Revathi and Kumar [62], authors provide a categorisation of anomalies according to the number of people and other objects involved. Three categories are defined: anomalies involving a single person with a single object, multiple people with multiple objects, and group behaviour.

Vishwakarma and Agrawal [73] analyse human action recognition in more general terms in video surveillance, although, they present an interesting taxonomy to classify complexity of topic-related algorithms. From a completely different point of view, two other works review physics- and hydrodynamics-based techniques [32, 56] for anomaly and event detection in crowds. Moore et al. [56] present a review of techniques for crowd analysis that consider huge crowds as “fluids” or “liquids”, which are bound to a series of rules and forces (e.g., repulsion and attraction) which explain the interactions among the particles that conform that fluid.

Jo et al. [32] further explore other physics-based techniques, and classify the works according to the categorisation presented in [30], which presents various “domains”: the image space domain, based on the analysis at the pixel, texture or object levels; the sociological domain, which accounts for the social interactions or “crowd mentality”; the level of services, where different crowd conditions are provided; or the computer graphics domain, which deals with realistic crowd simulation. Jacques Junior et al. [30] also classify crowd event detection techniques as either object-based, in which individuals are tracked and these tracks are used to analyse the situations [25, 33, 76]; or holistic-based [7, 53, 71], where the crowd is considered as a whole, and events are detected by extracting the major crowd flows from the monitored scene.

5 Summary

This chapter presents a review and comparative study of various topics in the area of crowd video analysis. The advantages and disadvantages of the state-of-the-art methods related to video analytics in crowded scenes have been detailed.

Tracking individuals in a high-density crowd has been addressed in recent years, as opposed to previously tracking individuals in sparse or even *ad-hoc* scenarios. A major advance is the introduction of high-level crowd motion pattern as a prior into a general framework [4, 63]. However, the problem of tracking still remains as a challenging problem in the area of computer vision. One major challenge for tracking in a crowded scene is inter-object occlusion due to the interactions of participants in a crowd. There remains a gap between the state-of-the-art and robust tracking of people in a crowded scene. Most recent trackers for crowds use Particle Filters, using different kinds of features; the use of self-similarity measures for this particular application can be of interest and deserves further research, given the results it achieved in other Computer Vision fields.

During recent years there has been substantial progress towards understanding crowd behaviour and abnormality detection based on modelling crowd motion pattern. However, these approaches capture general movement of a crowd but do not accurately detect details of individual movements. As a result, the current literature in understanding crowd motion is not ready to capture the motion pattern of an unstructured crowd scene where the motion of the crowd appears to be random [63]. Future research in this area requires localised modelling of crowd motion to capture different behaviours in unstructured crowded scenes. On the other hand, the understanding and modelling of crowd behaviour remains immature despite the considerable advances in human activity analysis. Progress in this area requires further advances in modelling or representation of a crowd event and recognition of these events in a natural environment.

References

1. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. *Comput. Vis. Image Underst.* **73**(3), 428–440 (1999)
2. Ali, I., Dailey, M.N.: Multiple human tracking in high-density crowds. In: *Advanced Concepts in Intelligent Vision Systems*, pp. 540–549 (2009)
3. Ali, S., Shah, M.: A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Minneapolis, Florida, pp. 1–6. IEEE, New York (2007)
4. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: *Proceedings of European Conference on Computer Vision*, Marseille, France, pp. 1–14. Springer, Berlin (2008)
5. Alper, Y., Omar, J., Mubarak, S.: Object tracking: a survey. *ACM Comput. Surv.* **38**(4), 13–58 (2006)
6. Andrade, E., Fisher, R.: Simulation of crowd problems for computer vision. In: *Proceedings of 19th International Conference on Pattern Recognition*, vol. 3, pp. 71–80 (2005)

7. Andrade, E., Fisher, R., Blunsden, S.: Modelling crowd scenes for event detection. In: Proceedings of 19th International Conference on Pattern Recognition, vol. 1, pp. 175–178 (2006)
8. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, pp. 1265–1272. IEEE, New York (2011)
9. Antonini, G., Martinez, S.V., Bierlaire, M., Thiran, J.P.: Behavioral priors for detection and tracking of pedestrians in video sequences. *Int. J. Comput. Vis.* **69**(2), 159–180 (1998)
10. Bai, K.: Particle filter tracking with mean shift and joint probability data association. In: 2010 International Conference on Image Analysis and Signal Processing (IASP), pp. 607–612. IEEE, New York (2010)
11. Barron, J., Fleet, D.J., Beauchemin, S.: Performance of optical flow techniques. *Int. J. Comput. Vis.* **12**(1), 43–77 (1994)
12. BenAbdelkader, C., Cutler, R., Nanda, H., Davis, L.: Eigengait: motion-based recognition of people using image self-similarity. Technical report (2001)
13. Boltes, M., Seyfried, A.: Collecting pedestrian trajectories. *Neurocomputing* **100**, 127–133 (2013)
14. Cai, Y., Pietikäinen, M.: Person re-identification based on global color context. In: Asian Conference on Computer Vision 2010 Workshops (2011)
15. Cai, Y., de Freitas, N., Little, J.J.: Robust visual tracking for multiple targets. In: Proceedings of Eighth European Conference on Computer Vision, vol. 3954, pp. 107–118. IEEE, New York (2006)
16. Chen, A.h., Yang, B.q., Chen, Z.g.: A timely occlusion detection based on mean shift algorithm. In: Deng, W. (ed.) *Future Control and Automation. Lecture Notes in Electrical Engineering*, vol. 173, pp. 51–56. Springer, Berlin (2012)
17. Dexter, E., Pérez, P., Laptev, I.: Multi-view synchronization of human actions and dynamic scenes. In: Proceedings of the British Machine Vision Conference 2009, British Machine Vision Association, pp. 122.1–122.11 (2009)
18. Dockstader, S.L., Tekalp, A.M.: Multiple camera tracking of interacting and occluded human motion. *Proc. IEEE* **89**(10), 1441–1455 (2001)
19. French, A., Naeem, A., Dryden, I., Pridmore, T.: Using social effects to guide tracking in complex scenes. In: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 212–217 (2007)
20. Gilbert, A., Bowden, R.: Multi person tracking within crowded scenes. In: Proceedings of Workshop on Human Motion, pp. 166–179 (2007)
21. Haselhoff, A., Hoehmann, L., Nunn, C., Meuter, M., Kummert, A.: On occlusion-handling for people detection fusion in multi-camera networks. In: Dziech, A., Czyżewski, A. (eds.) *Multimedia Communications, Services and Security. Communications in Computer and Information Science*, vol. 149, pp. 113–119. Springer, Berlin (2011)
22. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. *Phys. Rev. E* **51**(5), 4282–4286 (1995)
23. Horn, B.K.P., Schunck, B.G.: Determining optical flow. *Artif. Intell.* **17**, 185–203 (1981)
24. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev.* **34**(3), 334–352 (2004)
25. Hu, W., Xiao, X., Fu, Z., Dan, X., Tan, T., Steve, M.: A system for learning statistical motion patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(9), 1450–1464 (2006)
26. Hu, M., Ali, S., Shah, M.: Detecting global motion patterns in complex videos. In: Proceedings of International Conference on Pattern Recognition, Tempa, Florida, pp. 1–5. IEEE, New York (2008)
27. Hu, M., Ali, S., Shah, M.: Learning motion patterns in crowded scenes using motion flow field. In: Proceedings of International Conference on Pattern Recognition, Tempa, Florida, pp. 1–5. IEEE, New York (2008)
28. Hu, N., Bouma, H., Worring, M.: Tracking individuals in surveillance video of a high-density crowd. In: Proceedings of SPIE, vol. 8399, p. 839909 (2012)

29. Isard, M., Blake, A.: CONDENSATION conditional density propagation for visual tracking. *Int. J. Comput. Vis.* **29**(1), 5–28 (1998)
30. Jacques Junior, J.C.S., Musse, S.R., Jung, C.R.: Crowd analysis using computer vision techniques: a survey. *IEEE Signal Process. Mag.* (September), 66–77 (2010)
31. Jiang, Y., Zhang, P., Wong, S., Liu, R.: A higher-order macroscopic model for pedestrian flows. *Phys. A, Stat. Mech. Appl.* **389**(21), 4623–4635 (2010)
32. Jo, H., Chug, K., Sethi, R.J., Rey, M.: A review of physics-based methods for group and crowd analysis in computer vision. *J. Postdr. Res.* **1**(1), 4–7 (2013)
33. Johnson, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. *Image Vis. Comput.* **14**(8), 583–592 (1996)
34. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: Cross-view action recognition from temporal self-similarities. In: *Proceedings of the European Conference on Computer Vision 2008* (2008)
35. Junejo, I.N., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(1), 172–185 (2011)
36. Khan, S.M., Shah, M.: A multi-view approach to tracking people in dense crowded scenes using a planar homography constraint. In: *Proceedings of Workshop on Human Motion*, Graz, Austria, pp. 133–146 (2006)
37. Khan, S., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(3), 505–519 (2009)
38. Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(3), 505–519 (2009)
39. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(11), 1805–1918 (2005)
40. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2921–2928 (2009)
41. Kratz, L., Nishino, K.: Spatio-temporal motion pattern modelling of extremely crowded scenes. In: *The 1st International Workshop on Machine Learning for Vision-Based Motion Analysis*, Marseille, France (2008)
42. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Maimi Beach, Florida, pp. 1446–1453 (2009)
43. Kratz, L., Nishino, K.: Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, pp. 693–700 (2010)
44. Kwak, S., Nam, W., Han, B., Han, J.H.: Learning occlusion with likelihoods for visual tracking. In: *2011 IEEE International Conference on Computer Vision (ICCV)*, pp. 1551–1558 (2011)
45. Lee, L., Romano, R., Stein, G.: Monitoring activities from multiple video streams: establishing a common coordinate frame. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(8), 758–767 (2000)
46. Li, M., Zhang, Z., Huang, K., Tan, T.: Rapid and robust human detection and tracking based on omega-shape features. In: *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2545–2548 (2009)
47. Li, J., Lu, X., Ding, L., Lu, H.: Moving target tracking via particle filter based on color and contour features. In: *2010 2nd International Conference on Information Engineering and Computer Science (ICIECS)*, pp. 1–4. IEEE, New York (2010)
48. Luber, M., Stork, J.a., Tipaldi, G.D., Arras, K.O.: People tracking with human motion predictions from social forces. In: *2010 IEEE International Conference on Robotics and Automation*, pp. 464–469. IEEE, New York (2010)
49. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: *Proceedings of Image Understanding Workshop*, pp. 121–130 (1981)
50. Ma, Y., Cisar, P.: Activity representation in crowd. In: *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, Florida, USA, pp. 107–116. Springer, Berlin (2008)

51. Ma, L., Liu, J., Wang, J., Cheng, J., Lu, H.: A improved silhouette tracking approach integrating particle filter with graph cuts. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 1142–1145. IEEE, New York (2010)
52. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 1975–1981 (2010)
53. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, Maimi Beach, Florida, pp. 935–942. IEEE, New York (2009)
54. Michel, B., Gianluca, A., Mats, W.: Behavioural dynamics for pedestrians. In: Lecture Notes in Computer Science, pp. 1–18 (2003)
55. Mittal, A., Davis, L.S.: M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. *Int. J. Comput. Vis.* **51**(3), 189–203 (2003)
56. Moore, B.E., Ali, S., Mehran, R., Shah, M.: Visual crowd surveillance through a hydrodynamics lens. *Commun. ACM* **54**(12), 64–73 (2011)
57. Ni, Z., Sunderrajan, S., Rahimi, A., Manjunath, B.: Distributed particle filter tracking with online multiple instance learning in a camera sensor network. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 37–40. IEEE, New York (2010)
58. Okuma, K., Taleghani, A., Freitas, N.D., Little, J.J., Lowe, D.G.: A boosted particle filter: multitarget detection and tracking. In: Proceedings of Eighth European Conference on Computer Vision, pp. 28–39. IEEE, New York (2004)
59. Qi, Z., Ting, R., Husheng, F., Jinlin, Z.: Particle filter object tracking based on Harris-SIFT feature matching. *Proc. Eng.* **29**, 924–929 (2012)
60. Rani, M., Arumugam, G.: An efficient gait recognition system for human identification using modified ICA. *Int. J. Comput. Sci. Inf. Technol.* **2**(1), 55–67 (2010)
61. Reddy, V., Sanderson, C., Lovell, B.: Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: MLvMA Workshop, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, pp. 57–63. IEEE, New York (2011)
62. Revathi, A., Kumar, D.: A review of human activity recognition and behaviour understanding in video surveillance. *Comput. Sci. Inf. Technol.* **2**, 375–384 (2012)
63. Rodriguez, M., Ali, S., Kanade, T.: Tracking in unstructured crowded scenes. In: Proceedings of IEEE International Conference on Computer Vision, Kyoto, Japan, pp. 1389–1396. IEEE, New York (2009)
64. Saleemi, I., Hartung, L., Shah, M.: Scene understanding by statistical modeling of motion patterns. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 2069–2076 (2010)
65. Scott, S.: A system for tracking and recognizing multiple people with multiple camera. Technical report GIT-GVU-98-25, Georgia Institute of Technology (1998)
66. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE, New York (2007)
67. Shu-hong, C., Chun-hai, H.: Particle filter tracking algorithm based on multi-information fusion. In: 2009 International Conference on Information Engineering and Computer Science, ICIECS 2009, pp. 1–4. IEEE, New York (2009)
68. Sodemann, A.A., Ross, M.P., Borghetti, B.J.: A review of anomaly detection in automated surveillance. *IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev.* **42**(6), 1257–1272 (2012)
69. Sugano, H., Miyamoto, R.: Parallel implementation of pedestrian tracking using multiple cues on GPGPU. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 900–906. IEEE, New York (2009)
70. Tang, S.L., Kadim, Z., Liang, K.M., Lim, M.K.: Hybrid blob and particle filter tracking approach for robust object tracking. *Proc. Comput. Sci.* **1**(1), 2549–2557 (2010)
71. Thida, M., Eng, H.L., Monekosso, D.N., Remagnino, P.: Learning video manifold for segmenting crowd events and abnormality detection. In: Proceedings of 10th Asian Conference on Computer Vision, pp. 439–449. Springer, Berlin (2010)

72. Vezzani, R., Grana, C., Cucchiara, R.: Probabilistic people tracking with appearance models and occlusion classification: the ad-hoc system. *Pattern Recognit. Lett.* **32**(6), 867–877 (2011)
73. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behaviour understanding in video surveillance. *Vis. Comput.* (September) (2012)
74. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1030–1037. IEEE, New York (2010)
75. Wang, L., Hu, W., Tan, T.: Recent developments in human motion analysis. *Pattern Recognit.* **36**(3), 585–601 (2003)
76. Wang, X., Tieu, K., Grimson, E.: Learning semantic scene models by trajectory analysis. In: *Proceedings of European Conference on Computer Vision*, vol. 3, pp. 110–123 (2006)
77. Wang, P., Li, W., Zhu, W., Qiao, H.: Object tracking with serious occlusion based on occluder modeling. In: 2012 International Conference on Mechatronics and Automation (ICMA) pp. 1960–1965 (2012)
78. Wu, P., Kong, L., Zhao, F., Li, X.: Particle filter tracking based on color and SIFT features. In: 2008 International Conference on Audio, Language and Image Processing, pp. 932–937. IEEE, New York (2008)
79. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 2054–2060. IEEE, New York (2010)
80. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who are you with and where are you going? In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, pp. 1345–1352. IEEE, New York (2011)
81. Yang, Y., Liu, J., Shah, M.: Video scene understanding using multi-scale analysis. In: *Proceedings of IEEE International Conference on Computer Vision*, Kyoto, Japan, pp. 1669–1676. IEEE, New York (2009)
82. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. *ACM Comput. Surv.* **38**(4) (2006)
83. Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q.: Crowd analysis: a survey. *Mach. Vis. Appl.* **19**(5–6), 345–357 (2008)
84. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via structured multi-task sparse learning. *Int. J. Comput. Vis.* 1–17 (2012)
85. Zhang, C., Xu, J., Beaugendre, A., Goto, S.: A klt-based approach for occlusion handling in human tracking. In: *Picture Coding Symposium (PCS)*, 2012, pp. 337–340 (2012)
86. Zhong, Q., Qingqing, Z., Tengfei, G.: Moving object tracking based on codebook and particle filter. *Proc. Eng.* **29**, 174–178 (2012)

Intelligent Multimedia Surveillance

Current Trends and Research

Atrey, P.K.; Kankanhalli, M.; Cavallaro, A. (Eds.)

2013, IX, 192 p. 76 illus., 57 illus. in color., Hardcover

ISBN: 978-3-642-41511-1