Security and Surveillance

Shaogang Gong and Chen Change Loy and Tao Xiang

Abstract Human eyes are highly efficient devices for scanning through a large quantity of low-level visual sensory data and delivering selective information to one's brain for high-level semantic interpretation and gaining situational awareness. Over the last few decades, the computer vision community has endeavoured to bring about similar perceptual capabilities to artificial visual sensors. Substantial efforts have been made towards understanding static images of individual objects and the corresponding processes in the human visual system. This endeavour is intensified further by the need for understanding a massive quantity of video data, with the aim to comprehend multiple entities not only within a single image but also over time across multiple video frames for understanding their spatio-temporal relations. A significant application of video analysis and understanding is intelligent surveillance, which aims to interpret automatically human activity and detect unusual events that could pose a threat to public security and safety.

1 Introduction

There has been an accelerated expansion of Closed-Circuit TeleVision (CCTV) surveillance in recent years, largely in response to rising anxieties about crime and its threat to security and safety. Substantial numbers of surveillance cameras have been deployed in public spaces ranging from transport infrastructures (e.g. airports, underground stations), shopping centres, sport arenas to residential streets, serving as a tool for crime reduction and risk management. Conventional visual surveillance

Shaogang Gong
Queen Mary University of London, London E1 4NS, UK, e-mail: sgg@eecs.qmul.ac.uk

Chen Change Loy
Queen Mary University of London, London E1 4NS, UK, e-mail: ccloy@eecs.qmul.ac.uk

Tao Xiang
Queen Mary University of London, London E1 4NS, UK, e-mail: txiang@eecs.qmul.ac.uk

systems rely heavily on human operators to monitor activities and determine the actions to be taken upon the occurrence of an incident, e.g. tracking a suspicious target from one camera to another camera or alerting relevant agencies to areas of concern.

Unfortunately, many actionable incidents are simply miss-detected in such a manual system due to inherent limitations from deploying solely human operators eyeballing CCTV screens. Miss-detections could be caused by (1) excessive number of video screens to monitor, (2) boredom and tiredness due to prolonged monitoring, (3) lack of *a priori* and readily accessible knowledge for what to look for, (4) distraction by additional responsibilities such as other administrative tasks [24]. As a result, surveillance footages are often used merely as passive records or as evidence for post-event investigations. Miss-detections of important events can be perilous in critical surveillance tasks such as border control or airport surveillance. Technology providers and end-users recognise that manual process alone is inadequate to meet the need for screening timely and searching exhaustively colossal amount of video data generated from the growing number of cameras in public spaces. To fulfil such a need, video content analysis paradigm is shifting from a fully human operator model to a machine-assisted and automated model.

In the following, we describe applications and the latest advances in automated visual analysis of human activities for security and surveillance. Section 2 outlines some of the most common technologies in the market, and highlights technical challenges that limit the use and growth of video analytics software. Section 3 discusses state of the art video analytics techniques, which may help in advancing current security and surveillance applications.

2 Current Systems

There is a surge in demand in the last few years for automated video analysis technologies. This trend is persisting[1,2], mainly driven by the government initiatives and strong demands from retail and transportation sectors[3]. Increasing number of CCTV solutions are made available with some degree of automated analytic capabilities by suppliers from large-scale system integrators to small and medium enterprise (SME) software developers including IBM, Bosch, Pelco, GE Security, Honeywell, Siemens, ObjectVideo, IOImage, Aimetis, Sony, Panasonic, Nice, Vidient, March Network, Mate, Ipsotek, Citilog, Traficon, and BRS Labs [25, 22].

[1] Frost and Sullivan estimates that the video surveillance software market will reach $670.7 million annually by 2011 [21].

[2] The growing interest on video analytics is also evident from various industrial focus conferences such as the IMS Video Content Analysis Conferences (http://www.imsconferences.com).

[3] Research conducted by the British Industry Security Association demonstrated that video analytics technologies are deployed by the transport and retail sectors most frequently (http://www.bsia.co.uk/aboutbsia/cctv/O5E926740891).

Current video analytics find applications in various areas. For instance, IBM assists the Chicago City in laying out city-wide video analytics system based on the IBM Smart Surveillance Solution (S3) to detect suspicious activity and potential public safety concerns [14]. The City of Birmingham, Alabama also sets up a surveillance system equipped with artificial neural network based analytic software developed by the BRS Labs to detect suspicious and abnormal situations. Besides street-level surveillance, video analytics also find wide applications in the transport sector. For example, the Bosch Intelligent Video Analysis (IVA) software is employed at the Athens International Airport [10]. For border control, the Video Early Warning (VEW) software developed by ObjectVideo is used along the US border to locate suspicious individual or vehicles attempting to cross into the country [50]. Other government and commercial deployment of video analytics include installations of the IOImage video analytics software at the Israeli parliament, and the Aimetis VE Series at Volkswagen Sachsen.

Security has been the dominant driver for the development and deployment of video analytics solutions. Some common applications are:

1. **Intruder detection** often implies tripwire detection or fence trespassing detection, which alerts an operator if an intruder is detected crossing a virtual fence[4]. The underlying algorithm involves the extraction of foreground objects using background subtraction, followed by examining whether the foreground objects overlap with a pre-defined region in an image space. This application is useful to ensure perimeter control for sensitive and restricted areas such as limited-access buildings or train track areas, e.g. BRS Labs AISight[5].
2. **Unattended object detection** aims to ignore items attended by nearby person and only triggers an alarm when an item is deposited in a controlled area longer than a pre-defined time period, e.g. Honeywell video analytics[6].
3. **Loitering detection** aims to detect persons who stay in a controlled area for an extended period of time. This is often achieved by tracking an individual and recording the time stamps of appearance and disappearance of the person. Loitering detection is useful in bringing about attention on suspicious behaviour in advance to an actual security breach or intrusion, e.g. MarchNetworks VideoSphere[7].
4. **Tailgating detection** aims to detect illegal follow-through behaviour at access control points, e.g. doorways. It relies on individual tracking in conjunction with an access control system. Alert is generated for an immediate review by security personnel if multiple persons enter a restricted area while only one of them is authorised by the access control system, e.g. Mate video analytics[8].

[4] A set of real-world datasets and alarm definitions are released as the Image Library for Intelligent Detection Systems (i-LIDS), a UK government Home Office Scientific Development Branch (HOSDB) benchmark for video analytics systems [63], which has also been adopted by the US National Institute of Standards and Technology (NIST).

[5] http://www.brslabs.com/index.php?id=79

[6] http://www.honeywellvideo.com/support/library/videos/

[7] http://www.marchnetworks.com/Products/Video-and-Data-Analytics/

[8] http://mateusa.net/

5. **Crowd management** software monitors and collects statistics on the crowd volume by measuring the foreground occupancy level in a video. It can be used at transportation hubs and shopping malls to avoid overcrowding situations, e.g. Vident SmartCatch[9].

These applications provide some practical and useful solutions. Nevertheless, their effectiveness and success depend largely on rather stringent operational conditions in carefully controlled environments. There are growing concerns on the viability of using such analytics in real-world scenarios especially in unconstrained crowded public spaces. In particular, existing technologies for video content analysis largely rely on Video Motion Detection (VMD), hard-wired rules, and object-centred reasoning in isolation (i.e. object segmentation and tracking) with little if any context modelling. Such systems often suffer considerably high false alarm rate due to the changes in visual context, such as different weather conditions and gradual object behaviour drift over time. In addition, fully automated analysis of video data captured from public spaces is often intrinsically ill-conditioned due to large (and unknown) variations in video image quality, resolution, imaging noise, diversity of pose and appearance, and severe occlusion in crowded scenes. As a result, those systems that rely on hard-wired hypotheses and location-specific rules are likely to break down unexpectedly giving frequent false alarms, requiring elaborative re-configuration and careful parameter tuning by specialists, making system deployment non-scalable and hugely expensive. In the worst-case scenario, installed expensive video analytics systems are abandoned or otherwise infrequently used due to excessive operational burden and intolerable level of false alarms.

3 Emerging Techniques

Addressing the limitations of current systems demands more robust and intelligent computer vision solutions. In this section, we discuss several emerging video analysis techniques, which could help to remedy the problems with the existing video analytics technologies. We first highlight the recent developments in single view-based video analysis techniques, ranging from gauging individual intent (Section 3.1) to analysing crowd behaviour (Section 3.2). We then discuss the use of multiple cameras for cooperative monitoring of complex scenes (Section 3.3). Finally, we look into how one could exploit contextual information (Section 3.4) and learn from human feedback (Section 3.5) to facilitate more robust and smarter surveillance.

(a) (b) (c)

Fig. 1 (a) Facial expression, (b) head pose, and (c) gait may be used as behavioural cues to reveal human intention (images from [60], [52], [6]).

3.1 Intent Profiling

Psychological studies [16, 15] suggest that one's intention can be perceived from the microexpressions and incomplete motion cues. The findings have inspired the development of automated surveillance system to interpret human intent for making rapid and anticipatory detection of deceptive intention from visual observations. For instance, the US Department of Homeland Security undertakes an initiative to develop automated capabilities for the Fast Attribute Screening Test (FAST)[10] in order to link behavioural cues such as subtle changes in facial temperature to a variety of hidden emotions, and thereby spotting people being deceptive or planning for hostile acts.

Various computer vision studies have examined the possibility of inferring human emotion and intent based on temporal analysis of visual cues such as facial expression, head pose, body pose, and gait (see Fig. 1). In facial expression analysis [67], most systems extract either geometric-based [17] or appearance-based facial features, e.g. Gabor wavelets or local binary patterns [60], to recognise prototypic emotional expressions, such as anger or fear. Different from facial expressions, the head pose of a person may reveal one's focus of attention. Popular head pose estimation methods [47] include holistic template based approaches, i.e. classifying a pose direction based on the appearance of the entire face image, or local feature set based approaches, i.e. corresponding facial landmarks such as eyes and lips to a set of trained poses. Recent studies have attempted to estimate head pose in low-resolution images [8] as well as crowded surveillance videos [52]. In addition to head pose, body posture configuration [46] and gait [49] may also play an important role in human intent inference. In particular, by tracking the body posture of a person over time, we may discover angry or aggressive-looking postures, indicating threatening intentions. A common strategy to articulated body pose estimation is to exploit the mapping of kinematically constrained object parts to a pictorial structure for the appearance of body parts [20]. As opposed to body pos-

[9] http://www.vidient.com/solutions/transportation.php

[10] http://www.dhs.gov/xres/programs/gc_1218480185439.shtm

ture inference, gait analysis typically models the characterisation of periodic motion and spatio-temporal patterns of human silhouettes. Recent studies on gait analysis have been focusing on coping with variations caused by various covariate conditions (e.g. clothing and view angle) [6], and distinguishing abnormal walking styles (e.g. walking with objects attached to the legs) [55].

Gauging one's intent is challenging because behavioural cues are often incomplete, vary from person to person, and may last only a fraction of time. In addition, an image based analysis such as facial expression recognition becomes difficult given low-resolution video images captured from crowded public scenes. For accurate and robust intent inference, it is not only necessary to analyse the low-level imagery features and their inter-correlation, but also critical to model the high-level visual context, i.e. correlations among facial expression, head pose, body pose, and gait as well as their relationships with other entities in a scene.

3.2 Crowded Scene Analysis

(a) (b) (c)

Fig. 2 A crowded scene may be structured ((a) and (b)) where crowd motion is almost directionally coherent over time, or unstructured ((c)) where the motion of the crowd at any given location is multi-modal. Object tracking in crowded scenes is very challenging due to severe inter-object occlusion, visual appearance ambiguity, and complex interactions among objects.

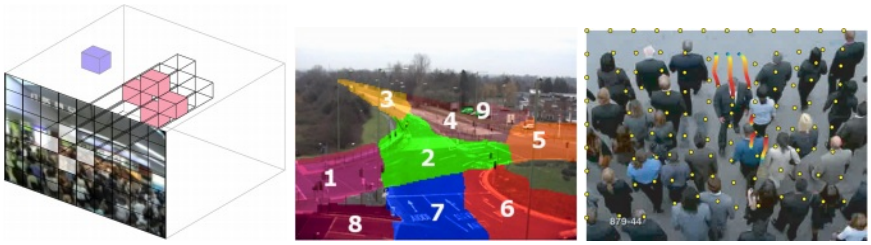
Crowded scene analysis is indispensable in most surveillance scenarios since most video analyses, e.g. intent profiling, have to be carried out in unconstrained and crowded environments (see Fig. 2). Crowded scene analysis can be categorised into three main problems: (1) crowd density estimation and people counting, (2) tracking in crowd, and (3) behaviour recognition in crowd. There exist commercial applications that support crowd density estimation for overcrowding detection, such as the solutions developed by iOmniscient and VideoIQ. However, commercial systems for tracking and behaviour analysis in crowded public scenes are almost non-existent. The main reason is that most existing commercial solutions generally rely on object-centred representation with little if any context modelling. Specifically, they generally assume reliable object localisation and detection as well as smooth object movement. These assumptions are often invalid in real-world surveillance settings characterised by severe inter-object occlusions due to excessive number of

objects in crowded scenes (Fig. 2). In this section, we focus our discussion on recent advances on behaviour analysis in crowded scenes, adopting a non-trajectory based representation. Tracking in crowd will be discussed in Section 3.4. Readers are referred to [32] for a review on crowd density estimation.

(a) (b)

Fig. 3 Examples of crowded scene analysis: (a) Action detection in crowd, an example of detection of 'standing' action in a busy underground station. (b) Crowd analysis involving multiple objects, a police car breaks a red light and turns right through opposing traffic. Here the right flow of the other traffic is a typical action, as is the left flow of the police car. However, their conjunction (forbidden by the traffic lights) is not (images from [62], [28]).

One of the key problems in crowd behaviour understanding is *action detection*, which aims to detect specific action, e.g. fighting or falling down in a crowded scene (Fig. 3(a)). The problem of action detection in crowd is largely unresolved as compared to the extensively studied action classification problem in well-defined environments [53]. Specifically, unlike action classification that assumes availability of pre-segmented single action sequence with fairly clean background, action detection in crowd does not assume well-segmented action clips. A model needs to search for an action of interest that can be overwhelmed by a large numbers of background activities in a cluttered scene. Existing approaches [33, 75] typically construct a set of action templates based on a single sample per action class. These templates are then used for matching given an unseen clip. The models may not be able to cope with large intra-class variations since only one sample per action class is used for a model. The intra-class variations can be captured using large numbers of training actions, but requiring manual annotations that can be time-consuming and unreliable. To generate sufficient training data without laborious manual annotation, different approaches have been proposed, e.g. the use of a multiple instance learning framework [29], a greedy k-nearest neighbour algorithm for automated annotation of positive training data [62], and a transfer learning framework [12] to generalise action detection models built from a source dataset to a new target dataset. For the detection strategy, most existing studies perform action detection by using sliding 3D search windows [29, 62]. This searching method, however, can be computationally prohibitive due to the enormous search space. This problem is addressed by Yuan et al. [76] using a 3D branch and bound searching method.



(a) (b) (c)

Fig. 4 After computing low-level activity features such as optical flow, there are different approaches to represent activity patterns. For example, (a) decomposition into local fixed-size spatio-temporal volumes, (b) decomposition into regions, each of which encapsulates location specific activities that differ from those observed in other regions, and (c) overlay of particles on the optical flow fields (images from [35] (Original video sequence is courtesy of Nippon Telegraph and Telephone Corporation) , [40], [45]).

Beyond action recognition that focuses on the behaviour of individuals, another important research area is crowd analysis, which aims to derive a collective understanding of behaviours and interactions of multiple objects co-exist in a common space (see Fig. 3(b)). Conventional methods [51,27] often start with individual tracking in the scene. Owing to the unreliability of tracking caused by extensive clutter and dynamic occlusions in crowded scenes, increasing numbers of studies approach the problem using a holistic representation to avoid explicit object segmentation and tracking. In particular, recent studies tend to represent activity patterns using pixel-level features, including foreground pixel changes [7,66], optical flow [28, 37], texture [42], and gradients of pixel intensities [35]. To construct a holistic representation given low-level pixel-based features, a model decomposes the scene and represents local activities as local fixed-size spatio-temporal volumes of features [34,35,28,70,45] (Fig. 4(a)). There are also studies [40,66] that decompose a scene into different regions, which are semantically relevant to the activity dynamics and structural knowledge of a scene (Fig. 4(b)). Alternatively, motivated by the studies on fluid dynamics, the notions of particle flow [3] and streak flow [44] are also exploited (Fig. 4(c)). These studies overlay a cloud of particles over optical flow fields and subsequently learn the dynamics and interactive forces of these moving particles for crowd segmentation [3,44] and abnormal crowd behaviour detection [45,73,44].

Given a representation of localised activity patterns, activity modelling is further considered for learning the space-time dependencies between local activities. To that end, suitable statistical learning models include dynamic Bayesian networks (DBNs) [35], Markov random field (MRF) [34], and probabilistic topic models (PTMs) [70, 28, 37]. Among them, PTMs such as Latent Dirichlet Allocation (LDA) [9] and Hierarchical Dirichlet Processes (HDP) [64] have gained increasing popularity. The PTMs are essentially bag of words models that perform clustering by concurrency. Specifically, local visual activities and video clips are often treated analogously as 'words' and 'documents'. Each video clip may be viewed as a mix-

ture of various 'topics', i.e. a cluster of co-occurring words in different documents. In general, PTMs are less demanding computationally and less sensitive to noise in comparison to DBNs due to the bag of words representation. This advantage, however, is gained at the expense of throwing away explicit temporal dependencies between local activities. Different solutions have been proposed to address this shortcoming, e.g. by introducing additional Markov chain to a topic model for modelling explicit temporal dependencies [28].

3.3 Cooperative Multi-Camera Network Surveillance

Multi-camera surveillance is another important and emerging research topic. In complex public scenes, multiple-camera network systems are more commonly deployed than single-camera systems. Specifically, disjoint cameras with non-overlapping field of view (FOV) are more prevalent, due to the desire to maximise spatial coverage in a wide-area scene whilst minimising the deployment cost. Most existing commercial applications for activity understanding and unusual event detection are designed for single-camera scenarios. Very few working systems are available for interpreting activity patterns across networked multiple disjoint camera views for global analysis and a coherent holistic situational awareness.

In this section, we highlight some efforts that have been made in the last few years by the computer vision community towards developing multi-camera video analytics, focusing on multi-camera object tracking and activity analysis.

Fig. 5 Partial observations of activities observed from different camera views: a group of people (highlighted in green boxes) get off a train [Cam 8, frame 10409] and subsequently take an upward escalator [Cam 5, frame 10443] which leads them to the escalator exit view [Cam 4, frame 10452]. Note that the same objects exhibit drastic appearance variations due to changes in illumination, camera viewpoint, and the distance between the objects and the cameras.

Object tracking across camera views is a major research topic due to its potential usefulness in visual surveillance, e.g. monitoring long-term activity patterns of targeted individuals. Current solutions mostly achieve inter-camera tracking by matching the visual appearance features and motion characteristics, e.g. speed, of a target object across views. The appearance features are often extracted from the entire individual body since biometric features such as facial appearance is no longer

reliable under typical surveillance viewing conditions. Inter-camera tracking, also known as *person re-identification*, aims to associate individuals observed at diverse locations in a camera network. Compared to multi-camera object tracking, methods devised for person re-identification generally ignore the temporal constraints across views and match objects based solely on appearance features. Due to the disparities in space, time, and viewpoint among disjoint cameras over different physical locations, objects travelling across such camera views often undergo drastic appearance variations (Fig. 5). To remedy the problem, different strategies have been proposed, including the mapping of colour distribution from one camera to another using a brightness transfer function [30, 54], exploiting contextual information extracted from surrounding people to resolve ambiguities [78], and computing more robust image features through incremental learning [23], boosting [26], or exploiting asymmetry/symmetry principles [19].

Robustness and accuracy in inter-camera tracking and person re-identification cannot be achieved ultimately by matching imagery information alone. It is essential to formulate and model knowledge about inter-camera relationships as contextual constraints in assisting object tracking and re-identification over significantly different camera views [13, 66]. The problem of inferring the spatial and temporal relationships between cameras is often known as *camera topology inference* [43,68], which involves the estimation of camera transition probabilities, i.e. how likely an object exiting a camera view would reappear in another camera view, and a inter-camera time delay distribution, i.e. travel time needed to cross a blind area. Recent studies on topology inference have been focusing on disjoint camera networks. A common strategy for learning the topology of a disjoint camera network is by matching individual object's visual appearance or motion trends. This is essentially similar to the multi-camera object tracking and the person re-identification tasks as discussed above. Once object correspondences are established using a large amount of observations, it would be straightforward to infer the paths and transition time distributions between cameras. However, without having to solve the correspondence problem explicitly, which is often nontrivial in itself, another popular strategy applicable to disjoint cameras is to infer inter-camera relationship through searching for a consistent temporal correlation from population activity patterns (rather than individual whereabouts) across views [43, 68, 66]. For example, Makris et al. [43] present an unsupervised method that accumulates evidence from a large set of cross-camera entrance/exit events, so as to establish a transition time distribution. A peak in the transition time distribution essentially implies a connection between the two camera views.

Global activity analysis across multiple camera views is another emerging problem to be solved, in which the goal is to build an activity model for understanding activities captured by multiple cameras holistically, e.g. performing unusual event detection in a global context. Performing global activity analysis in a public space through multiple cameras is non-trivial, especially with non-overlapping inter-camera views, in which global activities can only be observed partially with different views being separated by unknown time gaps. A straightforward approach to activity understanding and unusual event detection in multiple disjoint cameras is to

reconstruct the global path taken by an object by merging its trajectories observed in different views, followed by conventional single-view trajectory analysis [77]. With this approach, one must address the camera topology inference problem [43, 68] and the trajectory correspondence problem [30], both of which are still far from being solved. Wang et al. [71] propose an alternative trajectory-based method that bypasses the topology inference and correspondence problems by proposing a LDA-based co-clustering model. However, this model cannot cope with busy scenes and it is limited to capturing only co-occurrence relationships among activity patterns. In contrast to the trajectory-based approaches, Loy et al. [65] developed a method that automatically infers the unknown time delayed dependencies between local activities across views without relying on explicit object-centred segmentation and tracking. This technique is particularly useful in coping with low-quality public scene surveillance videos featuring severe inter-object occlusion therefore improving robustness and accuracy in multi-camera unusual event detection and object re-identification.

3.4 Context-aware Activity Analysis

(a) (b) (c)

Fig. 6 Visual context learning can benefit various surveillance tasks: (a) Tracking in crowd by leveraging the context learned from typical crowd motions, (b) arbitrating ambiguities in person re-identification tasks by exploiting the visual context extracted from surrounding people and objects, and (c) global unusual event detection in multiple disjoint cameras by learning a global dependency context (images from [56], [78], [65]).

Visual surveillance in public spaces is challenging due to severe occlusion, visual appearance variation, and temporal discontinuity. These factors contribute collectively in making visual observations noisy and incomplete, resulting in their interpretations ill-defined and ambiguous. To overcome this problem, a model needs to explore and discover extra knowledge about behavioural context from visual data. Activities in a public space are inherently *context-aware*, exhibited through implicit physical and social constraints imposed by the scene layout and correlated activities (and shared spaces) of other objects both in the same camera view and other

views. Strong psychophysical evidence [5, 58] suggests that visual contexts, which encompass spatio-temporal relations of an object with its surroundings, are crucial for establishing a clear comprehension of a scene. Current commercial video analytics solutions have yet to embrace visual context modelling whilst significant efforts have been made by the computer vision research community.

Object tracking in crowded public spaces is one of the application areas that can benefit greatly from visual context modelling. Various techniques have been proposed following this idea. For instance, tracking-by-detection [72, 11] exploits human or body part detection as categorical contextual information for more reliable tracking. There are also studies that exploit contextual information around a target object both spatially and temporally to facilitate more robust long-term tracking [48, 74]. In another study, Ali and Shah [4] exploit scene structure and behaviour of the crowd to assist appearance-based tracking in structured crowded scenes. The work is extended by Rodriguez et al. [56] and Kratz et al. [36] to unstructured crowded scenes, whereby tracking of individuals is aided by leveraging the contextual knowledge learned from typical multi-modal crowd motions (Fig. 6(a)). Visual context is also beneficial for resolving ambiguities from inter-camera tracking or person re-identification. For instance, Zheng et al. [78] embed contextual visual knowledge extracted from surrounding people into supporting descriptions for matching people across disjoint and distributed multiple views (Fig. 6(b)).

Visual context learning can also facilitate the detection of subtle unusual events otherwise undetectable in complex public scenes. For example, Li et al. [39] yield robust detection of unusual events with subtle visual difference but contextually incoherent. This system models both behaviour spatial and correlation context in a single wide-area camera view to provide situational awareness for where a behaviour may take place and how it is affected by other objects co-existing in the scene. Beyond a single camera view, activity understanding and unusual event detection in a multiple camera network can also benefit from the visual context learning [66, 65]. In particular, collective partial observations of an inferred global activity (not visually observable directly in a common space) are correlated and inter-dependent in that they take place following a certain temporal order even though with uncertain temporal gaps. Consequently, discovering the time-delayed correlations or dependencies between a set of visually disjoint partial observations can help to establish plausible and coherent visual context beyond individual camera views that facilitates more robust activity understanding (Fig. 6(c)).

3.5 Human in the Loop

A primary goal of a visual surveillance system is to detect genuine unusual events whilst ignoring distractors. Most unusual event detection methods [70, 45, 34] employ an outlier detection strategy, in which a model is trained using normal events through unsupervised one-class learning. Events that deviate statistically from the resulting normal profile are deemed unusual. This strategy offers a practical way

of bypassing the problems of imbalanced class distribution and inadequate unusual event training samples. However, the outlier detection strategy may subject to a few inextricable limitations as pointed out by Loy et al. [41]. Specifically, without human supervision, unsupervised methods have difficulties in detecting subtle unusual events that are visually similar to a large numbers of normally behaving objects co-existing in a scene. In addition, surveillance video of public spaces are highly cluttered with large number of nuisance distractors, often appearing visually similar to genuine unusual events of interest. Relying on information extracted from imagery data alone is computationally difficult to distinguish a genuine unusual event from noise. The usefulness of machine detected events can benefit from further examination using human expert knowledge. From statistical model learning perspective, constructing a model that encompasses 'all' normal events is inherently difficult. Given limited (and often partial) observation, some outlying regions of a normal class may be falsely detected as being unusual (and of interest) if no human feedback is taken into account for arbitrating such false alarms.

To overcome this inherent limitation of unsupervised learning from incomplete data, other sources of information need be exploited. Human feedback is a rich source of accumulative information that can be utilised to assist in resolving ambiguities during class decision boundary formation. An attractive approach to learn a model from human feedback is by employing an *active learning* strategy [59]. Active learning aims to follow a set of predefined query criteria to select the most critical and informative point for human feedback on labelling verification. This strategy for active selection of human verification on some but not all machine detected events allows a model to learn quickly with far fewer samples compared to passive random labelling strategy. Importantly, it helps in resolving ambiguities of interest when lacking visual distinctiveness, leading to more robust and accurate detection of subtle unusual events.

There have been very few active learning systems proposed for activity understanding and unusual event detection. Sillito and Fisher [61] formulate a method to harnesses human feedback on-the-fly for improving unusual event detection performance. Specifically, human approval is sought if a newly observed instance deviates statistically from the learned normal profile. If the suspicious instance is indeed normal, it will be included in the re-training process, or else it will be flagged as anomaly. In a more recent study, Loy et al. [41] propose a stream-based multi-criteria model for active learning from human feedback. In particular, the model makes a decision on-the-fly on whether to request human verification on unsupervised detection. The model selects adaptively two active learning criteria, likelihood criterion and uncertainty criterion, to achieve (1) discovery of unknown event classes and (2) refinement of classification boundary. The system shows that active learning helps in resolving ambiguities in detecting genuine unusual events of interest, leading to a more robust and accurate detection of subtle unusual events compared to the conventional outlier detection strategy.

4 Discussion

Current video surveillance technologies mostly suffer a high false alarm rate, oversensitive to visual context changes due to hard-wired rules, and poor scalability to crowded public scenes. Emerging techniques can help in mitigating some of these problems. In particular, video analytics can benefit from recent development in computer vision research for intent profiling, non-trajectory based representation in crowded scene analysis, multi-camera network cooperative activity monitoring, visual context modelling, and human-in-the-loop learning. There are other notable emerging trends in both algorithm and hardware development, which can also improve visual analytics for surveillance and security.

Robust and transfer video analysis aim to construct computer vision algorithms learning adaptively over long duration of time and across locations in order to cope with weather conditions, large environmental changes (e.g. different seasons in a calendar year), camera changes, and transitions of activity dynamics. Knowledge learned in a particular scene can be transferred selectively to new scenes without the need to invoke a new learning process from the beginning again.

Multi-sensor surveillance aims to exploit information from multiple heterogeneous sensors for collaborative analysis. Utilising different visual devices can be of benefit, including a combination of pan-tilt-zoom (PTZ) cameras, thermal cameras, stereo cameras, time-of-flight cameras, or wearable cameras. Non-visual sensors such as audio sensors, positioning sensors, and motion sensors can also be integrated into such a heterogeneous system in order to assist surveillance tasks, e.g. cooperative object detection and tracking using multiple active PTZ cameras [18] and wearable cameras [2].

On-the-fly variable level-of-detail content search can benefit from recent proliferation on high-resolution and low-cost cameras. Activity and behaviour based focus of attention can be developed to facilitate capabilities for dynamic sensing of visual content at variable level of details for on-the-fly automatic searching of interesting events and object in high-resolution, face recognition, and expression analysis from long distance in a crowded space. This can be exploited by either the deployment of selective high-resolution cameras or massively deployed random forest of redundant low-cost cameras. The use of higher resolution videos also demands tractable and specialised algorithms that are able to run in individual camera nodes, e.g. on a field-programmable gate array (FPGA) in a camera, to share the computational loads of the centralised processing server.

5 Further Reading

Interested readers are referred to the following further readings:

- [25] for a general overview of the video surveillance market, the architecture of a surveillance system, and the technology status of video analytics.

- [69, 38] for surveys on action and activity recognition.
- [32] for a survey on crowd analysis.
- [31, 1] for system perspectives on multiple camera activity analysis.
- [57] for trends on surveillance hardware development.

References

1. H. Aghajan and A. Cavallaro, editors. *Multi-Camera Networks: Principles and Applications*. Elsevier, 2009.
2. A. Alahia, P. Vandergheynsta, M. Bierlaireb, and M. Kunt. Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 2010.
3. S. Ali and M. Shah. A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.
4. S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *European Conference on Computer Vision*, pages 1–24, 2008.
5. M. Bar. Visual objects in context. *Nature Reviews: Neuroscience*, 5:617–629, 2004.
6. K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, 2010.
7. Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurrences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2458–2465, 2009.
8. B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *British Machine Vision Conference*, 2009.
9. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
10. Bosch. Athens international airport. http://www.boschsecurity.co.uk/, 2001.
11. M. D. Breitenstein. *Visual Surveillance - Dynamic Behavior Analysis at Multiple Levels*. PhD thesis, ETH Zurich, 2009.
12. L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1998–2005, 2010.
13. K.-W. Chen, C.-C. Lai, Y.-P. Hung, and C.-S. Chen. An adaptive learning method for target tracking across multiple cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
14. I. Corporation. Command, control, collabo-rate: public safety solutions from IBM. Solution Brief, 2009.
15. P. Ekman. Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3(1):34–38, 1992.
16. P. Ekman and W. V. Friesen. *Unmasking the Face*. Consulting Psychologists Press, 2 edition, 1984.
17. I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.
18. I. Everts, N. Sebe, and G. Jones. Cooperative object tracking with multiple PTZ cameras. In *International Conference on Image Analysis and Processing*, pages 323–330, 2007.
19. M. Farenzena, L. Bazzani, A. Perina, M. Cristani, and V. Murino. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360 –2367, 2010.
20. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

21. Frost & Sullivan. Video surveillance software emerges as key weapon in fight against terrorism. Press release - http://www.frost.com/.
22. Frost & Sullivan. Eyes on the network - Understanding the shift toward network-based video surveillance in Asia, 2007.
23. A. Gilbert and R. Bowden. Incremental, scalable tracking of objects inter camera. *Computer Vision and Image Understanding*, 111(1):43–58, 2008.
24. P. M. Gill, A. Spriggs, J. Allen, M. Hemming, P. Jessiman, D. Kara, J. Kilworth, R. Little, and D. Swain. Control room operation: findings from control room observations. Home office online report 14/05, Home Office, 2005.
25. V. Gouaillier and A.-E. Fleurant. Intelligent video surveillance: Promises and challenges. Technological and commercial intelligence report, CRIM and Technôpole Defence and Security, 2009.
26. D. Gray and H. Tao. Viewpoint invariant pedestrain recognition with an ensemble of localized features. In *European Conference on Computer Vision*, pages 262–275, 2008.
27. S. Hongeng, R. Nevatia, and F. Brémond. Video-based event recognition: Activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, 2004.
28. T. Hospedales, S. Gong, and T. Xiang. A Markov clustering topic model for mining behaviour in video. In *IEEE International Conference on Computer Vision*, pages 1165–1172, 2009.
29. Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *IEEE International Conference on Computer Vision*, 2009.
30. O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 26–33, 2005.
31. O. Javed and M. Shah. *Automated Multi-camera Surveillance: Theory and Practice*. Springer, 2008.
32. J. C. S. J. Junior, S. R. Musse, and C. R. Jung. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, pages 66–77, September 2010.
33. Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.
34. J. Kim and K. Grauman. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928, 2009.
35. L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1453, 2009.
36. L. Kratz and K. Nishino. Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 693–700, 2010.
37. D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? Discovering spatio-temporal dependencies in dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1951–1958, 2010.
38. G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(5):489–504, 2009.
39. J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *British Machine Vision Conference*, pages 193–202, 2008.
40. J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *European Conference on Computer Vision*, pages 383–395, 2008.
41. C. C. Loy, T. Xiang, and S. Gong. Stream-based active unusual event detection. In *Asian Conference on Computer Vision*, 2010.
42. V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

43. D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 205–210, 2004.
44. R. Mehran, B. E. Moore, and M. Shah. A streakline representation of flow in crowded scenes. In *European Conference on Computer Vision*, 2010.
45. R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behaviour detection using social force model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2009.
46. T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.
47. E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.
48. H. T. Nguyen, Q. Ji, and A. W. M. Smeulders. Spatio-temporal context for robust multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):52–64, 2007.
49. M. S. Nixon, T. Tan, and R. Chellappa. *Human identification based on gait*. Springer, 2005.
50. ObjectVideo. Hardening U.S. borders. http://www.objectvideo.com/, 2003.
51. N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.
52. J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *British Machine Vision Conference*, 2009.
53. R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.
54. B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *British Machine Vision Conference*, 2008.
55. Y. Ran, Q. Zheng, R. Chellappa, and T. M. Strat. Applications of a simple characterization of human gait in surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(4):1009–1020, 2010.
56. M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *IEEE International Conference on Computer Vision*, 2009.
57. R. Schneiderman. Trends in video surveillance give dsp an apps boost [special reports]. *IEEE Signal Processing Magazine*, 27(6):6–12, 2010.
58. O. Schwartz, A. Hsu, and P. Dayan. Space and time in visual context. *Nature Reviews Neuroscience*, 8:522–535, 2007.
59. B. Settles. Active learning literature survey. Technical report, University of WisconsinMadison, 2010.
60. C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.
61. R. Sillito and R. Fisher. Semi-supervised learning for anomalous trajectory detection. In *British Machine Vision Conference*, 2008.
62. P. Siva and T. Xiang. Action detection in crowd. In *British Machine Vision Conference*, 2010.
63. i. Team. Imagery library for intelligent detection systems (i-LIDS); a standard for testing video based detection systems. In *Annual IEEE International Carnahan Conferences Security Technology*, pages 75–80, 2006.
64. Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
65. C. C. Loy, T. Xiang, and S. Gong. Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In *IEEE International Conference on Computer Vision*, pages 120–127, 2009.
66. C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010.
67. Y.-L. Tian, T. Kanade, and J. F. Cohn. *Facial expression analysis*, chapter 11. Springer, 2005.
68. K. Tieu, G. Dalley, and W. E. L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *IEEE International Conference on Computer Vision*, pages 1842–1849, 2005.

69. P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities - a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.
70. X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):539–555, 2009.
71. X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):56–71, 2010.
72. B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.
73. S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2054–2060, 2010.
74. M. Yang, Y. Wu, and G. Hua. Context-aware visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1195–1209, 2008.
75. W. Yang, Y. Wang, and G. Mori. Efficient human action detection using a transferable distance function. In *Asian Conference on Computer Vision*, 2009.
76. J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2442–2449, 2009.
77. E. E. Zelniker, S. Gong, and T. Xiang. Global abnormal behaviour detection using a network of CCTV cameras. In *IEEE International Workshop on Visual Surveillance*, 2008.
78. W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *British Machine Vision Conference*, 2009.