Pattern Recognition Letters 34 (2013) 3–19

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Intelligent multi-camera video surveillance: A review

Xiaogang Wang *

Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

ARTICLE INFO

Article history:
Available online 20 July 2012

Keywords:
Multi-camera video surveillance
Multi-camera calibration
Topology of camera networks
Multi-camera tracking
Object re-identification
Multi-camera activity analysis

ABSTRACT

Intelligent multi-camera video surveillance is a multidisciplinary field related to computer vision, pattern recognition, signal processing, communication, embedded computing and image sensors. This paper reviews the recent development of relevant technologies from the perspectives of computer vision and pattern recognition. The covered topics include multi-camera calibration, computing the topology of camera networks, multi-camera tracking, object re-identification, multi-camera activity analysis and cooperative video surveillance both with active and static cameras. Detailed descriptions of their technical challenges and comparison of different solutions are provided. It emphasizes the connection and integration of different modules in various environments and application scenarios. According to the most recent works, some problems can be jointly solved in order to improve the efficiency and accuracy. With the fast development of surveillance systems, the scales and complexities of camera networks are increasing and the monitored environments are becoming more and more complicated and crowded. This paper discusses how to face these emerging challenges.

1. Introduction

Intelligent video surveillance has been one of the most active research areas in computer vision. The goal is to efficiently extract useful information from a huge amount of videos collected by surveillance cameras by automatically detecting, tracking and recognizing objects of interest, and understanding and analyzing their activities. Video surveillance has a wide variety of applications both in public and private environments, such as homeland security, crime prevention, traffic control, accident prediction and detection, and monitoring patients, elderly and children at home. These applications require monitoring indoor and outdoor scenes of airports, train stations, highways, parking lots, stores, shopping malls and offices. There is an increasing interest in video surveillance due to the growing availability of cheap sensors and processors, and also a growing need for safety and security from the public. Nowadays there are tens of thousands of cameras in a city collecting a huge amount of data on a daily basis. Researchers are urged to develop intelligent systems to efficiently extract information from large scale data.

The view of a single camera is finite and limited by scene structures. In order to monitor a wide area, such as tracking a vehicle traveling through the road network of a city or analyzing the global activities happening in a large train station, video streams from multiple cameras have to be used. Many intelligent multi-camera video surveillance systems have been developed (Collins et al., 2001; Aghajan and Cavallaro, 2009; Valera and Velastin, 2004). It is a multidisciplinary field related to computer vision, pattern recognition, signal processing, communication, embedded computing and image sensors. This paper reviews the recent development of relevant technologies from the perspective of computer vision. Some key computer vision technologies used in multi-camera surveillance systems are shown in Fig. 1.

1. Multi-camera calibration maps different camera views to a single coordinate system. In many surveillance systems, it is a key pre-step for other multi-camera based analysis.
2. The topology of a camera network identifies whether camera views are overlapped or spatially adjacent and describes the transition time of objects between camera views.
3. Object re-identification is to match two image regions observed in different camera views and recognize whether they belong to the same object or not, purely based the appearance information without spatio-temporal reasoning.
4. Multi-camera tracking is to track objects across camera views.
5. Multi-camera activity analysis is to automatically recognize activities of different categories and detect abnormal activities in a large area by fusing information from multiple camera views.

Different modules support one another and the arrows in Fig. 1 show the information flow between them.

While some existing reviews Valera and Velastin (2004) and Aghajan and Cavallaro (2009) tried to cover all the aspects of

* Tel.: +852 39438283; fax: +852 26035558.
E-mail address: xgwang@ee.cuhk.edu.hk

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.07.005

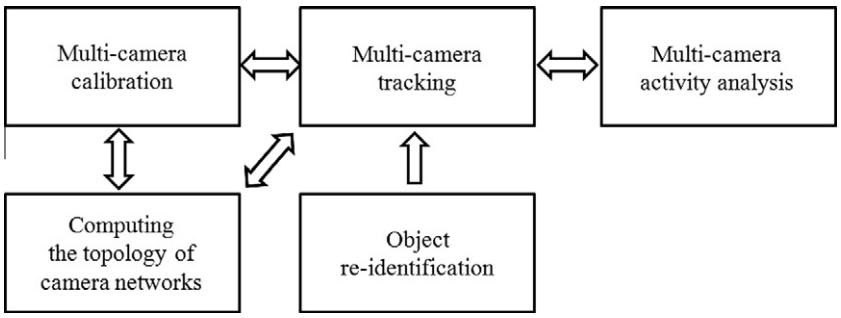

Fig. 1. Some technologies in intelligent multi-camera video surveillance. The arrows indicate the information flow between different modules.

architectures, technologies and applications, this paper emphasizes the connection and integration of these key computer vision and pattern recognition technologies in various environments and application scenarios and reviews their most recent development. Many existing surveillance systems solve these problems sequentially according to a pipeline. However, recent research works show that some of these problems can be jointly solved or even be skipped in order to overcome the challenges posed by certain application scenarios. For example, while it is easy to compute the topology of a camera network after cameras are well calibrated, some approaches are proposed to compute the topology without camera calibration, because existing calibration methods have various limitations and may not be efficient or accurate enough in certain scenarios. On the other hand, the topology information can help with calibration. If it is known that two camera views have overlap, the homography between them can be computed in an automatic manner. Therefore, these two problems are jointly solved in some approaches. Multi-camera tracking requires matching tracks obtained from different camera views according to their visual and spatio-temporal similarities. Matching the appearance of image regions is studied in object re-identification. The spatio-temporal reasoning requires camera calibration and the knowledge of topology. Some studies show that the complete trajectories across camera views can be used to calibrate cameras and to compute the topology. Therefore, multi-camera tracking can be jointly solved with camera calibration and inference of the topology. Multi-camera tracking is often a pre-step for multi-camera activity analysis, which uses the complete tracks of objects over the camera network as features. It is also possible to directly model activities in multiple camera views without tracking object across camera views. Once the models of activities are learned, they can provide useful information for multi-camera tracking, since if two tracks are classified as the same activity category, it is more likely for them to be the same object. A good understanding of the relationship of these modules helps to design optimal multi-camera video surveillance meeting the requirements of different applications.

Intelligent multi-camera video surveillance faces many challenges with the fast growth of camera networks. A few of them are briefly mentioned below. More detailed discussions are found in later sessions.

- A multi-camera video surveillance system may be applied to many different scenes and have various configurations. As the scales of camera networks increase, it is expected that the multi-camera surveillance systems can self-adapt to a variety of scenes with less human intervention. For example, it is very time consuming to manually calibrate all the cameras on a large network and the human effort has to be repeated when the configuration of the camera network changes. Therefore, automatic calibration is preferred. Object re-identification and multi-camera activity analysis prefer unsupervised approaches in order to avoid manually labeling new training samples scenes and camera views change.
- The topology of a large camera network could be complex and the fields of views of cameras are limited by scene structures. Some camera views are disjointed and may cover multiple ground planes. These bring great challenges for camera calibration, inference of topology and multi-camera tracking.
- There are often large changes of viewpoints, illumination conditions and camera settings between different camera views. It is difficult to match the appearance of objects across camera views.
- Many scenes of high security interest, such as airports, train stations, shopping malls and street intersections are very crowded. It is difficult to track objects over long distances without failures because of frequent occlusions among objects in such scenes. Although some existing surveillance systems work well in sparse scenes, there are many challenges unsolved in their applications to crowded environments.
- In order to monitor a wide area with a small number of cameras and to acquire high resolution images from optimal viewpoints, some surveillance systems employ both static cameras and active cameras, whose panning, tilting and zooming (PTZ) parameters are automatically and dynamically controlled by the system. Calibration, motion detection, object tracking and activity analysis with hybrid cameras face many new challenges compared with only using static cameras.

This paper reviews the five key computer vision and pattern recognition technologies (i.e., multi-camera calibration, computing the topology of camera views, multi-camera tracking, object re-identification and multi-camera activity analysis) from Sections 2–6. Cooperative video surveillance both with static and active cameras is discussed in Section 7. Detailed descriptions of their technical challenges and comparison of different solutions are provided under each topic. Finally some unsolved challenges and future research directions are discussed in Section 8.

2. Camera calibration

Camera calibration is a fundamental problem in computer vision and is indispensable in many video surveillance applications. There has been a huge literature on calibrating camera views with respect to a 3D world coordinate system (Faugeras, 1993; Triggs, 1999; Jones et al., 2002; Hartley and Zisserman, 2004). They estimate both the intrinsic parameters (such as focal length, principal point, skew coefficients and distortion coefficients) and extrinsic parameters (such as the position of the camera center and the camera's orientation in world coordinates) of cameras. In video surveillance, it

often assumes that objects move on a common ground plane. These approaches require manually labeling salient points in the scene and recording their real coordinates in the 3D world. The required wide site survey is time-consuming, especially when the number of cameras is large. It is also difficult to measure 3D points which are not laid on the ground plane in wide surveillance scenes.

Besides manually selecting 3D points, there are other automatic ways of calibrating cameras. Cameras can be calibrated with objects whose 3D geometry is known (Tsai, 1986; Sturm and Maybank, 1999; Liebowitz and Zisserman, 1999; Heikkila, 2000; Zhang, 2000; Faugeras and Luong, 2001; Teramoto and Xu, 2002; Agrawal and Davis, 2003). Zhang (2000) proposes an approach of easily calibrating a camera by observing a known planar template with unknown motion. Both the camera and the planar template can be freely moved. It has a closed-form solution with good accuracy. Although this approach has been widely used in many application scenarios, calibrated templates are not available in wide-field surveillance scenes because their projections are of very small sizes on the image plane and supply poor accuracy for calibration. Some approaches (Beardsley and Murray, 1992; Cipolla et al., 1999; Liebowitz et al., 1999; Caprile and Grimson, 1990; Deutscher et al., 2002; Wong et al., 2003; Colombo et al., 2005; Krahnstoever and Mendonca, 2005) use vanishing points (which are points onto which parallel lines appear to converge in a perspective projected image) from static scene structures, such as buildings and landmarks, to recover intrinsic parameters from a single camera and extrinsic parameters from multiple cameras. They employ constraints from geometric relationships, such as parallelism and orthogonality, which commonly exist in architectural structures. In the absence of inherent scene structures, Lv et al. (2002, 2006) estimate vanishing points from object motions. They obtain the needed line segments by tracking the head and feet positions of a walking person. Zhang et al. (2008) assume the camera height and estimate three vanishing points corresponding to three orthogonal directions in the 3D world coordinate system based on motion and appearance of moving objects. It can recover both intrinsic and extrinsic camera parameters. Bose and Grimson (2003) track vehicles and detect constant velocities along linear paths to realize ground plane rectification instead of recovering the intrinsic and extrinsic parameters of cameras. Solar shadows of objects are commonly observed in natural environments and they can also be used to estimate the intrinsic and extrinsic parameters of cameras as well as the orientation of the light source (Antone and Bosse, 2004; Lu et al., 2005; Cao and Foroosh, 2006; Junejo and Foroosh, 2008). Cao and Foroosh (2006) use multiple views of objects and their shadows for camera calibration. Junejo and Foroosh (2008) use the shadow trajectories of two stationary objects during the course of a day to locate the physical location of the camera (GPS coordinates) and the date of image acquisition.

If two camera views have substantial overlap, a homography between them can be computed with calibration (Stein and Medioni, 1992; Thompson et al., 1993; Cozman and Krotkov, 1997; Stein, 1999; Lee et al., 2000; Black et al., 2002; Brown and Lowe, 2003; Baker and Aloimonos, 2003; Stauffer and Tieu, 2003; Lowe, 2004; Jannotti and Mao, 2006; Sheikh and Shah, 2008). Many approaches manually or automatically select and match static features from 2D images to compute an assumed homography between two camera views and calibrate multiple camera views to a single global ground plane (Brown and Lowe, 2003; Baker and Aloimonos, 2003; Jannotti and Mao, 2006). The selected features are typically corner points, such as Harris corners (Harris and Stephens, 1988) and Scale-Invariant Feature Transform (SIFT) points (Lowe, 2004). They are matched by local descriptors which characterize texture or shape of their neighborhoods to establish correspondences. Comparisons of various keypoint detectors and local descriptors can be found in (Salti et al., 2011; Mikolajczyk and Schmid, 2005). The matching needs to be robust to the variations of viewpoints and lightings between camera views. The automatically obtained pairwise correspondences between feature points may include a significant amount of false matches. RANSAC Lacey et al., 2000 is used to find the homography that brings the largest number of feature points into match. There are also approaches of computing the homography based on the tracks of objects (Caspi and Irani, 2000; Lee et al., 2000; Stauffer and Tieu, 2003; Sheikh and Shah, 2008; Pflugfelder and Bischof, 2010). Lee et al. (2000) track objects simultaneously in partially overlapped camera views and use the object centroids as potential point correspondences to recover the homography between two camera views with planar geometric constraints on the moving objects. When the scene is sparse, the number of possible correspondences is small according to a temporal constraint that two corresponding tracked image centroids should be observed around the same time. A robust RANSAC variant is used to find a subset of centroids that best fits the homography. It faces problems when the scene is crowded. It is assumed that all the objects move on a single ground plane which is aligned with multiple camera views according to the computed homographies of multiple camera pairs. The 3D camera configuration and ground plane position and orientation are recovered up to a scale factor. The cameras are not necessarily well synchronized and the geometric constraints can align the tracking data in time. Caspi et al. (2006) extend Lee's method without restricting objects to a single ground plane. They enforce consistent matching of all the centroid points along track sequences instead of only a few pairs of centroids. Stauffer and Tieu (2003) jointly solve the problem of tracking objects across camera views and computing the homographies between overlapping camera views. An example of calibrated camera views from tracked objects is shown in Fig. 2. Pflugfelder and Bischof (2010) propose an approach of simultaneously estimating the translations between two synchronized but disjoint cameras and the track of a moving object in the 3D space. It requires correspondences of tracks observed in different camera views.

3. Computing the topology of camera views

Topology identifies camera views that are overlapped or spatially adjacent. Spatial adjacency means that there is no other view-field between the two camera views and hence there may potentially exist an inter-connecting pathway directly connecting tracks of objects observed in the two camera views. When an object leaves a camera view, it may reappear in some of other adjacent camera views with certain probabilities. Due to the constraints of scene structures and the configurations of camera networks, the topology of camera views could be complex. The camera views can be overlapped or disjoint, adjacent or far away from each other. There are "blind areas" between two adjacent but disjoint camera views, which makes multi-camera tracking difficult. The scene of a camera view can be modeled with structures such as source regions (where objects enter camera views), sink regions (where objects exit from camera views), and the paths connecting sources and sinks. Therefore, the topology can be described in a more detailed way with a network, where nodes are sources and sinks and edges are paths (within or across camera views) connecting sources and sinks. An example is shown in Fig. 3. These scene structures can be manually input or automatically learned from surveillance data (Stauffer, 2003; Makris et al., 2004; Wang et al., 2008).

The knowledge of the topology is important to assist tracking object across camera views (Kettnaker and Zabih, 1999). According to the topology information, the tracker of one camera can "hand-over" the track to the tracker in another adjacent camera view. The topology network can be augmented by associating an edge with a

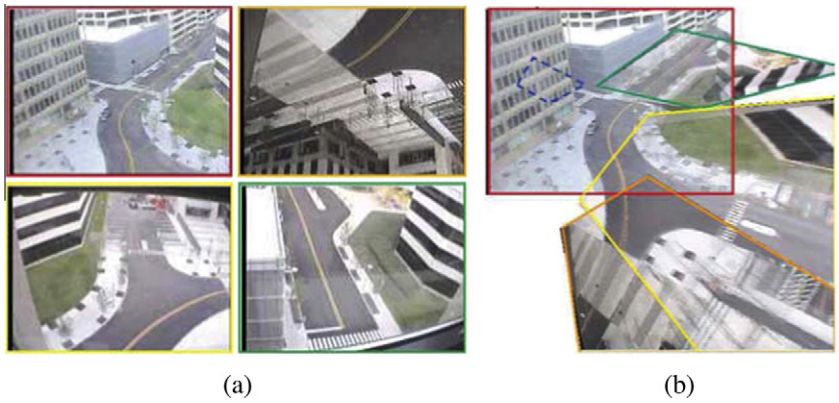

Fig. 2. (a) Four camera views. (b) The camera views are aligned to a ground plane after automatically computing the homographies between adjacent camera views using the approach in (Stauffer and Tieu, 2003). The figure is reproduced from Stauffer and Tieu (2003).

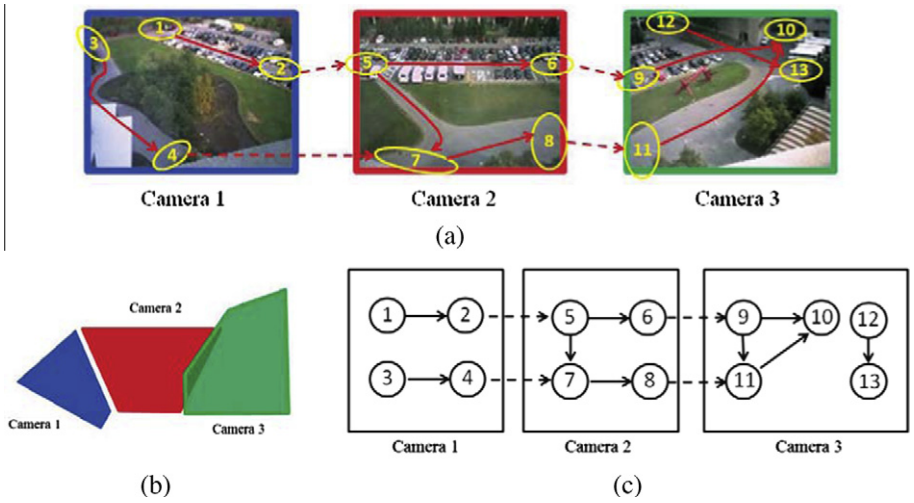

Fig. 3. Example of the topology of camera views. (a) Three camera views, their sources and sinks (indicated by yellow circles with numbers), and the paths between sources and sinks (the solid arrows indicate paths within camera views and the dash arrows indicate paths crossing camera views). (b) The topology of the three camera views. (c) The topology of sources and sinks in multiple camera views. Nodes indicate sources and sinks and edges indicate paths. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distribution of transition time between a sink and a source. These distributions can be learned from training data in supervised or unsupervised ways. When an object disappears from a sink region in a camera view, we can predict when and where the object will reappear in another camera view utilizing the topology network. This spatio-temporal reasoning can solve a lot of ambiguity during multi-camera tracking.

If cameras are already calibrated with a single 3D world coordinate system, the topology can be computed in a straightforward way. The spatial adjacency can be found through geometric analysis and the viewfields of cameras. Otherwise, it has to be inferred from training data. The proposed approaches are in two categories: correspondence-based (Kettnaker and Zabih, 1999; Javed et al., 2003) and correspondence-free (Ellis et al., 2003; Makris et al., 2004). Correspondence means the knowledge that tracks observed in different camera views actually correspond to the same object. It can be obtained manually or with some object identification technologies (such as a license plate reader or face recognition). Automatic object identification may be difficult especially in far-field video surveillance where objects are small in size. For the first category, Javed et al. (2003) use Parzen windows to estimate the distribution of inter-camera transition time from some training tracks with manually labeled correspondence. The learned distribution is used to improve multi-camera tracking.

For correspondence-free approaches, Ellis et al. (2003), Makris et al. (2004) learn the temporal transitions between sources and sinks from the cross-correlation between disappearing events and appearing events. Sources and sinks are not linked if their appearing and disappearing events are statistically independent. It assumes that if a sink and a source are adjacent, a pair of disappearing and reappearing events observed from them and caused by the same object should have a temporal difference less than T seconds. It collects all pairs of disappearing and reappearing events satisfying this temporal constraint and computes a distribution of the transition time between the source and the sink. It also assumes that this distribution only has a single mode and exhaustively searches for the location of the mode. It does not work well in the cases when the scene is busy or objects on the same path travel with different speeds. Both cases lead to multi-modal distributions of the transition time. Tieu et al. (2005) infer the topology of non-overlapping camera networks by measuring statistical dependency between observations, such as transition time and color appearance of objects, in different camera views under a information-theoretic framework. It is assumed that adjacent camera views have a large degree of dependence. The statistical dependence is measured using non-parametric estimation and the uncertainty of correspondence is integrated out in a Bayesian manner. It can be applied to multi-modal transition distributions.

4. Object tracking across camera views

Multi-camera tracking consists of two parts: (1) intra-camera tracking, *i.e.* tracking objects within a camera view; and (2) inter-camera tracking, *i.e.* associating the tracks of objects observed in different camera views. There is a huge literature on intra-camera tracking and a comprehensive survey can be found in (Yilmaz et al., 2006). This section focuses on inter-camera tracking, which is more challenging because (1) the prediction of the spatio-temporal information of objects across camera views is much less reliable than in the same camera view and (2) the appearance of objects may undergo dramatic changes because of variations of many factors, such as camera settings, viewpoints and lighting conditions, in different camera views.

4.1. Inter-camera tracking based on multi-camera calibration

The most typical way of multi-camera tracking is to track objects in a 3D coordinate system (Mikic et al., 1998; Dockstader and Tekalp, 2001; Li et al., 2002; Focken and Stiefelhagen, 2002; Mittal and Davis, 2003; Pflugfelder and Bischof, 2007) or on a single global ground plane (Chang and Gong, 2001; Black et al., 2002; Otsuka and Mukawa, 2004; Hu et al., 2006; Fleuret et al., 2008; Straw et al., 2010) or based on the homography between camera views (Lee et al., 2000; Caspi and Irani, 2000; Khan and Shah, 2006; Eshel and Moses, 2008) after calibration. Tracks of objects observed in different camera views are stitched based on their spatial proximity in the 3D coordinate system or on the common ground plane. It is usually assumed that the topology of camera views and the camera calibration are already solved before the tracking stage (Cai and Aggarwal, 1996). However, there also exist approaches which jointly infer the topology of camera views, calibrate cameras, and track objects across camera views (Stauffer and Grimson, 2000; Rahimi et al., 2004). They assume that inter-camera tracking can also help with the inference of topology and camera calibration. Rahimi et al. (2004) simultaneously recover the calibration parameters of cameras and track object across disjoint camera views under a Bayesian formulation. Stauffer and Tieu (2003) jointly infer the topology of camera views, estimate the homography between camera views and establish the correspondences of object tracks. If cameras are uncalibrated but have overlapping fields of views (FOV), finding the limits of FOV of each camera as visible in the other cameras can help with association of tracks. Khan and Shah (2003) propose a method to automatically recover FOV lines, which are the boundaries of the FOV of a camera in another camera views, by observing the motions of objects. If the FOV lines are known, it is possible to disambiguate among multiple possibilities for correspondence.

In some video surveillance scenarios, there is a need to track a large number of objects in crowded environments, where occlusions happen frequently due to the interactions among objects. Multi-camera tracking can better solve the challenge of occlusions, because it fuses the information from multiple camera views for robust tracking. For example, when an object is occluded in one of camera views, tracking can be switched to a better view without occlusions by predicting the existence of occlusions in camera views (Utsumi et al., 1998; Sogo and Ishiguro, 2000; Dockstader and Tekalp, 2001; Mittal and Davis, 2003). Cai and Aggarwal (1996) measure the tracking confidence, which is low when an object is occluded. When the tracking confidence is below a certain threshold, tracking is switched to an optimal camera view with the highest tracking confidence. Fleuret et al. (2008) predict occlusions with a generative model and a probabilistic occupancy map. Otsuka and Mukawa (2004) estimate the occlusion structures based on an explicit model of the geometric structure of the process that creates occlusions between objects. It is formulated as a recursive Bayesian estimation problem and implemented by particle filtering. With calibration, the observations from multiple camera views can be mapped to points in a single 3D world coordinate system. Some observations are missed if objects are occluded in some camera views. The Kalman filter (Mikic et al., 1998; Black et al., 2002), the extended Kalman filter (Straw et al., 2010) and the particle filter (Otsuka and Mukawa, 2004; Perez et al., 2004; Kim and Davis, 2006) are used to track objects in the 3D world coordinate system with occlusion handling. If the 3D coordinates are not available, the homography constraint between camera views can also be used to solve occlusions (Khan and Shah, 2006; Eshel and Moses, 2008).

4.2. Inter-camera tracking with appearance cues

Most of the approaches discussed above assume that adjacent camera views have overlap and therefore the spatial proximity of tracks in the overlapping area can be computed. In order to track objects across disjoint camera views, appearance cues have to been integrated with spatio-temporal reasoning (Alexander and Lucchesi, xxxx; Huang and Russell, 1997; Pasula et al., 1999; Veenman et al., 2001; Javed et al., 2003; Shafique and Shah, 2003; Morariu and Camps, 2006; Jiang et al., 2007; Song and Roy-Chowdhury, 2008; Hamid et al., 2010; Kuo et al., 2010). Various frameworks have been proposed. The Bayesian formulation is a natural way to integrate multiple types of features. It computes the posterior of object matching given evidence observed in different camera views. Huang and Russell (1997) propose a Bayesian approach to integrate the colors and the sizes of objects with velocities, arrival time and lane positions to track vehicles between two camera views. It models the probabilities of predicting the appearance or spatio-temporal features of objects observed in one camera view conditioned on their observations in the other camera view. Pasula et al. (1999) extend this approach to track objects across a large number of camera views. Instead of modeling the conditional probabilities of features between two camera views, it introduces hidden variables to characterize the intrinsic properties of appearance and spatio-temporal features in a Bayesian network. Javed et al. (2003) employ kernel density estimators to estimate the probability of an object entering a camera view with a certain travel time given the location and velocity of its exit from another camera view. It requires training data whose correspondences are labeled. The change of appearance between camera views is computed as the distance between color histograms. The probability of color distance is modeled as a Gaussian distribution which is learned for each pair of camera views from the training data. Matei et al. (2011) integrate appearance and spatio-temporal likelihoods within a multi-hypothesis framework. Instead of adopting a Bayesian approach, Morariu and Camps (2006) use manifold learning to match the appearance of objects across camera views. High dimensional images are mapped to low dimensional manifolds which are learned from sequences of observations. The manifolds of different camera views are aligned by capturing the temporal correlations between sequences. With the aligned manifolds, it extracts the intrinsic coordinates of the observed objects and establishes their correspondences.

The spatio-temporal relationships and appearance relationships between camera views may change dynamically and therefore their models need to be updated adaptively. For example, the lighting conditions change throughout the day. The travel time of vehicles between camera views changes with the amounts of traffic on a road network within different periods of a day. Collecting reliable training samples is a major challenge for online updating models since manually labeled correspondences are not available at runtime. In (Huang and Russell, 1997), the parameters of appearance models are online updated under the Expectation–Maximization (EM) framework. Javed et al. (2003) update the probability models

using the online kernel density estimation (Lambert et al., 1999). Chen et al. (2008) proposed an online unsupervised approach to learn both spatio-temporal and appearance relationships for a camera network. It incrementally refines the clustering results of sources and sinks, and learns the appearance models by combining the spatio-temporal information and MCMC sampling. Kuo et al. (2010) use the Multiple Instance Learning (MIL) (Dietterich et al., 1997) to online learn a discriminative appearance model. The spatio-temporal constraints of tracks observed in two camera views can provide some weakly labeled training samples which include some potentially associated pairs of tracks and exclude impossible associations. The selected potentially associated pairs have false positives as noise. MIL can accommodate the ambiguity of labeling during the model learning process.

4.3. Solving correspondences across multiple camera views

Each camera view may capture a set of multiple objects within a short period of time. Tracking objects across multiple camera views leads to solving the correspondences of tracks among multiple sets of candidates. Given the similarities between tracks obtained in different camera views as discussed above, a most likely assignment problem remains to be solved under the constraint that a track in a camera view can match with at most of one tracks in another camera view. If there are only two camera views, this problem can be solved by the Hungarian algorithm (Kuhn, 1956) or be formulated as a weighted bipartite graph matching problem (Cox and Hingorani, 1994; Alexander and Lucchesi, xxxx; Veenman et al., 2001; Javed et al., 2003). The Hungarian algorithm requires computing a cost matrix based on the pairwise similarities between tracks obtained in two different camera views. Its complexity is $O(n^3)$ where n is the number of tracks. If it is formulated as a bipartite graph matching problem, each track is represented as a vertex in the graph. The weight of an edge connect two tracks in different camera views is their similarity. Bipartite graph matching is to find M disjoint paths in the graph and each path indicates association of tracks of the same object. It can be solved with a complexity of $O(n^{2.5})$ (Hopcroft and Karp, 1973). If there are more than two camera views, solving this problem is NP hard. Various optimization approaches have been proposed to find suboptimal solutions. In (Shafique and Shah, 2003; Hamid et al., 2010), various K-partite graph matching algorithms have been proposed to solve this problem. Wu et al. (2009) formulate the problem of finding correspondences across multiple camera views as a multidimensional assignment problem and solve it with a greedy randomized adaptive search procedure. Jiang et al. (2007) formulate it as a multi-path search problem and solve it with a proposed linear programming relaxation scheme.

5. Object re-identification

In some application scenarios, the topology of a camera network and tracking information are not available, especially when the cameras are far in distance and the environments are crowded. For example, only the snapshots of objects instead of tracks captured by different cameras are available. In this case spatio-temporal reasoning is not feasible or accurate for inter-camera tracking. In recent years, a lot of research work (Nakajima et al., 2003; Bird et al., 2005; Javed et al., 2005; Shan et al., 2005; Shan et al., 2005; Gheissari et al., 2006; Hu et al., 2006; Guo et al., 2007; Wang et al., 2007; Prosser et al., 2008; Guo et al., 2008; Hamdoun et al., 2008; Lin and Davis, 2008; Gray and Tao, 2008; Shan et al., 2008; Schwartz and Davis, 2009; Zheng et al., 2009; Farenzena et al., 2010; Prosser et al., 2010) has been done on matching objects such as vehicles and pedestrians observed in different camera views only using visual information without spatio-temporal reasoning. It is assumed that the observations of a pedestrian are captured in the same day and therefore his or her clothes or shape do not change much. Objects can be matched with a single shot (Javed et al., 2005; Shan et al., 2005; Wang et al., 2007; Lin and Davis, 2008; Gray and Tao, 2008; Schwartz and Davis, 2009; Zheng et al., 2009; Farenzena et al., 2010) or multiple shots (Nakajima et al., 2003; Gheissari et al., 2006; Bird et al., 2005; Hamdoun et al., 2008). This problem is called object re-identification. Studying object re-identification separately from inter-camera tracking helps to better understand the capability of object matching using visual features alone. Once it has been well investigated, it can be integrated with spatial and temporal reasoning at the later stages which can further prune the candidate sets to be matched. Object re-identification is very challenging. The same object observed in different camera views undergo significant variations of resolutions, lightings, poses and viewpoints. Because objects captured by surveillance cameras are often small in size and a lot of visual details such as facial components are indistinguishable in images, some of them look similar in appearance. Examples of observed pedestrians in different camera views are shown in Fig. 4. The ambiguities increase when the number of objects to be distinguished increases. Therefore, features and distance metrics used to match image regions need to be highly discriminative and robust to those inter-camera variations.

5.1. Features for object re-identification

The appearance of objects is usually characterized in three aspects, color, shape and texture. They are reviewed below. A single type of features are not powerful enough to capture the subtle differences of all pairs of objects. They are usually combined and weighted differently according to their discriminative power.

5.1.1. Color

Color histograms of the whole image regions are widely used as global features to match objects across camera views because they are robust to the variations of poses and viewpoints (Orwell et al., 1999; Krumm et al., 2000; Mittal and Davis, 2003; Park et al., 2006; Cheng and Piccardi, 2006). However, they also have the weakness that they are sensitive to the variations of lighting conditions and photometric settings of cameras and that their discriminative power is not high enough to distinguish a large number of objects. Various color spaces such RGB, Lab, HSV and Log-RGB have been investigated and compared in (Wang et al., 2007). By removing the lightness component in the HSV color space, the color variation across camera views can be greatly reduced. The Log-RGB color space is less sensitive to photometric transformations. It computes the first directional derivatives of the logarithm of the colors, which are essentially the ratios of neighboring colors. The color of a pixel is formed as the product of the incident illumination and the surface albedo. Since illumination remains constant in local regions, the ratios of neighboring colors can effectively remove the lighting component. Mittal and Davis (2003) apply Gaussian color models to solve the correspondences of color modes between camera views. Other color invariants (Cheng and Piccardi, 2006; Slater and Healey, 1996; Weijer and Schmid, 2006) are also proposed. In order to enhance the discriminative power, the image region of an object is partitioned into local regions, color histograms within local regions are computed and concatenated as features for object matching (Park et al., 2006).

5.1.2. Shape

Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005; Wang et al., 2007; Schwartz and Davis, 2009) characterizes local

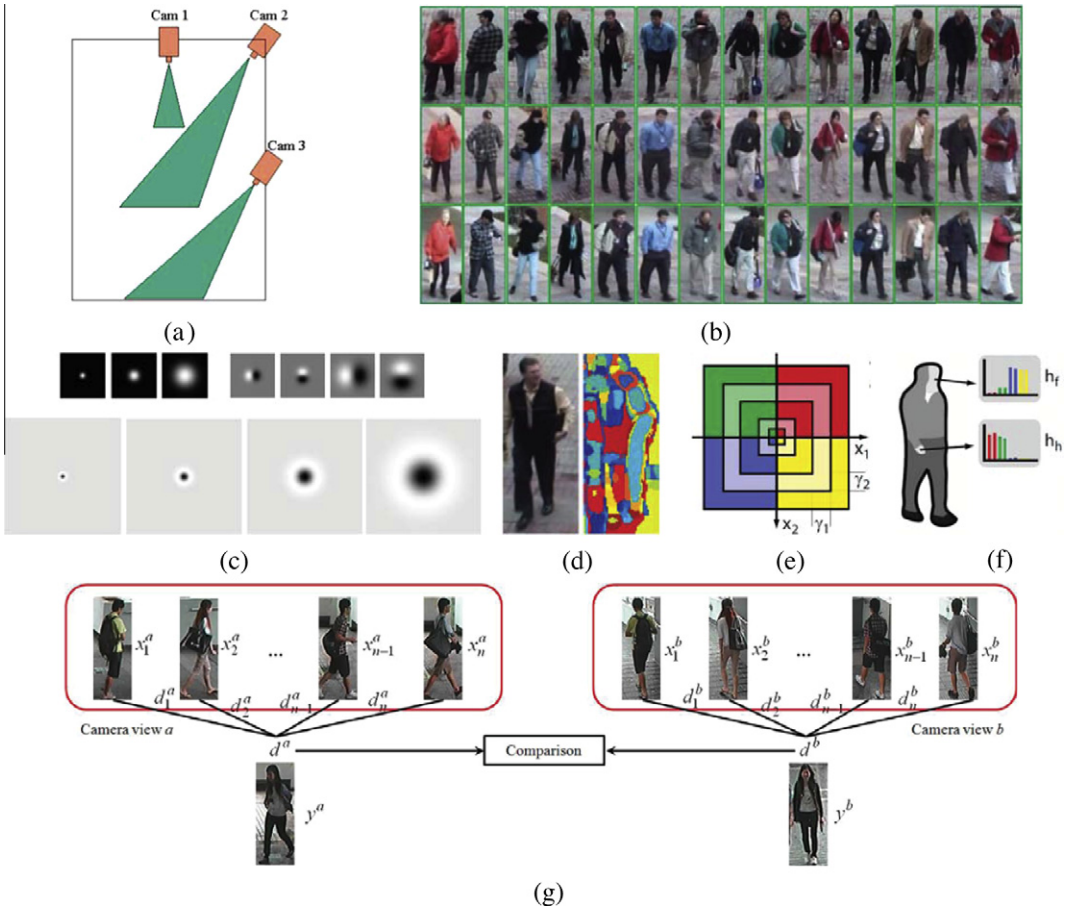

Fig. 4. Object re-identification across camera views. (a) A network with three cameras. (b) Examples of pedestrians observed in the three camera views. They are selected from the dataset introduced in (Wang et al., 2007). (c) A set of filter-banks proposed in (Winn et al., 2005). (d) Labels (indicated by different colors) when applying the filter-banks in (e) to the image on the left. (e) Example of a spatial kernel. (f) When the spatial kernel is place at different parts of a person (face and hand) different histograms of visual words are obtained. They can be used as discriminative features for object matching. (g) Exemplar-based object re-identification.

shapes by capturing edges and gradient structures. It computes the histograms of gradient orientations within cells which are placed on a dense grid and undergo local photometric normalization. It is robust to small translations and rotations of object parts. Shape context proposed by Belongie et al. (2002) characterizes both global and local shape structures. It is used to partition human bodies into constitutes for person re-identification by learning a shape dictionary in (Wang et al., 2007). There are also other models (Agarwal and Triggs, 2006; Carneiro and Lowe, 2006) proposed to characterize the geometric configuration of different local parts of objects.

5.1.3. Texture

Many filters, such as Gabor filter (Daugman et al., 1985) and other linear filter-banks (Winn et al., 2005; Varma and Zisserman, 2005; Leung and Malik, 1999), and local descriptors, such as SIFT (Lowe, 2004), color SIFT (Abdel-Hakim and Farag, 2006), Local Binary Patterns (LBP) (Ojala et al., 2002), Speeded Up Robust Feature (SURF) (Bay et al., 2006), Maximally Stable Extremal Regions (MSER) (Forssen, 2007), region covariance (Tuzel et al., 2006) and spin images (Lazebnik et al., 2003), have been proposed to characterize local texture and they can be applied to object re-identification (Hamdoun et al., 2008). These filters or descriptors can be applied to sparse feature points or on a dense grid. Their responses are usually quantized into visual words according to a pre-learned visual dictionary. A set of linear filter-banks proposed by Winn et al. is shown in Fig. 4(c). It combines Gaussians, Laplacian of Gaussians and first order derivatives of Gaussians in the Lab color space. Labels of quantized visual words with this set of filter-banks are shown in Fig. 4(d). With a bag-of-features model, the histogram of visual words of the whole image region is used as features for object matching. However, this feature is not discriminative enough. For example, it cannot distinguish a person wearing a white jacket and blue pants with one wearing a blue jacket and white pants. Therefore more features are proposed to capture the spatial distributions of visual words. On the other hand, the proposed features have to be invariant to the variations of poses and viewpoints when encoding the spatial information. Wang et al. (2007) propose shape and appearance context which computes the co-occurrence of shape words and visual words. It segments deformable objects into L parts using the shape context and a learned shape dictionary. Using a spatial kernel, which partitions the image domain into M subregions, as shown in Fig. 4(e), it models the spatial distributions of visual words relative to each of the object parts. When the spatial kernel is placed on one object part, the histograms of visual words within the subregions of the spatial kernel are computed. The $L \times M$ histograms are used as visual features for object matching. There are also other features such as correlograms (Huang et al., 1997) and correlations (Savarese et al., 2006) to capture the co-occurrence of visual words over spatial kernels.

5.1.4. Spatio-temporal features

Gheissari et al. (2006) propose an approach of utilizing local motion features for person re-identification. It establishes the correspondence between parts of different persons through

spatio-temporal segmentation with model fitting. Features for person re-identification are extracted by combining normalized colors and salient edgel histograms within different body parts.

5.1.5. Exemplar-based representations

Instead of directly matching visual features, some approaches (Shan et al., 2005; Guo et al., 2007) propose exemplar-based representations to overcome the dramatic variations of viewpoints. An graphical illustration is shown in Fig. 4(g). For two camera views a and b, n representative pairs $\{(\boldsymbol{x}_1^a, \boldsymbol{x}_1^b), \ldots, (\boldsymbol{x}_n^a, \boldsymbol{x}_n^b)\}$ are selected as exemplars. $\boldsymbol{x}_i^a$ and $\boldsymbol{x}_i^b$ are observations of the same object i captured in a and b respectively. If a sample $\boldsymbol{y}^a$ is observed in a, it is embedded amongst the representative exemplars also observed in a, and it is represented as a n-dimensional vector $\boldsymbol{d}^a = (d_1^a, \ldots, d_n^a)$, where d_i^a is the distance between $\boldsymbol{y}^a$ and $\boldsymbol{x}_i^a$ by matching their visual features. If a different sample $\boldsymbol{y}^b$ is observed in b, a vector $\boldsymbol{d}^b$ is obtained in the same way. If the change of viewpoints is large, it is more reliable to compare $\boldsymbol{d}^a$ and $\boldsymbol{d}^b$ than $\boldsymbol{y}^a$ and $\boldsymbol{y}^b$. The underlying assumption is that if an object to be identified is similar to one of the exemplar objects i, its observations in a and b should be similar to $\boldsymbol{x}_i^a$ and $\boldsymbol{x}_i^b$ respectively, and therefore both d_i^a and d_i^b are small no matter how different the two viewpoints are. It means that $\boldsymbol{d}^a$ and $\boldsymbol{d}^b$ are similar if $\boldsymbol{y}^a$ and $\boldsymbol{y}^b$ are the observations of the same object. However, it requires a set of representative exemplars for any pair of camera views and costs more manually labeling effort.

5.2. Learning for object re-identification

The photometric transformation between two camera views can be learned. Javed et al. (2005), Prosser et al. (2008) learn the Brightness Transfer Functions (BTFs) and the bi-directional Cumulative Brightness Transfer Functions (CBTF), which map the color observed in one camera view to that in another camera view, from training examples which are collected from a pair of camera views and whose correspondences are known. Porikli (2003) and Porikli et al. (2003) propose a non-parametric function to model color distortion between camera views using correlation matrix analysis and dynamic programming. Gilbert and Bowden (2006) incrementally and jointly learn the color mapping and the spatio-temporal transitions between camera views. It does not require manually labeled training examples with correspondences. These two types of transformations are complementary and support each other during the learning process.

Some approaches learn the similarity/distance metrics or select an optimal subset of features to match image regions observed in different camera views. Schwartz and Davis (2009) propose an approach of projecting high dimensional features to a low dimensional discriminant latent space by Partial Least Squares reduction (Wold, 1985). It weights features according to their discriminative power to best distinguish the observations of one object with those of others in a one-against-all scheme. Lin and Davis (2008) learn a different pairwise dissimilarity profile which best distinguishes a pair of persons. It is assumed that a feature may be crucial to discriminate two very similar objects but not be effective for other objects. Therefore it is easier to train discriminative features in a pairwise scheme. However, these two approaches require that all the objects to be re-identified have examples in the training set. If a new object is to be re-identified at the testing stage, the discriminant latent space or the dissimilarities have to be re-trained. Zheng et al. (2011) propose a *Probabilistic Relative Distance Comparison* model. It formulates object re-identification as a distance learning problem and maximizes the probability that a pair of true match has a smaller distance than a wrong match pair. The learned distance metric can be generalized to objects outside the training set. In (Gray and Tao, 2008; Prosser et al., 2010) boosting and RankSVM are used to select an optimal subset of features for matching objects across camera views. Shan et al. (2005, 2008) propose an unsupervised approach to learn discriminative edge measures for vehicle matching.

6. Multi-camera activity analysis

Activity analysis is a key task in video surveillance. It classifies activities into different categories and discovers typical and abnormal activities. The proposed approaches fall into two categories. The supervised approaches (Murata and Properties, 1989; Bobick and Ivanov, 1998; Oliver et al., 2000; Smith et al., 2005) require manually labeling training samples. However, since the observations of activities change dramatically in different camera views, it often requires relabeling training samples when these approaches are applied to different camera views. This limits their scalability and adaptability. On the other hand, it is very difficult to make these approaches robust to viewpoint transformation without the process of retraining. Video surveillance systems need to process video streams captured from a large number of cameras. The scales of camera networks are fast increasing nowadays. Therefore, people prefer unsupervised approaches (Brand and Kettnaker, 2000; Song et al., 2003; Wang et al., 2006, 2009) which can automatically learn the models of activities without labeling training samples. They can easily adapt to different scenes with little human intervention.

In far-field video surveillance, objects are small in size and the captured videos are of low resolution and poor quality. It is difficult to compute sophisticated features, such as poses, gestures, and appearance of objects. The activities of objects are mainly distinguished by their moving patterns. In many surveillance systems (Johnson and Hogg, 1995; Stauffer and Grimson, 2000; Oliver et al., 2000; Haritaoglu et al., 2000; Brand and Kettnaker, 2000; Medioni et al., 2001; Honggeng and Nevatia, 2001; Hu et al., 2004; Wang et al., 2006; Morris and Trivedi, 2008; Wang et al., 2008, 2011), objects are first detected and tracked and the activity of an object is then treated as sequential movements along its tracks. Usually only positions of objects are recorded along tracks, which are called trajectories. With positions and velocities as features, the motion patterns of trajectories can distinguish many different activity categories in far-field settings. Some examples are shown in Fig. 5. The activities of objects are regularized by scene structures, such as paths, sources and sinks. Many approaches (Keogh and Pazzani, 2000; Makris and Ellis, 2002; Porikli, 2003; Junejo et al., 2004; Fu et al., 2005; Zhang et al., 2006; Wang et al., 2011) have been proposed to cluster trajectories of objects into different activity categories without supervision. If a trajectory does not fit any of the typical activity models, it is detected as abnormality.

A natural way of doing activity analysis in multiple camera views is to first track objects across camera views and then use the complete trajectory of an object observed in different camera views for activity analysis with similar approaches developed for activity analysis in single camera views. For example, Zelniker et al. (2008) cluster stitched trajectories from multiple camera views and detected abnormalities. However, as discussed earlier, tracking objects across camera views require inferring the topology of camera views, calibrating camera views, and solving the correspondence problem, which are challenging especially when scene structures and the configurations of camera networks are quite arbitrary. The camera views may have any combination of large, little, or even no overlap. The objects may move on one or multiple ground planes. Some approaches (Wang et al., 2008, 2010; Loy et al., 2009) are proposed for activity analysis in multiple camera views without tracking objects across camera views. They will be discussed in Section 6.1.

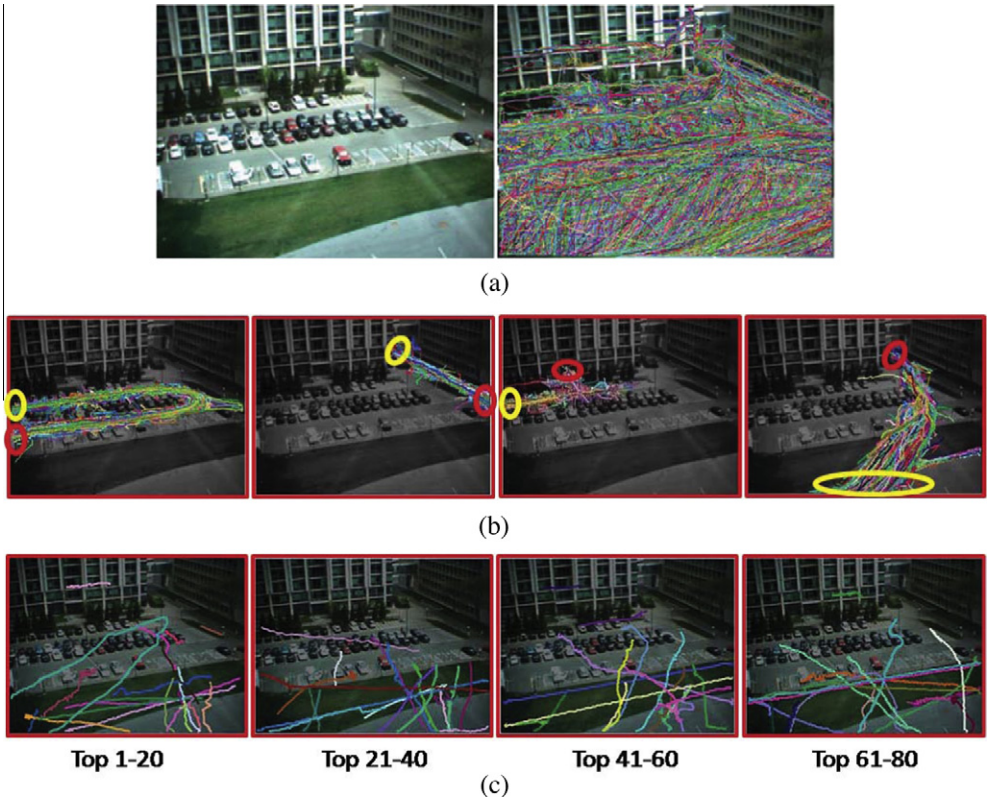

Fig. 5. (a) More than 40,000 trajectories collected in a parking lot and observed in a single camera view. Random colors indicate different trajectories. (b) Trajectories are clustered into different motion patterns which correspond to different activity categories using the approach proposed in (Wang et al., 2011). For example, the first motion pattern can be explained as vehicles making u-turns. The third motion pattern can be explained as pedestrians coming out of the building and leaving the parking lot. The activities are regularized by scene structures. Red and yellow circles indicate sources and sinks. (c) Top 80 abnormal trajectories detected using the approach proposed in (Wang et al., 2011). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In near-fields, more features of objects, such as color, texture, shape, gestures and movements of body parts can be observed. Therefore, activities can be analyzed with more categories and with more detailed features. It is called action recognition in this paper. These features change dramatically when they are observed in different camera views. Some examples are shown in Fig. 9. Many approaches (Rao et al., 2002; Junejo et al., 2008, 2011; Yilmaz and Shah, 2005; Syeda-Mahmood et al., 2001; Parameswaran and Chellappa, 2006; Shen and Foroosh, 2008; Ogale et al., 2006; Li et al., 2007; Weinland et al., 2007; Yan et al., 2008; Farhadi and Tabrizi, 2008; Liu et al., 2011) are proposed to make action recognition robust to the change of camera views. They will be discussed in Section 6.3.

6.1. Correspondence-free multi-camera activity analysis

Wang et al. (2010) propose an approach of jointly modeling activities in multiple camera views using a topic model and a trajectory network without requiring solving the challenging correspondence problem. It is assumed that the cameras are synchronized but uncalibrated, and the topology of their fields of views is unknown and arbitrary. Objects are tracked in each camera view independently, however, without inter-camera tracking. The goal is to learn the model of an activity category with distributions in all the camera views and to cluster trajectories in all the camera views without supervision. An example is shown in Fig. 6.

As shown in Fig. 7, a network is built by connecting trajectories observed in different camera views based on their temporal extents. Each node on the network is a trajectory. If two trajectories are observed in different camera views and their temporal extents are close, they are connected by an edge. An edge on the network indicates a possible correspondence candidate only based on the temporal information of trajectories. However, this network does not solve the correspondence problem, because many edges are actually false correspondences.

Topic models (Hofmann, 1999; Blei et al., 2003) were originally proposed for document analysis. Under topic models, words such as "professor" and "university" which often co-occur in the same documents, are clustered into one topic such as "education". In (Wang et al., 2010), trajectories are treated as documents, observations on trajectories are treated as words, and activity classes are treated as topics. Observations are quantized into words according to their locations and moving directions. Each activity class has a joint distribution over locations and moving directions in all the camera views, and corresponds to a path commonly taken by objects. Only considering single camera views separately, if two word values, which are indices of locations and moving directions, often co-occur on the same trajectories (documents), they are on the same path (belonging to the same topic). Trajectories passing through the same paths belong to the same activity classes. Therefore, the models of activities can be learned in single camera views according to the tracking information with topic models. In (Wang et al., 2010), it is further assumed that if two trajectories in different camera views are connected by an edge on the network, which means that they may correspond to the same object since they are observed by cameras around the same time, they tend to have similar distributions over activities. Thus based on such temporal correlation, the distributions of an activity class (the path of objects) in different camera views can be jointly modeled.

An example is shown in Fig. 7(d). Trajectories a and b are observed in different camera views and connected by an edge. Points on trajectories are assigned to activity classes by fitting the models

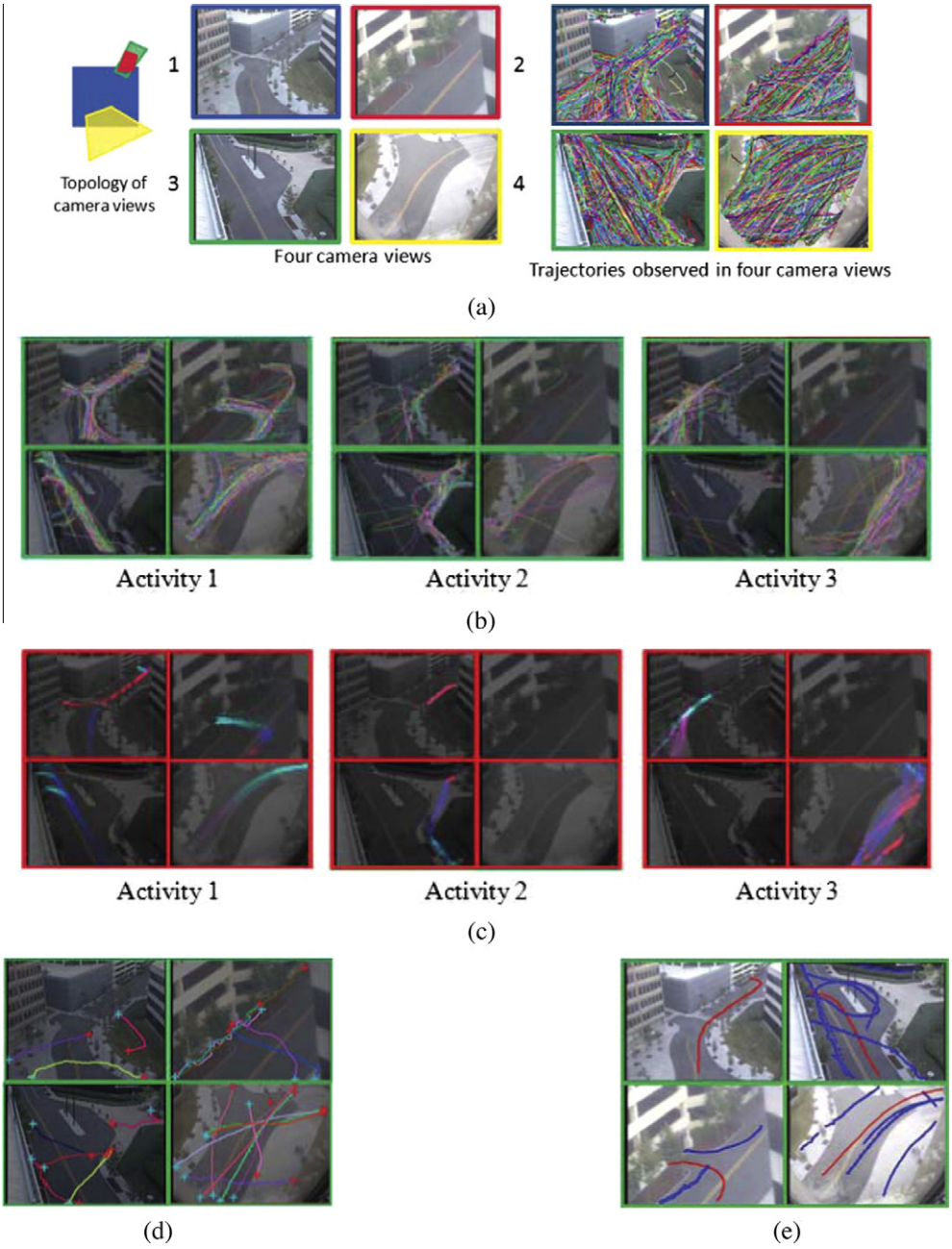

Fig. 6. (a) Four camera views, their topology and the trajectories observed in these camera views. (b) All the trajectories are clustered into different activity categories using the approach proposed in (Wang et al., 2010). In each cluster, the trajectories are observed in different camera views but belong to the same activities. For example, in activity 1, vehicles move first from the top-right to the bottom-left along the road observed in camera view 4, and then they move upward along the road in camera view 1. Some of them enter the parking lot as observed in camera view 2. Some continue to move along the road as observed in camera view 3. In activity 3, pedestrians walk along the sidewalk and are observed in camera views 1 and 4. (c) The models of activities learned without supervision. Each model has a joint distribution over positions and moving directions in all the four camera views. (d) The detected abnormal trajectories which do not fit any of the learned activity models. (e) A trajectory of activity 1 is observed in the view of camera 1. Around the same time, trajectories in other camera views are observed and plotted. The red trajectories all belong to activity 1 shown in (c). Thus it is more likely for them to be the same object.

of activities. These models in each of the individual camera views can be learned using the topic model according to the tracking information within single camera views. However, the goal is to learn the joint distribution of each activity model in all the camera views. The smoothness constraint requires that the distributions of a and b over activities are similar in order to have a smaller penalty. In this example, both a and b have a larger distribution on activity 1, so the models of activity 1 in the two different camera views can be associated.

Many public places of high security interest are extremely crowded. It is difficult to accurately detect and track objects in such environments. In recent years, many activity analysis approaches (Wang et al., 2007, 2009; Loy et al., 2009) have been proposed for video surveillance in crowded environments without tracking objects. Loy et al. (2009) propose an approach of activity analysis with multiple non-overlapping and uncalibrated camera views in a busy scene without intra- or inter- camera tracking. Activities are represented as features of local motions. They decompose each camera

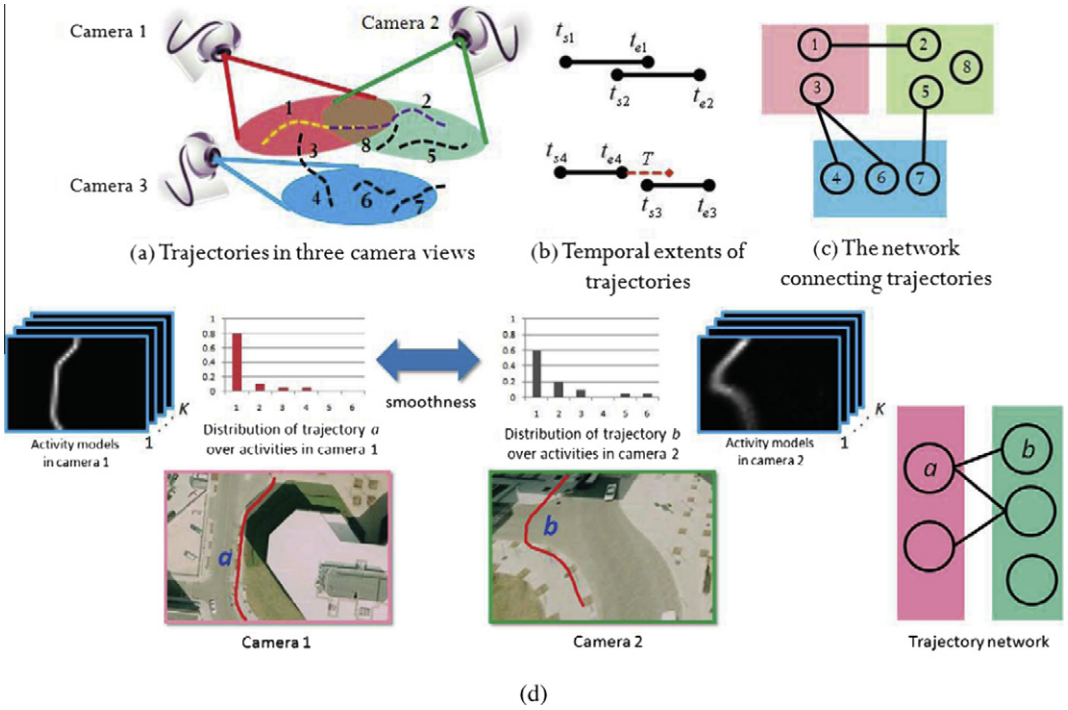

Fig. 7. Correspondence-free activity analysis. (a)–(c): Example of building a network connecting trajectories observed in multiple camera views. (d) Example to describe the high-level picture of the model proposed in (Wang et al., 2010).

view into semantic regions according to the similarity of local spatio-temporal motion patterns as shown in Fig. 8(b). The temporal and causal relationships between activities of semantic regions within and across camera views are detected and quantified using Cross Canonical Correlation Analysis. The proposed approach can automatically infer the topology of the semantic regions as well as the camera network (as shown in Fig. 8(c) and (d)), and can model the global activities over the whole camera network by linking visual evidence collected in multiple camera views.

6.2. Using activity models to improve tracking and object re-identification across camera views

As discussed above, the models of activities in all the camera views can be learned without correspondence among trajectories in an unsupervised way. Once they are learned, they can be used to solve the correspondence problem by providing prior information. If two trajectories belong to the same activity category, it is more likely for them to be the same object. An example is shown

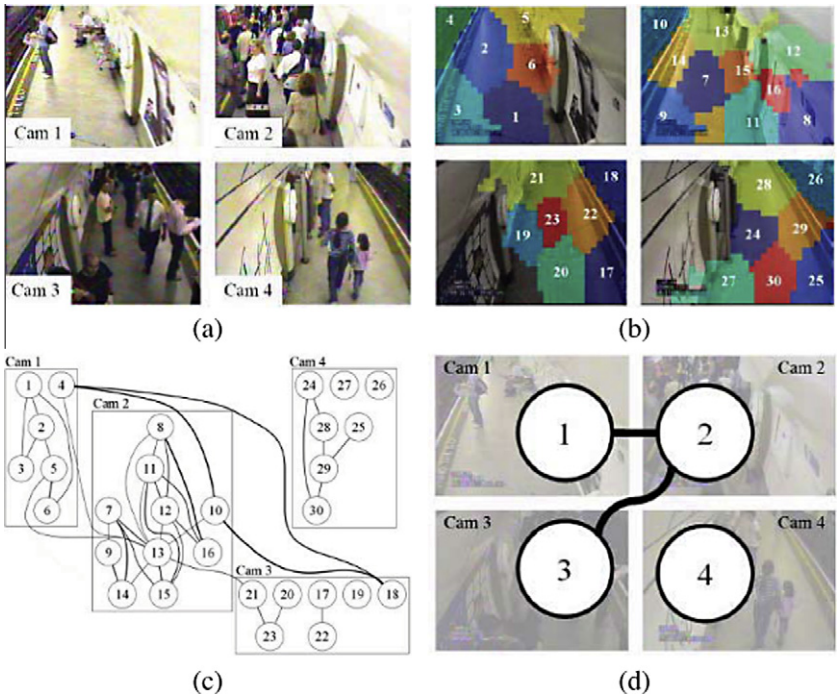

Fig. 8. Example of multi-camera activity correlation analysis (Loy et al., 2009). (a) Four non-overlapping camera views. (b) Semantic regions segmented within each of the camera views. (c) The inferred topology of semantic regions. (d) The inferred topology of camera views. The figure is reproduced from Loy et al. (2009).

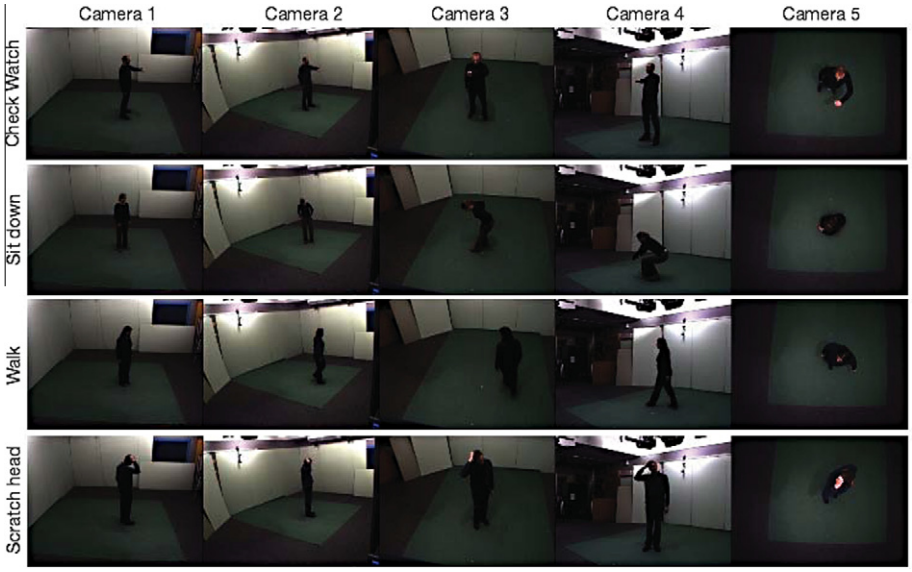

Fig. 9. Examples of human actions observed in five different camera views from the IXMAS multi-view data set (Weinland et al., 2006).

in Fig. 6(e). So the information on activity categories can dramatically reduce the search space when solving the correspondence problem. In (Wang et al., 2010), the distance between two trajectories is defined as the Jensen-Shannon divergence of their distributions over activity categories. The correspondence problem is solved by the Hungarian algorithm (Kuhn, 1956). Berclaz et al. (2008) integrate activity models into a multi-camera tracking system in order to improve the tracking performance. Each of the activity models is represented by a *behavioral map* that encodes, for each ground plane location, the probability of an object moving into one of the adjacent positions at the next frame. The probability of an object switching between different behavior maps (i.e. activity models) is also modeled. The behavior maps are combined with the multi-people tracking algorithm proposed in (Fleuret et al., 2007) under HMM. The multi-camera activity correlation analysis proposed in (Loy et al., 2009) can improve object re-identification across camera views by providing the contextual information of the temporal and causal relationships between regional activities. It effectively reduces the search space and resolve the ambiguities among objects with similar appearance.

6.3. Human action recognition in multiple camera views

Multi-camera activity analysis in near-fields faces the great challenge that the change of viewpoints causes large variations both on appearance and motions of human actions. Some examples are shown in Fig. 9. Most research effort has been made along two directions: (1) proposing features which are invariant to the variation of view points; and (2) reducing the gap between viewpoints through learning.

Various viewpoint invariant features are proposed for human action recognition. Many of them are based on trajectories extracted from human bodies (Rao et al., 2002; Parameswaran and Chellappa, 2006; Shen and Foroosh, 2008; Yilmaz and Shah, 2005; Syeda-Mahmood et al., 2001). Rao et al. (2002) first track human body parts (such as hands) and then use the spatio-temporal curvatures of 2-D trajectories as features, which capture the dramatic changes in speed and direction of actions. Parameswaran and Chellappa (2006) track body joints and find a set of *canonical poses* where at least five body joints are approximately aligned on the same plane. For each canonical pose, two view-invariants are computed. The periodic occurrences of canonical poses and the dynamic trajectories in a view-invariance space are used as the representation for action recognition. Shen and Foroosh (2008) represent an action as a set of *pose transitions* defined by a set of triplets of body joints. Each triplet forms a moving plane observed by a fixed camera and it can be characterized by a fundamental matrix across frames. It is shown that some ratios among the elements in the fundamental matrix are invariant to view points and can be used to match plane motions across camera views. In (Yilmaz and Shah, 2005; Syeda-Mahmood et al., 2001), trajectories of landmark points are extracted. Through computing correspondence of landmarks, fundamental matrix constraints are imposed for matching actions in a stationary camera view and a moving camera view. However, in these approaches the requirement of accurately tracking body parts, joints or landmarks under different viewpoints is challenging. Besides trajectories, other feature representations also can be used such as silhouettes (Weinland et al., 2007), and self-similarities (Junejo et al., 2008, 2011). Weinland et al. (2007) fully reconstruct the 3D models of human actions from silhouettes seen from multiple cameras using an exemplar-based HMM (Frey et al., 2000). At the recognition stage, actions observed from a single camera view can be efficiently recognized without information of the viewpoint as *a priori*. The parameters of viewpoints are estimated as latent variables. Yan et al. (2008) develop a 4D action shape model, which is a sequence of 3D shapes constructed from multi-view silhouette sequences. Spatio-temporal action features are computed by analyzing differential geometric properties of the 4D shapes. Obtaining silhouettes requires background segmentation which is difficult in cluttered scenes or with moving cameras. Junejo et al. (2008, 2011) propose an action descriptor that captures the structure of temporal similarities and dissimilarities in a video sequence based on the observation that self-similarities of action sequences over time show stability under the changes of viewpoints. The self-similarity is computed from pairwise distances between image features in different frames. It does not require tracking or background subtraction.

Liu et al. (2011) propose a transfer learning framework for human action recognition across camera views. Many approaches model an action as a bag of visual words in each of the two camera views (Liu et al., 2009). Such a feature representation is sensitive to view changes. Therefore, some higher level features which can be shared across camera views are further learned in (Liu et al., 2011). A bipartite graph is built to model two view-dependent vocabularies, and then the two vocabularies are co-clustered into

visual-word clusters called bilingual-words, which are the representations of high-level features, through bipartite graph partitioning based on the co-occurrence of visual words in the training videos. A bag-of-bilingual-words is used to represent an action for recognition. It bridges the semantic gap between view-dependent vocabularies. Also under the transfer learning framework, Farhadi and Tabrizi (2008) employ Maximum Margin Clustering (Xu et al., 2004) to generate split-based features in a source view, and then a predictor is trained to predict split-based features in the target view using unlabeled but temporally aligned training video pairs in both the source view and the target view. The split-based features are transferable across views in this way. The drawback of these approaches is that the learned feature representations are only applicable to a fixed pair of camera views. When the viewpoints change, they have to be trained again. Weinland et al. (2010) handle viewpoint changes by learning classifiers on training examples taken from various views without being limited to fixed viewpoints.

7. Cooperative video surveillance with static and active cameras

Many techniques discussed above are applied to static cameras. With a limited number of static cameras to monitor a large area, the observed objects are often small in size and there exist gaps between camera views. By including active cameras, whose panning, tilting and zooming (PTZ) parameters are automatically and dynamically controlled by the systems, the performance of video surveillance can be significantly improved (Collins et al., 2001, 2002; Matsuyama and Ukita, 2002; Gonzalez-Galvan et al., 2002; Kurihara et al., 2002; Naish et al., 2003; Bakhtari et al., 2009). The hybrid systems both with static and active cameras can observe a wider area with a smaller number of cameras by constantly changing the fields of views of the active cameras according to a scanning plan (Sakane et al., 1987; Levit et al., 1992; Ye and Tsotsos, 1999; Matsuyama et al., 1999; Marcenaro et al., 2000; Gonzalez-Galvan et al., 2002). Once objects of security interest are detected, their images can be captured with higher resolutions by automatically zooming in of the active cameras (Woo and Capson, 2000; Izo et al., 2007). It also allows to online choose the optimal viewpoints for object detection, tracking and recognition (Cowan and Kovesi, 1988; Tarabanis et al., 1990; Kim et al., 1995; Piexoto et al., 2000). However, the complexity of the hybrid systems also considerably increases. They have to face some new challenges, some of which are briefly mentioned below. Some of the topics are closely related to the research areas of Active Vision (Bajcsy and passive perception, 1985; Aloimonos et al., 1988; Aloimonos, 1993; Blake and Yuille, 1993; Bakhtari et al., 2009) and Vision-Based Robot Control (Agin et al., 1979; Weiss et al., 1987; Hutchinson et al., 1996; Chaumette et al., 2006, 2007).

1. Online calibration of active cameras static cameras (Chen et al., 2009). It requires high efficiency and no human intervention is allowed at the online stage.
2. Background modeling of active cameras (Kang et al., 2003; Azzari et al., 2005; Bevilacqua and Azzari, 2006, 2007; Sankaranarayanan and Davis, 2008). Background subtraction (Piccardi, 2004) is widely used for detecting moving objects in video surveillance with static cameras. However, it becomes more challenging for active cameras whose background constantly change because of camera motions.
3. Designing a scanning plan according to which active cameras navigate the environments until objects of interest are detected (Davis et al., 2006).
4. Coordinating active cameras and static cameras to improve tracking performance in terms of minimizing the cost and maximizing the accuracy (Matsuyama and Ukita, 2002; Micheloni et al., 2005; Bakhtari et al., 2007, 2009). In order to keep tracking an object without breaks, one camera needs to hand over the object to another camera. In order to minimize the amount of data to be processed, a sensing strategy needs to dynamically activate an optimal subset of cameras in response to the motion of objects to serve the tracking purpose. Sensing strategies also need to be planned to maneuver the cameras into optimal poses and to reduce the uncertainty of tracking. The coordination methods take into account both object motion characteristics and mechanical camera dynamics.

There are also other issues, such as activity analysis with hybrid cameras (Singh and Atrey, 2008) and visualization of videos and analysis results (Morison et al., 2009) to be considered. Some detailed discussions are provided in the sections below.

7.1. Background modeling of active cameras

Most approaches for background modeling of active cameras compute a mosaic of the background scene. A mosaic background is a compound image built through aligning a large number of frames captured by the active camera when it is hinged and freely rotates around its optical center and transforming them onto a common reference plane according to the geometric model of the active camera. When a new frame is captured online, it is registered to the mosaic background and moving objects are detected by comparing their differences. Computational cost is one of the major concerns for realtime applications, where online image registration is time consuming. Some approaches simplify the geometric transform model from projective transform to rigid or affine transform and utilize the pan/tilt/zoom information to speed up registration (Winkelman and Patras, 2004; Hayman and Eklundh, 2003). Various approaches are proposed to reduce the registration errors (Bhat et al., 2000; Bartoli et al., 2002; Bevilacqua et al., 2005). Since frames may be captured under different lighting conditions, they need photometric calibration when being composed together (Mann, 1996; Tsin et al., 2001; Capel et al., 2001).

7.2. Object tracking with active cameras

Object tracking with an active camera involves two iterative steps: *perception* and *action*. The perception step uses the PTZ parameters of the camera obtained from the action step to update the background model and locate the moving object. The action step uses the object location obtained from the perception step to control the camera (Murray and Basu, 1994). An active camera can continually track an object while keeping it centered in the camera view. Therefore, there is a need to efficiently map image pixel coordinates from tracking to their pan-tilt orientations in the word coordinate system in order to adjust the camera to a new location which corresponds to the centroid of the object (Sankaranarayanan and Davis, 2008).

Since multiple objects move freely in the scene, the surveillance system has to adaptively determine which cameras should track which objects considering the dynamic behaviors of objects and the current sates of cameras. This real-time dynamic resource allocation is solve as an optimization problem (Tarabanis et al., 1995; Miura and Ikeuchi, 1998). The optimization criterion includes maximizing the visibility measure (i.e. the tracked objects are less occluded in the camera views) (Bakhtari et al., 2006; Mackay and Benhabib, 2008), maximizing the distinctiveness of the appearance of objects from the background (Snidaro et al., 2003), the importance of objects (*i.e.*, objects of higher security interest have a higher priority to obtain the resource) (Izo et al., 2007), and minimizing the changes of cameras' positions from one time to the next (Sakane et al., 1987). The number and the types of cameras utilized

for data acquisition and the optimal positions and orientations of cameras need to be decided as well (Bakhtari et al., 2009).

8. Discussion and conclusions

By employing distributed camera networks, video surveillance systems substantially extend their capabilities and improve their robustness through data fusion and cooperative sensing. With multi-camera surveillance systems, activities in wide areas are analyzed, the accuracy and robustness of object tracking are improved by fusing data from multiple camera views, and one camera handovers objects to another camera to realize tracking over long distances without break. As the sizes and complexities of camera networks fast increase, there are higher requirements on the robustness, reliability, scalability, transferability, self-adaptability and less human intervention of intelligent multi-camera video surveillance systems. The discussed computer vision and pattern recognition technologies are closely related to each other. While most conventional surveillance systems assume one directional information flow, recent studies show that different modules actually can support each other. For example, activity modeling can improve inter-camera tracking and multi-camera tracking provides information for camera calibration and inference of the topology of camera views. Jointly solving some of these problems not only improves the robust and accuracy but also reduces human intervention. Jointly modeling these problems at different hierarchical levels in more principled ways is an important issue for further investigation. Significant progress on intelligent multi-camera video surveillance has been achieved in recent years. While some problems, such as calibrating camera views with significant overlap and computing their topology, have been well studied, some need more research effort in the future. It is still challenging to calibrate camera views which are disjoint and where objects move on multiple ground planes. Object re-identification is relatively new and its performance is still far from satisfactory. The accuracy of state-of-the-art is below 20% on the VIPeR dataset (Gray et al., 2007). This bring challenges for inter-camera tracking when spatio-temporal reasoning is unreliable and it has to more rely on appearance matching. Video surveillance in crowded environments started to draw a lot of attention in the past five years because it is very challenging and highly valuable for public security. Most existing works on this topic assume a single camera view. Although it is well known that multi-camera surveillance systems can better solve occlusions and scene clutters, not much research work has been done on designing the topology of camera networks and cooperating hybrid cameras to avoid occlusions in extremely crowded environments. Most published results on camera calibration, inference of topology, object re-identification, tracking and activity analysis are based on small camera networks. However, larger scale camera networks are needed for future research. Both benchmark datasets and comprehensive experimental evaluations on very large scale camera networks are needed in the future research. In conclusion, this paper reviews some key computer vision and pattern recognition technologies utilized in intelligent multi-camera video surveillance and emphasizes their connections and integration. The most recent development of these technologies is discussed and different solutions are compared. It provides detailed descriptions of major challenges for each of the key technologies. We believe that this review will encourage new research work in the fast growing area.

Acknowledgements

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Projects Nos. CUHK417110 and CUHK417011) and National Natural Science Foundation of China (Project No. 61005057).

References

Abdel-Hakim, A.E., Farag, A.A., 2006. Csift: A sift descriptor with color invariant characteristics. In: Proc. European Conf. Computer Vision.
Agarwal, A., Triggs, B., 2006. Hyperfeatures – multilevel local coding for visual recognition. In: Proc. European Conf. Computer Vision.
Aghajan, H., Cavallaro, A. (Eds.), 2009. Multi-Camera Networks: Concepts and Applications. Elsevier.
Agin, G.J., 1979. Real Time Control of a Robot with a Mobile Camera, Tech. Rep. SRI International Technical Note.
Agrawal, M., Davis, L., 2003. Complete camera calibration using spheres: Dual space approach. In: Proc. IEEE Internat. Conf. Computer Vision.
Alexander, H., Lucchesi, C., xxxx. Matching algorithms for bipartite graphs. Relatorio Tecnico 700.
Aloimonos, Y. (Ed.), 1993. Active Perception. Lawrence Erlbaum Associates Publisher.
Aloimonos, J., Weiss, I., Bandyopadhyay, A., 1988. Active vision. Int. J. Comput. Vision 1, 333–356.
Antone, M., Bosse, M., 2004. Calibration of outdoor cameras from cast shadows. In: Proc. IEEE Internat. Conf. Systems, Man and Cybernetics.
Azzari, P., Stefano, D.L., Bevilacqua, A., 2005. An effective real-time mosaicing algorithm apt to detect motion through background subtraction using a ptz camera. In: Proc. Advanced Video and Signal Based Surveillance.
Bajcsy, R., 1985. Active perception vs. passive perception. In: Proc. IEEE Workshop on Computer Vision: Representation and Control.
Baker, P., Aloimonos, Y., 2003. Calibration of a multicamera network. In: Proc. Omnivis 2003: Omnidirectional Vision and Camera Networks.
Bakhtari, A., Benhabib, B., 2007. An active vision system for multitarget surveillance in dynamic environments. IEEE Trans. Syst. Man Cybernet. 37, 190–198.
Bakhtari, A., Naish, M.D., Eskandari, M., Croft, E.A., Benhabib, B., 2006. Active-vision-based multisensor surveillance – An implementation. IEEE Trans. Syst. Man Cybernet. 36, 668–680.
Bakhtari, A., Mackay, M., Benhabib, B., 2009. Active-vision for the autonomous surveillance of dynamic, multi-object environments. J. Intell. Robot Syst. 54, 567–593.
Bartoli, A., Dalal, N., Bose, B., Horaud, R., 2002. From video sequences to motion panoramas. In: Proc. IEEE Workshop on Motion and Video Computing.
Bay, H., Tuytelaars, T., Gool, L.V., 2006. Surf: Speed up robust features. In: Proc. European Conf. Computer Vision.
Beardsley, P., Murray, D., 1992. Camera calibration using vanishing points. In: Proc. British Machine Vision Conf.
Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Machine Intell. 24, 509–512.
Berclaz, J., Fleuret, F., Fua, P., 2008. Multi-camera tracking and atypical motion detection with behavioral maps. In: Proc. European Conf. Computer Vision.
Bevilacqua, A., Azzari, P., 2006. High-quality real time motion detection using ptz cameras. In: Proc. Advanced Video and Signal Based Surveillance.
Bevilacqua, A., Azzari, P., 2007. A fast and reliable image mosaicing technique with application to wide area motion detection. In: Proc. Internat. Conf. Image Analysis and Recognition.
Bevilacqua, A., Stefano, L.D., Azzari, P., 2005. An effective real-time mosaicing algorithm apt to detect motion through background subtraction using a ptz camera. In: Proc. Advanced Video and Signal Based Surveillance.
Bhat, K.S., Saptharishi, M., Khosla, P.K., 2000. Motion detection and segmentation using image mosaics. In: Proc. IEEE Internat. Conf. Multimedia and Expo.
Bird, N., Masoud, O., Papanikolopoulos, N., Issacs, A., 2005. Detection of loitering individuals in public transportation areas. IEEE Trans. Intell. Transport. Syst. 6, 167–177.
Black, J., Ellis, T.J., Rosin, P., 2002. Multi view image surveillance and tracking. In: Proc. IEEE Workshop on Motion and Video Computing.
Blake, A., Yuille, A., 1993. Active Vision. MIIT Press..
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Machine Learn. Res. 3, 993–1022.
Bobick, A.F., Ivanov, Y.A., 1998. Action recognition using probabilistic parsing. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Bose, B., Grimson, E., 2003. Ground plane rectification by tracking moving objects. In: Proc. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.
Brand, M., Kettnaker, V., 2000. Discovery and segmentation of activities in video. IEEE Trans. Pattern Anal. Machine Intell. 22, 844–851.
Brown, M., Lowe, D., 2003. Recognising panoramas. In: Proc. IEEE Internat. Conf. Computer Vision.
Cai, Q., Aggarwal, J.K., 1996. Tracking human motion in structured environments using a distributed-camera system. IEEE Trans. Pattern Anal. Machine Intell. 21, 1241–1247.
Cao, X., Foroosh, H., 2006. Camera calibration and light source orientation from solar shadows. In: Journal of Computer Vision and Image Understanding.
Capel, D.P., 2001. Image Mosaicing and Super-resolution, Ph.D. Thesis. University of Oxford.
Caprile, B., Grimson, V., 1990. Using vanishing points for camera calibration. Internat. J. Comput. Vision 4, 127–139.

Carneiro, G., Lowe, D., 2006. Sparse flexible models of local features. In: Proc. European Conf. Computer Vision.
Caspi, Y., Irani, M., 2000. A step towards sequence-to-sequence alignment. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Caspi, Y., Simakov, D., Irani, M., 2006. Feature-based sequence-to-sequence matching. Internat. J. Comput. Vision 68, 53–64.
Chang, T.H., Gong, S., 2001. Tracking multiple people with a multi-camera system. In: Proc. IEEE Internat. Conf. Computer Vision.
Chaumette, F., Hutchinson, S., 2006. Visual servo control, Part I: Basic approaches. IEEE Robot. Automat. Mag. 13, 82–90.
Chaumette, F., Hutchinson, S., 2007. Visual servo control, Part ii: Advanced approaches. IEEE Robot. Automat. Mag. 14, 109–118.
Chen, K., Lai, C., Hung, Y., Chen, C., 2008. An adaptive learning method for target tracking across multiple cameras. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition, 2008.
Chen, C., Yao, Y., Dira, A., Koschan, A., Abidi, M., 2009. Cooperative mapping of multiple ptz cameras in automated surveillance systems. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Cheng, E.D., Piccardi, M., 2006. Matching of objects moving across disjoint cameras. In: Proc. IEEE Internat. Conf. Image Processing.
Cipolla, R., Drummond, T., Drummond, D.P., 1999. Camera calibration from vanishing points in images of architecural scenes. In: Proc. British Machine Vision Conf.
Collins, R.T., Lipton, A.J., Fujiyoshi, H., Kanade, T., 2001. Algorithms for cooperative multisensor surveillance. Proc. IEEE 89, 1456–1477.
Collins, R., Amidi, O., Kanade, T., 2002. An active camera system for acquiring multi-view video. In: Proc. IEEE Internat. Conf. Image Processing.
Colombo, C., Bimbo, A.D., Pernici, F., 2005. Metric 3d reconstruction and texture acquisition of surfaces of revolution from a single uncalibrated view. IEEE Trans. Pattern Anal. Machine Intell. 27, 99–114.
Cowan, C., Kovesi, P., 1988. Automatic sensor placement from vision task requirements. IEEE Trans. Pattern Anal. Machine Intell. 10, 407–416.
Cox, I.J., Hingorani, S.L., 1994. An efficient implementation and evaluation of reid's multiple hypothesis tracking algorithm for visual tracking. In: Proc. IEEE Internat. Conf. Pattern Recognition.
Cozman, F., Krotkov, E., 1997. Automatic mountain detection and pose estimation for teleoperation of lunar rovers. In: Proc. IEEE Internat. Conf. Robotics and Automation.
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Daugman, J.G., 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J. Opt. Soc. Am. A 2, 1160–1169.
Davis, J.W., Morison, A.M., Woods, D.D., 2006. Building adaptive camera models for video surveillance. In: Proc. Advanced Video and Signal Based Surveillance.
Deutscher, J., Isard, M., MacCormick, J., 2002. Automatic camera calibration from a single manhttan image. In: Proc. European Conf. Computer Vision.
Dietterich, T.G., Lathrop, R.H., Lozano-Perez, T., 1997. A pharmaceutical, solving the multiple-instance problem with axis-parallel rectangles. Artif. Intell. 89, 31–71.
Dockstader, S.L., Tekalp, A.M., 2001. Multiple camera tracking of interacting and occluded human motion. Proceedings of IEEE 89, 1441–1455.
Ellis, T.J., Makris, D., Black, J., 2003. Learning a multicamera topology. In: Proc. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.
Eshel, R., Moses, Y., 2008. Homography based multiple camera detection and tracking of people in a dense crowd. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010. Person re-identification by symmetry-driven accumulation of local features. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Farhadi, A., Tabrizi, M.K., 2008. Learning to recognize activities from the wrong view point. In: Proc. European Conf. Computer Vision.
Faugeras, O., 1993. Three Dimensional Computer Vision: A Geometric Viewpoint. MIT Press.
Faugeras, O., Luong, Q.T., 2001. The Geometry of Multiple Images. MIT Press.
Fleuret, F., Berclaz, J., Lengagne, R., Fua, P., 2007. Multi-camera people tracking with a probabilistic occupancy map. IEEE Trans. Pattern Anal. Machine Intell. 30, 267–282.
Fleuret, F., Berclaz, J., Lengagne, R., Fua, P., 2008. Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. Pattern Anal. Machine Intell. 30, 267–282.
Focken, D., Stiefelhagen, R., 2002. Towards vision-based 3d people tracking in a smart room. In: Proc. IEEE Internat. Conf. Multimodal Interfaces.
Forssen, P.E., 2007. Maximally stable colour regions for recognition and matching. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Frey, B.J., Jojic, N., 2000. Learning graphical models of images, videos and their spatial transformations. Proc. Uncertainty Artif. Intell., 184–191.
Fu, Z., Hu, W., Tan, T., 2005. Similarity based vehicle trajectory clustering and anomaly detection. In: Proc. IEEE Internat. Conf. Image Processing.
Gheissari, N., Sebastian, T.B., Rittscher, J., Hartley, R., 2006. Person reidentification using spatiotemporal appearance. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Gilbert, A., Bowden, R., 2006. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In: Proc. European Conf. Computer Vision.
Gonzalez-Galvan, E.J., Pazos-Flores, F., Skarr, S., Cardebas-Galindo, A., 2002. Camera pan/tilt to eliminate the workspace-size/pixel-resolution tradeoffs with camera-space manipulation. Robotics Comput. Integrated Manufact. 18, 95–104.
Gray, D., Tao, H., 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proc. European Conf. Computer Vision.
Gray, D., Brennan, S., Tao, H., 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.
Guo, Y., Shan, Y., Sawhney, H., Kumar, R., 2007. Peet: Prototype embedding and embedding transition for matching vehicles over disparate viewpoints. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Guo, Y., Rao, C., Samarasekera, S., Kim, J., Kumar, R., Sawhney, H., 2008. Matching vehicles under large pose transformations using approximate 3d models and piecewise mrf model. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B., 2008. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: Proc. IEEE Conference on Distributed Smart Cameras.
Hamid, R., Kumar, R.K., Grundmann, M., Kim, K., Essa, I., Hodgins, J., 2010. Player localization using multiple static cameras for sports visualization. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Haritaoglu, I., Harwood, D., Davis, L.S., 2000. W4: Real-time surveillance of people and their activities. IEEE Trans. Pattern Anal. Machine Intell. 22, 809–830.
Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: Proc. Alvey Vision Conference.
Hartley, R.I., Zisserman, A., 2004. Multiple View Geometry in Computer Vision. Cambridge University Press.
Hayman, E., Eklundh, J., 2003. Statistical background subtraction for a mobile observer. In: Proc. IEEE Internat. Conf. Computer Vision.
Heikkila, J., 2000. Geometric camera calibration using circular control points. In: IEEE Trans. Pattern Analysis and Machine Intell.
Hofmann, T., 1999. Probabilistic latent semantic analysis. In: Proc. Uncertainty in Artificial Intelligence.
Honggeng, S., Nevatia, R., 2001. Multi-agent event recognition. In: Proc. IEEE Internat. Conf. Computer Vision.
Hopcroft, J., Karp, R., 1973. An $n^2.5$ algorithm for maximum matchings in bipartite graphs. SIAM J. Computing.
Huang, T., Russell, S., 1997. Object identification in a bayesian context. In: Proc. Internat. Joint Conf. Artificial Intelligence.
Huang, J., Kumar, S.R., Mitra, M., Zhu, M., Zabih, R., 1997. Image indexing using color correlograms. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Trans. Systems, Man, Cybernetics – Part C: Applications and Reviews 34, 334–352.
Hu, W., Hu, M., Zhou, X., Tan, T., Luo, J., Maybank, S., 2006. Principal axis-based correspondence between multiple cameras for people tracking. IEEE Trans. Pattern Anal. Machine Intell. 28, 663–671.
Hutchinson, S.A., Hager, G.D., Corke, P.I., 1996. A tutorial on visual servo control. IEEE Trans. Robotics Automat. 12, 651–675.
Izo, T., 2007. Visual Attention Models for Far-field Scene Analysis, Ph.D. Thesis. MIT.
Jannotti, J., Mao, J., 2006. Distributed calibration of smart cameras. In: Proc. Workshop on Distributed Smart Cameras.
Javed, O., Rasheed, Z., Shafique, K., Shah, M., 2003. Tracking across multiple cameras with disjoint views. In: Proc. IEEE Internat. Conf. Computer Vision.
Javed, O., Shafique, K., Shah, M., 2005. Appearance modeling for tracking in multiple non-overlapping cameras. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Jiang, H., Fels, S., Little, J., 2007. A linear programming approach for multiple object tracking. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Johnson, N., Hogg, D., 1995. Learning the distribution of object trajectories for event recognition. In: Proc. British Machine Vision Conf.
Jones, G.A., Renno, J.R., Remagnino, P., 2002. Auto-calibration in multiple-camera surveillance environments. In: Proc. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.
Junejo, I., Foroosh, H., 2008. Estimating geo-temporal location of stationary cameras using shadow trajectories. In: Proc. European Conf. Computer Vision.
Junejo, I., Javed, O., Shah, M., 2004. Multi feature path modeling for video surveillance. In: Proc. IEEE Internat. Conf. Pattern Recognition.
Junejo, I., Dexter, E., Laptev, I., 2008. Cross-view action recognition from temporal self-similarities. In: Proc. European Conf. Computer Vision.
Junejo, I., Dexter, E., Laptev, I., Perez, P., 2011. Cross-view action recognition from temporal self-similarities. IEEE Trans. Pattern Anal. Machine Intell. 33, 172–185.
Kang, S., Paik, J., Koschan, A., Abidi, B., Abidi, M.A., 2003. Real-time video tracking using ptz cameras. In: Prof. SPIE Internat. Conf. Quality Control by Artificial Vision.
Keogh, E., Pazzani, M., 2000. Scaling up dynamic time scaling up dynamic time. In: Proc. SIGKDD.
Kettnaker, V., Zabih, R., 1999. Bayesian multi-camera surveillance. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Khan, S., Shah, M., 2003. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Trans. Pattern Anal. Machine Intell. 25, 1355–1360.

Khan, S.M., Shah, M., 2006. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: Proc. European Conf. Computer Vision.
Kim, K., Davis, L., 2006. Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In: Proc. European Conf. Computer Vision.
Kim, Y., Lee, K., Lim, C., Park, K., 1995. A study of the implementation of moving object tracking. In: Proc. SPIE Conf. Visual Communications and Image Processing.
Krahnstoever, N., Mendonca, P.R.S., 2005. Bayesian autocalibration for surveillance. In: Proc. IEEE Internat. Conf. Computer Vision.
Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S., 2000. Multi-camera multi-person tracking for easyliving. In: Proc. IEEE Workshop on Visual Surveillance.
Kuhn, H.W., 1956. Variants of the hungarian method for assignment problems. Naval Res. Logist. Quart. 3, 253–258.
Kuo, C., Huang, R., 2010. Nevatia, Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In: Proc. European Conf. Computer Vision.
Kurihara, K., Hoshino, S., Yamane, K., Nakamura, Y., 2002. Optical motion capture system with pan-tilt camera tracking and real-time data processing. In: Proc. IEEE Internat. Conf. Robotics and Automation.
Lacey, A.J., Pinitkarn, N., Thacker, N.A., 2000. An evaluation of the performance of ransac algorithms for stereo camera calibration. In: Proc. British Machine Vision Conf.
Lambert, C.G., Harrington, S.E., Harvey, C.R., Glodjo, A., 1999. Efficient on-line nonparametric kernel density estimation. Algorithmica 25, 37–56.
Lazebnik, S., Schmid, C., Ponce, J., 2003. A sparse texture representation using affineinvariant regions. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Lee, L., Romano, R., Stein, G., 2000. Monitoring activities from multiple video streams: Establishing a common coordinate frame. IEEE Trans. Pattern Anal. Machine Intell. 22, 758–768.
Leung, T., Malik, J., 1999. Recognizing surfaces using three-dimensional textons. In: Proc. IEEE Internat. Conf. Computer Vision.
Levit, T., Agosta, J., Binford, T., 1992. Bayesian methods for interpretation and control in multi-agent vision systems. In: Proc. SPIE Conf. Applications of AI X: Machine Vision and Robotics.
Liebowitz, D., Zisserman, A., 1999. Combining scene and auto-calibration constraints. In: Proc. IEEE Internat. Conf. Computer Vision.
Liebowitz, D., Criminisi, A., Zisserman, A., 1999. Creating architecural models from images. In: Proc. Euro-Graphics.
Li, Y., Hilton, A., Illingworth, J., 2002. A relaxation algorithm for real-time multiple view 3d-tracking. Image Vision Comput. 20, 841–859.
Li, R., Tian, T., Sclaroff, S., 2007. Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. In: Proc. IEEE Internat. Conf. Computer Vision.
Lin, Z., Davis, L., 2008. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In: Proc. Internat. Symposium on Advances in Visual Computing.
Liu, J., Yang, Y., Shah, M., 2009. Learning semantic visual vocabuaries using diffusion distance. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Liu, J., Shah, M., Kuipers, B., Savarese, S., 2011. Corss-view action recognition via view knowledge transfer. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Internat. J. Comput. Vision 60 (2), 91–110.
Loy, C., Xiang, T., Gong, S., 2009. Multi-camera activity correlation analysis. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Lu, F., Shen, Y., Cao, X., Foroosh, H., 2005. Camera calibration from two shadow trajectories. In: Proc. IEEE Internat. Conf. Pattern Recognition.
Lv, F., Zhao, T., Nevatia, R., 2002. Self-calibration of a camera from video of a walking human. In: Proc. IEEE Internat. Conf. Image Processing.
Lv, F., Zhao, T., Nevatia, R., 2006. Camera calibration from video of a walking human. IEEE Trans. Pattern Anal. Machine Intell. 28, 1513–1518.
Mackay, M., Benhabib, B., 2008. Active-vision system reconfiguration for form recognition in the presence of dynamic obstacles. Articulated Motion Deformable Objects, 188–207.
Makris, D., Ellis, T., 2002. Path detection in video surveillance. Image Vision Comput. 20, 859–903.
Makris, D., Ellis, T., Black, J., 2004. Bridging the gaps between cameras. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Mann, S., 1996. Pencigraphy with agc: Joint parameter estimation in both domain and range of functions in same orbit of the projective wyckoff group. In: Proc. IEEE Internat. Conf. Image Processing.
Marcenaro, L., Oberti, F., Regazzoni, C., 2000. Change detection methods for automatic scene analysis using mobile surveillance cameras. In: Proc. IEEE Internat. Conf. Image Processing.
Matei, B., Sawhney, H., Samarasekera, S., 2011. Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Matsuyama, T., Ukita, N., 2002. Real-time multitarget tracking by a cooperative distributed vision system. Proc. IEEE 90, 1136–1150.
Matsuyama, T., Wada, T., Tokai, S., 1999. Active image capturing and dynamic scene visualization by cooperative distributed vision. In: Advanced Multimedia Content Processing.
Medioni, G., Cohen, I., BreAmond, F., Hongeng, S., Nevatia, R., 2001. Event detection and analysis from video streams. IEEE Trans. Pattern Anal. Machine Intell. 23, 873–889.
Micheloni, C., Foresti, G.L., Snidaro, L., 2005. A network of co-operative cameras for visual surveillance. IEE Proc. Vision Image Signal Process. 152, 205–212.
Mikic, I., Santini, S., Jain, R., 1998. Video processing and integration from multiple cameras. In: Porc. DARPA Image Understanding Workshop.
Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Machine Intell. 27, 1615–1630.
Mittal, A., Davis, L.S., 2003. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. Internat. J. Comput. Vision 51, 189–203.
Miura, J., Ikeuchi, K., 1998. Task-oriented generation of visual sensing strategies in assembly tasks. IEEE Trans. Pattern Anal. Machine Intell. 20, 126–138.
Morariu, V.I., Camps, O.I., 2006. Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Morison, A.M., Woods, D.D., Davis, J.W., 2009. How panoramic visualization can support human supervision of intelligent surveillance. In: Annual Meeting of the Human Factors and Ergonomics Society.
Morris, B.T., Trivedi, M.M., 2008. A survey of vision-based trajectory learning and analysis for surveillance. IEEE Trans. Circuits Syst. Video Technol. 18, 1114–1127.
Murata, T., 1989. Petri nets: Properties, analysis and applications. Proc. IEEE 77, 541–579.
Murray, D., Basu, A., 1994. Motion tracking with an active camera. IEEE Trans. Pattern Anal. Machine Intell. 19, 449–454.
Naish, M.D., Croft, E.A., Benhabib, B., 2003. Coordinated dispatching of proximity sensors for the surveillance of manoeuvring targets. Robotics Comput. Integrated Manufact. 19, 283–299.
Nakajima, C., Pontil, M., Heisele, B., Poggio, T., 2003. Full-body person recognition system. Pattern Recognition 36, 1977–2006.
Ogale, A., Karapurkar, A., Aloimonos, Y., 2006. View-invariant modeling and recognizing of human actions using grammars. In: Proc. Workshop on Dynamic Vision.
Ojala, T., Pietikäinen, M., Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Machine Intell., 971–987.
Oliver, N., Rosario, B., Pentland, A., 2000. A bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal. Machine Intell. 22, 831–843.
Orwell, J., Remagnino, P., Jones, G.A., 1999. Multiple camera color tracking. In: Proc. IEEE Workshop on Visual Surveillance.
Otsuka, K., Mukawa, N., 2004. Multi-view occlusion analysis for tracking densely populated objects based on 2d visual angles. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Parameswaran, V., Chellappa, R., 2006. View invariance for human action recognition. Int. J. Comput. Vision 66, 83–101.
Park, U., Jain, A., Kitahara, I., Kogure, K., Hagita, N., 2006. Vise: Visual search engine using multiple networked cameras. In: Proc. IEEE Internat. Conf. Pattern Recognition.
Pasula, H., Russell, S., Ostland, M., 1999. Tracking many objects with many sensors. In: Proc. Internat. Joint Conf. Artificial Intelligence.
Perez, P., Vermaak, J., Blake, A., 2004. Data fusion for visual tracking with particles. Proc. IEEE 92, 495–513.
Pflugfelder, R., Bischof, H., 2007. People tracking across two distant self-calibrated cameras. In: Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance.
Pflugfelder, R., Bischof, H., 2010. Localization and trajectory reconstruction in surveillance cameras with nonoverlapping views. IEEE Trans. Pattern Anal. Machine Intell. 32, 709–721.
Piccardi, M., 2004. Background subtraction techniques: a review. In: Proc. IEEE Internat. Conf. Systems, Man and Cybernetics.
Piexoto, P., Batista, J., Araujo, H., 2000. Integration of information from several vision systems for a common task of surveillance. J. Robot Autom. Syst. 31, 99–108.
Porikli, F., 2003. Inter-camera color calibration by correlation model function. In: Proc. IEEE Internat. Conf. Image Processing.
Porikli, F., Divakaran, A., 2003. Multi-camera calibration, object tracking and query generation. In: Proc. IEEE Internat. Conf. Multimedia and Expo.
Prosser, B., Gong, S., Xiang, T., 2008. Multi-camera matching using bi-directional cumulative brightness transfer function. In: Proc. British Machine Vision Conf..
Prosser, B., Zheng, W., Gong, S., Xiang, T., Mary, Q., 2010. Person re-identification by support vector ranking. In: Proc. British Machine Vision Conf.
Rahimi, A., Dunagan, B., Darrell, T., 2004. Simultaneous calibration and tracking with a network of non-overlapping sensors. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Rao, C., Yilmaz, A., Shah, M., 2002. View-invariant representation and recognition of actions. Int. J. Comput. Vision 50, 203–226.
Sakane, S., Sato, T., 1987. Kakikura, Model-based planning of visual sensors using a hand-eye action simulator: Heaven. In: Proc. Internat. Conf. Advanced Robotics.
Salti, S., Tombari, F., Stefano, L.D., 2011. A performance evaluation of 3d keypoint detectors. In: Proc. Internat. Conf. 3D Imaging, Modeling, Processing, Visualization and Transmission.
Sankaranarayanan, K., Davis, J.W., 2008. An efficent active camera model for video surveillance. In: Proc. IEEE Workshop on Applications of Computer Vision.

Savarese, S., Winn, J., Criminisi, A., 2006. Discriminative object class models of appearance and shape by correlatons. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Schwartz, W., Davis, L., 2009. Learning discriminative appearance-based models using partial least sqaures. In: Proc. XXII SIBGRAPI.
Shafique, K., Shah, M., 2003. A non-iterative greedy algorithm for multi-frame point correspondence. IEEE Trans. Pattern Anal. Machine Intell. 27, 51–65.
Shan, Y., Sawhney, H., Kumar, R., 2005. Unsupervised learning of discriminative edge measures for vehicle matching between non-overlapping cameras. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Shan, Y., Sawhney, H., Kumar, R., 2005. Vehicle identification between non-overlapping cameras without direct feature matching. In: Proc. IEEE Internat. Conf. Computer Vision.
Shan, Y., Sawhney, H.S., Kumar, R., 2008. Unsupervised learning of discriminative edge measures for vehicle matching between nonoverlapping cameras. IEEE Trans. Pattern Anal. Machine Intell. 30, 700–711.
Sheikh, Y.A., Shah, M., 2008. Trajectory association across multiple airborne cameras. IEEE Trans. Pattern Anal. Machine Intell. 30, 361–367.
Shen, Y., Foroosh, H., 2008. View invariant action recognition using fundamental ratios. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Singh, V.K., Atrey, P.K., 2008. Cooperitive multi-camera surveillance using model predictive control. Machine Vision Appl. 19, 375–393.
Slater, D., Healey, G., 1996. The illumination-invariant recognition of 3d objects using local color invariants. IEEE Trans. Pattern Anal. Machine Intell. 18, 206–210.
Smith, P., Lobo, N., Shah, M., 2005. Temporalboost for event recognition. In: Proc. IEEE Internat. Conf. Computer Vision.
Snidaro, L., Niu, R., Varsjmey, P.K., Foresti, G.L., 2003. Automatic camera selection and fusion for outdoor surveillance under changing weather conditions. In: Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance.
Sogo, T., Ishiguro, H., 2000. Real-time target localization and tracking by n-ocular stereo. In: Proc. iEEE Workshop on Omnidirectional Vision.
Song, B., Roy-Chowdhury, A.K., 2008. Robust tracking in a camera network: A multi-objective optimization framework. IEEE J. Select. Top. Signal Process. 2, 582–596.
Song, Y., Goncalves, L., Perona, P., 2003. Unsupervised learning of human motion. IEEE Trans. Pattern Anal. Machine Intell. 25, 814–827.
Stauffer, C., 2003. Estimating tracking sources and sinks. In: Proceedings of the Second IEEE Workshop on Event Mining.
Stauffer, C., Grimson, W.E.L., 2000. Learning patterns of activity using real-time tracking. In: IEEE Trans. Pattern Anal. Machine Intell.
Stauffer, C., Tieu, K., 2003. Automated multi-camera planar tracking correspondence modeling. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Stein, G.P., 1999. Tracking from multiple view points: Self-calibration of space and time. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Stein, F., Medioni, G., 1992. Map-based localization using the panoramic horizon. In: Proc. IEEE Internat. Conf. Robotics and Automation.
Straw, A.D., Branson, K., Neumann, T.R., Dickinson, M.H., 2010. Multi-camera realtime 3d tracking of multiple flying animals. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Sturm, P., Maybank, S., 1999. On plane-based camera calibration: a general algorithm, singularities, applications. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Syeda-Mahmood, T., Vasilescu, T., Sethi, S., 2001. Recognizing action events from multiple viewpoints. In: Proc. IEEE Workshop on Detection and Recognition of Events in Video.
Tarabanis, K., Tsai, R., Allen, P., 1990. Analytical characterization of the feature detectability constraints of resolution, focus, and field of view for vision planning. CVGIP: Image Understand. 59, 340–358.
Tarabanis, K.A., Allen, P.K., Tsai, R.Y., 1995. A survey of sensor planning in computer vision. IEEE Trans. Robot. Automat. 11, 84–104.
Teramoto, H., Xu, G., 2002. Camera calibration by a single image of balls: from conic to the absolute conic. In: Proc. Asian Conf. Computer Vision.
Thompson, W., Henderson, T., Colvin, T., Dick, T., Valiquette, C., 1993. Vision-based localization. In: Proc. ARPA Image Understanding Workshop.
Tieu, K., Dalley, G., Grimson, E., 2005. Inference of non-overlapping camera network topology by measuring statistical dependence. In: Proc. IEEE Internat. Conf. Computer Vision.
Triggs, B., 1999. Camera pose and calibration from 4 or 5 known 3d points. In: Proc. IEEE Internat. Conf. Computer Vision.
Tsai, R., 1986. An efficient and accurate camera calibration technique for 3d machine vision. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Tsin, Y., Ramesh, V., Kanade, T., 2001. Statistical calibration of ccd imaging process. In: Proc. IEEE Internat. Conf. Computer Vision.
Tuzel, O., Porikli, F., Meer, P., 2006. Region covariance: A fast descriptor for detection and classification. In: Proc. European Conf. Computer Vision.
Utsumi, A., Mori, H., Ohya, J., Yachida, M., 1998. Multiple-view-based tracking of multiple humans. In: Proc. IEEE Internat. Conf. Pattern Recognition.
Valera, M., Velastin, S.A., 2004. Intelligent distributed surveillance systems: A review. IEE Proc. 152, 193–204.
Varma, M., Zisserman, A., 2005. A statistical approach to texture classification from single images. Internat. J. Comput. Vision 62, 61–81.
Veenman, C., Reinders, M., Backer, E., 2001. Resolving motion correspondence for densely moving points. IEEE Trans. Pattern Anal. Machine Intell. 23, 54–72.
Wang, Y., Jiang, T., Drew, M.S., Li, Z., Mori, G., 2006. Unsupervised discovery of action classes. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Wang, X., Tieu, K., Grimson, W.E.L., 2006. Learning semantic scene models by trajectory analysis. In: Proc. European Conf. Computer Vision.
Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P., 2007. Shape and appearance context modeling. In: Proc. IEEE Internat. Conf. Computer Vision.
Wang, X., Ma, X., Grimson, E., 2007. Unsupervised activity perception by hierarchical bayesian models. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Wang, X., Tieu, K., Grimson, E., 2008. Correspondence-free multi-camera activity analysis and scene modeling. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Wang, X., Ma, K.T., Ng, G., Grimson, E., 2008. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Wang, X., Ma, X., Grimson, W.E.L., 2009. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans. Pattern Anal. Machine Intell. 31, 539–555.
Wang, X., Tieu, K., Grimson, E., 2010. Correspondence-free activity analysis and scene modeling in multiple camera views. IEEE Trans. Pattern Anal. Machine Intell. 32, 56–71.
Wang, X., Ma, K.T., Ng, G., Grimson, E., 2011. Trajectory analysis and semantic region modeling using nonparametric bayesian models. Internat. J. Comput. Vision 95, 287–312.
Weijer, J., Schmid, C., 2006/ Coloring local feature extraction. In: Proc. European Conf. Computer Vision.
Weinland, D., Ronfard, R., Boyer, E., 2006. Free viewpoint action recognition using motion history volumes. J. Comput. Vision Image Understand. 104, 249–257.
Weinland, D., Boyer, E., Ronfard, R., 2007. Action recognition from arbitrary views using 3d exemplars. In: Proc. IEEE Internat. Conf. Computer Vision.
Weinland, D., Ozuysal, M., Fua, P., 2010. Making action recognition robust to occlusions and viewpoint changes. In: Proc. European Conf. Computer Vision.
Weiss, L.E., Sanderson, A.C., Neuman, C.P., 1987. Dynamic sensor-based control of robots with visual feedback. IEEE J. Robotics Automat. RA-3, 404–417.
Winkelman, F., Patras, I., 2004. Online globally consistent mosaicing using an efficient representation. IEEE Trans. Systems Man Cybernet 4, 3116–3121.
Winn, J., Criminisi, A., Minka, T., 2005. Object categorization by learned universal visual dictionary. In: Proc. IEEE Internat. Conf. Computer Vision.
Wold, H., 1985. Partial least squares. Encyclopedia Statist. Sci. 6, 581–591.
Wong, K.Y., Mendonca, R.S.P., Cipolla, R., 2003. Camera calibration from surfaces of revolution. IEEE Trans. Pattern Anal. Machine Intell. 25, 147–161.
Woo, D., Capson, D., 2000. 3d visual tracking using a network of low-cost pan/tilt cameras. In: Proc. Can. Conf. Electrical and Computer Engineering.
Wu, Z., Hristov, N.I., Hedrick, T.L., 2009. Tracking a large number of objects from multiple views. In: Proc. IEEE Internat. Conf. Computer Vision.
Xu, L., Neufeld, J., Larson, B., Schuurmans, D., 2004. Maximum margin clustering. In: Proc. Neural Information Processing Systems Conf.
Yan, P., Khan, S.M., Shah, M., 2008. Learning 4d action feature models for arbitrary view action recognition. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Ye, Y., Tsotsos, K., 1999. Sensor planning for 3d object search. J. Comput. Vision Image Understand. 73, 145–168.
Yilmaz, A., Shah, M., 2005. Recognizing human actions in video acquired by uncalibrated moving cameras. In: Proc. IEEE Internat. Conf. Computer Vision.
Yilmaz, A., Javed, O., Shah, M. (2006). Object tracking: A survey. ACM Computing Surveys 38.
Zelniker, E.E., Gong, S., Xiang, T., 2008. Global abnormal behaviour detection using a network of cctv cameras. In: Poc. IEEE Workshop on Visual Surveillance.
Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Machine Intell. 22, 1330–1340.
Zhang, Z., Huang, K., Tan, T., 2006. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In: Proc. IEEE Internat. Conf. Pattern Recognition.
Zhang, Z., Li, M., Huang, K., Tan, T., 2008. Practical camera auto-calibration based on object appearance and motion for traffic scene visual surveillance. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.
Zheng, W., Gong, S., Xiang, T., 2009. Associating groups of people. In: Proc. British Machine Vision Conf.
Zheng, W., Gong, S., Xiang, T., 2011. Person re-identification by probabilistic relative distance comparison. In: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition.