Int J Comput Vis (2010) 90: 106–129
DOI 10.1007/s11263-010-0347-5

# Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding

Chen Change Loy · Tao Xiang · Shaogang Gong

Received: 12 October 2009 / Accepted: 20 April 2010 / Published online: 4 May 2010

**Abstract** We propose a novel approach to understanding activities from their partial observations monitored through multiple non-overlapping cameras separated by unknown time gaps. In our approach, each camera view is first decomposed automatically into regions based on the correlation of object dynamics across different spatial locations in all camera views. A new Cross Canonical Correlation Analysis (xCCA) is then formulated to discover and quantify the time delayed correlations of regional activities observed within and across multiple camera views in a single common reference space. We show that learning the time delayed activity correlations offers important contextual information for (i) spatial and temporal topology inference of a camera network; (ii) robust person re-identification and (iii) global activity interpretation and video temporal segmentation. Crucially, in contrast to conventional methods, our approach does not rely on either intra-camera or inter-camera object tracking; it thus can be applied to low-quality surveillance videos featured with severe inter-object occlusions. The effectiveness and robustness of our approach are demonstrated through experiments on 330 hours of videos captured from 17 cameras installed at two busy underground stations with complex and diverse scenes.

**Keywords** Visual surveillance · Correlation modelling · Time delay estimation · Person re-identification · Camera topology inference · Multi-camera activity modelling

C.C. Loy (✉) · T. Xiang · S. Gong
School of EECS, Queen Mary University of London, London, UK
e-mail: ccloy@eecs.qmul.ac.uk

T. Xiang
e-mail: txiang@eecs.qmul.ac.uk

S. Gong
e-mail: sgg@eecs.qmul.ac.uk

## 1 Introduction

In recent years there has been increasing deployment of multiple camera systems in wide-area public spaces such as airports, underground stations, shopping complexes and road junctions. To facilitate more efficient multi-camera surveillance and reduce the burden on human operators, growing research efforts have been undertaken on automated activity understanding in camera networks, focusing on camera topology inference (Makris et al. 2004; Tieu et al. 2005; van den Hengel et al. 2006), person re-identification (Javed et al. 2003, 2005; Prosser et al. 2008; Gheissari et al. 2006; Gray and Tao 2008; Hu et al. 2006a; Zheng et al. 2009), and global activity analysis (Lee et al. 2000; Wang et al. 2010; Zelniker et al. 2008). In topology inference, the aim is to infer spatial and temporal relationships between cameras. The task of person re-identification is concerned with associating people observed at different camera views. As for global activity analysis, one wishes to understand activities captured by multiple cameras holistically by building global activity models. These three problems are non-trivial, especially given multiple disjoint cameras with non-overlapping views, in which global activities can only be observed partially with different views being separated by unknown time gaps. In particular, the unknown and often large separation of cameras in space and over time increases the uncertainties in activity understanding due to drastic feature variations and temporal discontinuity in visual observations.

Let us first define the term ‘activity’ before we discuss the motivations of our approach on multi-camera activity understanding. In this paper, we categorise activities into global activities and regional activities. A regional activity refers to an activity that takes place locally in a single region of a camera view. For instance, passengers walking next to a train track or sitting on benches on a platform are regional

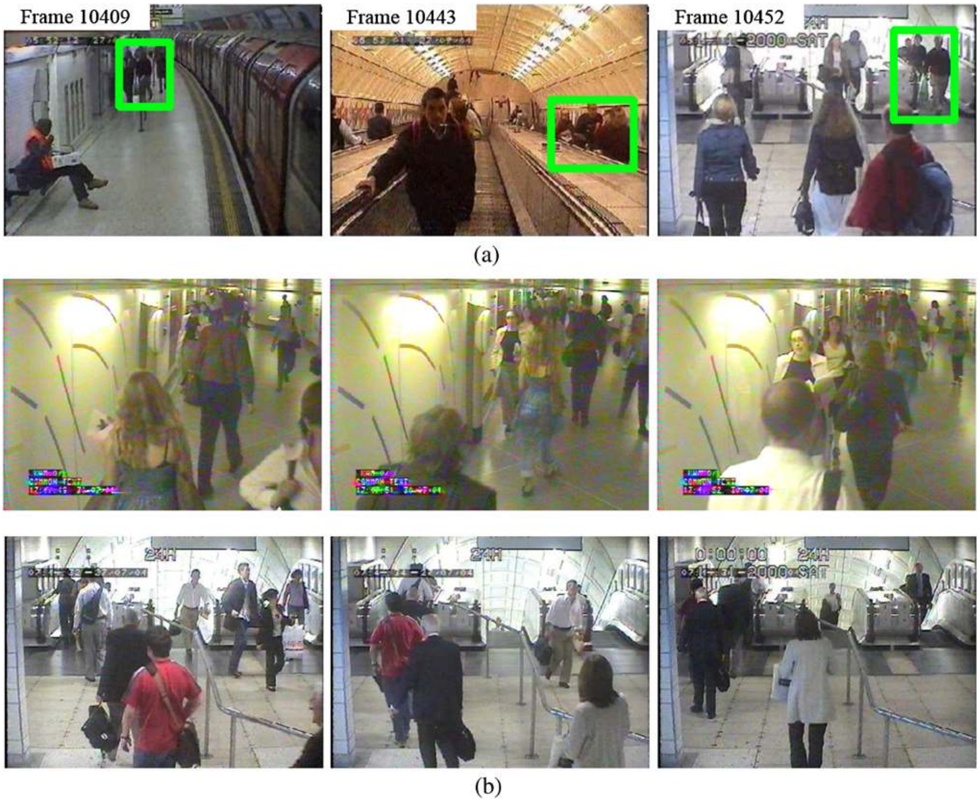


**Fig. 1** (**a**) Partial observations of activities observed from different camera views often form a chain of inter-correlated spatio-temporal patterns: a group of people (highlighted in green boxes) get off a train [Cam 8, frame 10409] and subsequently take an upward escalator [Cam 5, frame 10443] which leads them to the escalator exit view [Cam 4, frame 10452]. (**b**) Three consecutive frames captured from two different cameras at 0.7 frames per second (fps). An object can pass through the whole view in just three frames. In addition, severe inter-object occlusion and low-quality video are among the key factors that render object tracking infeasible

activities. A global activity, on the other hand, is defined as an activity that involves correlated partial observations of multiple regional activities across multiple cameras. For example, a global activity of train departure may involve co-existing activities taking place at different regions such as the movements of a train at a track area of a platform, passengers moving towards the exits of the platform, and passengers leaving station via escalators (see Fig. 1(a)).

The key to activity understanding in multiple non-overlapping cameras lies on how well we can link the partial observations of an activity together for complete and global interpretation. Specifically, activities of an object in a public space are inherently context-aware, exhibited through constraints imposed by scene layout and the correlated activities of other objects both in the same camera view and other views. Consequently the partial observations of a global activity are correlated in that they take place following a certain temporal order with unknown temporal gaps caused by the spatial distances between camera views. In other words, these partial observations often form *a chain of inter-correlated spatio-temporal patterns*, spanning across different regions in a networked global view space (see Fig. 1(a) for an example). It is therefore necessary to discover and quantify the correlations between these partial observations in terms of both temporal order and temporal delays, which provides important contextual information on the global activities across multiple camera views. While considerable work has been done on multi-camera activity understanding, none of them has addressed the problem of discovering and modelling multi-camera activity correlations with unknown time delays (see Sect. 2 for detailed discussion).

An obvious solution to multi-camera activity correlation analysis and global activity understanding seems to be tracking objects within and across camera views. Indeed, most previous methods rely on either intra-camera (within camera) tracking to detect entry and exit events for modelling transition time distribution, or inter-camera (between cameras) tracking for object/trajectory association (Zelniker et al. 2008; Makris et al. 2004; Wang et al. 2010). These methods generally assume reliable object localisation and detection as well as smooth object movement. However, these assumptions are often invalid in real-world surveillance settings featured with severe occlusions caused by *excessive number of objects* in the scene and *very low temporal and spatial resolution*.[1] Particularly, in a typical public scene as shown in Fig. 1(b), the sheer number of objects with complex activities causes severe inter-object occlusions continuously, leading to temporal discontinuity of trajectories. Tracking is further compounded by the typically low temporal resolution of surveillance video, where large spatial

[1] Many multi-camera surveillance systems record videos at less than 5 fps to optimise data bandwidth and storage space (Kruegle 2006; Cohen et al. 2006).

displacement is observed in moving objects between consecutive frames.

In this paper, we propose a novel approach to modelling time delayed correlations among multi-camera activities without relying on either intra-camera or inter-camera tracking. Specifically, since a complex scene naturally consists of multiple local scene regions that encompass distinctive activities, each camera view is first decomposed automatically into regions, across which different spatio-temporal activity patterns are observed. A novel Cross Canonical Correlation Analysis (xCCA) framework is then formulated to discover and quantify correlation and temporal relationships of *arbitrary order* among these multi-camera regional activities. As opposed to object centred approaches, our xCCA learns activity correlations by exploiting the underlying spatial and temporal correlation of regional activities in a holistic manner and avoids object tracking under challenging surveillance conditions.

The proposed approach is employed to address three fundamental problems in multi-camera activity understanding:

(i) Estimate the spatial topology (i.e. between-camera spatial relationships) and more importantly the temporal topology of a camera network, that is, the temporal relationships (e.g. the unknown delay time) between inter-correlated partial observations taking place in different camera views.
(ii) Facilitate more robust and accurate person re-identification among different camera views, by resolving ambiguities and uncertainties that arise due to large and unknown separation among cameras both spatially and temporally.
(iii) Interpret global activity, and perform video temporal segmentation by linking visual evidences collected from different camera views.

We demonstrate the effectiveness of the proposed approach using 330 hours of videos captured at 0.7 fps from two busy underground stations with eight and nine camera views respectively, all of which feature crowded scene and complex activities.

The rest of the paper is structured as follows: Sect. 2 reviews related work to highlight the contributions of this study. The proposed framework is explained in Sect. 3. Results are reported and discussed in Sect. 4. Finally, the paper concludes with some suggestions for further investigation in Sect. 5.

## 2 Previous Work

Much work on activity correlation modelling has been devoted to single camera scenario. In most cases, activity correlations are modelled among limited number of individual objects (Oliver et al. 2000; Du and Chen 2007; Gong and Xiang 2003). More recently, Li et al. (2008) propose to model the co-occurrence of activities observed in different regions of a wide-area scene. Although promising results are reported, these methods are not suitable for multiple camera scenarios since the time delays between activities are ignored.

For *multi-camera activity analysis*, there have been a few attempts, but only limited to modelling co-occurrence or first-order temporal relationships between activities. Zhou and Kimber (2006) model activities across views using a Coupled Hidden Markov Model (CHMM), which can only handle first-order dependencies without considering relationships of arbitrary order. Wang et al. (2010) employ intra-camera tracking to extract trajectories from each camera view and group them into global activities using topic models extended from the Latent Dirichlet Analysis (LDA) (Blei et al. 2003), with a restriction that only co-occurrence relationships between activities within a fixed temporal threshold can be modelled. In contrast to existing methods, the xCCA framework proposed in this work is capable of capturing correlation and temporal relationships of *arbitrary order* without relying on object tracking. Moreover, our approach is able to cope with co-existence of large number of objects both within and across camera views.

Apart from global activity analysis, considerable efforts have been devoted to *camera network topology inference* (Javed et al. 2003; Makris et al. 2004; Tieu et al. 2005) and *person re-identification* (Javed et al. 2003, 2005; Gheissari et al. 2006; Prosser et al. 2008). To infer the topology of a camera network or re-identify a person over multiple cameras, existing methods generally follow two approaches: (i) matching individual object visual appearance or motion trends such as movement speed; (ii) exploiting distribution of transition times of entry and exit events.

The first approach, e.g. (Javed et al. 2003), relies on the availability of reliable visual and motion features from target to achieve inter-camera object association. In practice, this approach suffers from significant feature variations across camera views due to changes in illumination (both intra- and inter-camera), camera orientation, and person appearance caused by pose change (see Fig. 1(a) for example). Although various strategies have been proposed (Javed et al. 2005; Gheissari et al. 2006; Gray and Tao 2008; Zheng et al. 2009) to adapt and rectify feature variation, object feature matching remains a notoriously difficult problem under real-world surveillance conditions. This is because that even if feature variation can be rectified or reduced, reliable features for matching may still not be available due to severe inter-object occlusions, typical in a busy public space. In contrast, our approach circumvents unreliable feature matching by inferring the spatial and temporal relationships between regions across camera views, which also provides an important contextual cue in addition to visual appearance for person association. Furthermore, most existing methods perform supervised training by assuming known object correspondences, which is difficult to establish automatically and reliably. In comparison, our approach is unsupervised without relying on any prior knowledge on object correspondence.

The second approach, e.g. (Makris et al. 2004; Tieu et al. 2005), avoids explicit feature matching by modelling transition time distribution between entry and exit events detected in different camera views. These methods are based on the assumption that individual exit and entry events can be detected by exploiting starting and ending points of object trajectories. However, as we shall see in Sect. 4.4, object tracking in a busy public scene is extremely unreliable especially when the spatial and temporal resolutions of the video are low. Our approach overcomes this problem by discovering and quantifying correlation and temporal relationships between activities in different camera views without relying on either intra-camera or inter-camera tracking. It therefore can be applied under the most challenging public scene viewing conditions.

It is worth pointing out that van den Hengel et al. (2006) also attempted to infer the camera topology without relying on object tracking. Their method starts with full connectivity among camera views and gradually eliminates linkages among image regions that exhibit simultaneous object occupancy and vacancy. This method, however, only examines static object co-occurrence over time without considering the temporal relationships among multi-camera activities. Importantly, it ignores possible connections between non-overlapping views. It is thus limited to learning only the connectivity among overlapping camera views. On the contrary, our approach takes temporal relationships among activities into account. In addition, although we focus on disjoint cameras in this study, the proposed approach can be readily used for camera views with any degrees of overlapping.

In summary, the main novelties of the proposed approach are two-fold:

(i) It is capable of discovering and quantifying the correlation and temporal relationships of arbitrary order among local activities across different camera views. To our best knowledge, this study is the first attempt to model time delayed activity correlations among multiple cameras.

(ii) It does not rely on either inter-camera or intra-camera tracking. Therefore it is robust to occlusions and can be applied to crowded scenes of low spatial and temporal resolutions.

Compared to our earlier version of this work (Loy et al. 2009), there are three changes in methodology for addressing a number of key limitations of our CVPR approach that prevent it from being scalable on more complicated and challenging multi-camera scenes. First, we formulate in this paper a robust background model to cope with sudden changes in global intensity level within camera views in surveillance videos (Sect. 3.1). This results in more reliable features for the proposed time delayed correlation analysis. Second, we formulate a new topology inference approach that considers both time delay and correlation strength (Sect. 3.3). Third, we employ a more principled approach to compute objects' spatio-temporal relationships for context-aware person re-identification (Sect. 3.4). On experimental evaluation, a new dataset is added to the one used in Loy et al. (2009), which was captured from a multi-camera site that exhibits more complex behaviours and contains more diverse scenes. Our experimental results suggested the performance of the approach formulated in this work is superior to that of the approach in Loy et al. (2009).

## 3 Multi-camera Activity Correlation Analysis

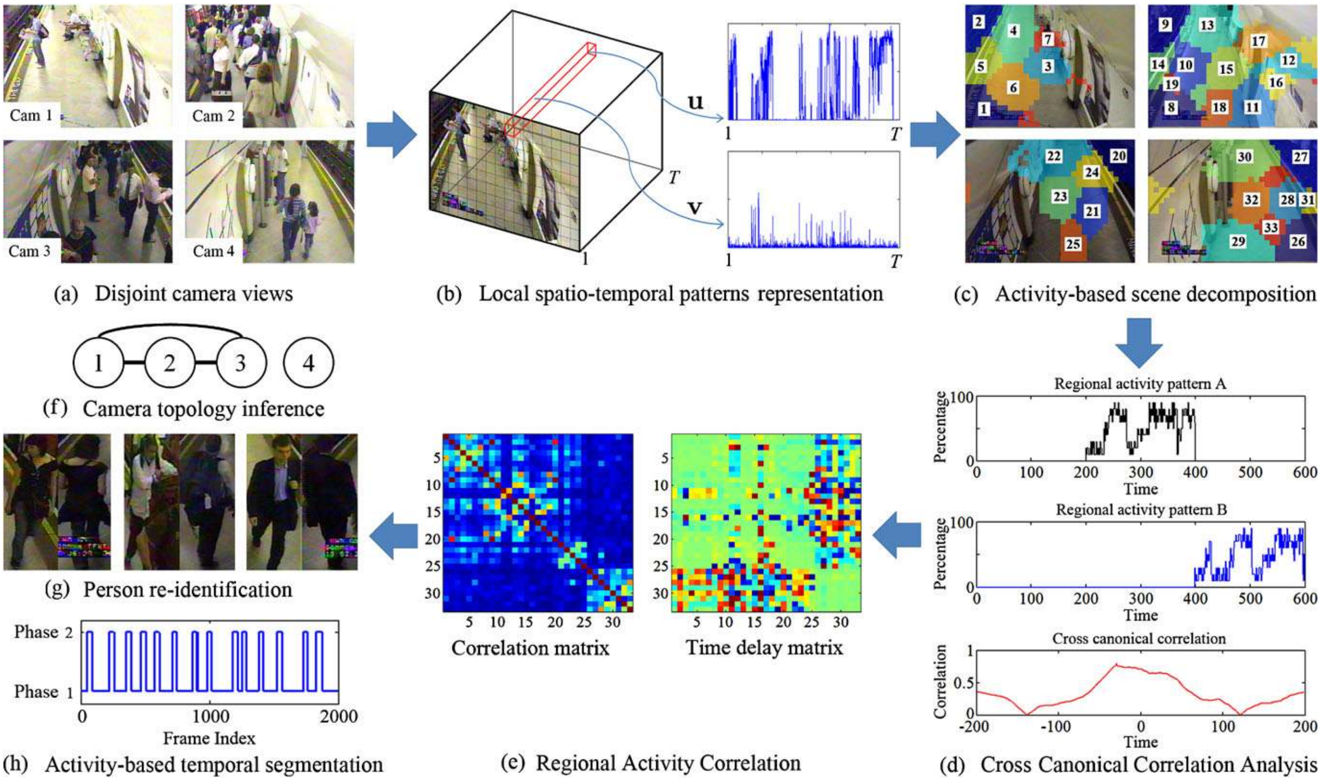
The key components of the proposed approach is illustrated in Fig. 2. Given disjoint camera views in a camera network (Fig. 2(a)), local spatio-temporal patterns are first extracted and represented as time-series data from each camera view (Fig. 2(b)). The patterns are then used as input to our activity-based scene decomposition method to segment the scenes into regions (Fig. 2(c)), from which the regional activity patterns are extracted. Subsequently, the Cross Canonical Correlation Analysis (xCCA) is performed to infer inter-region time-delayed correlations (Fig. 2(d)). Regional activity correlations are then discovered and quantified (Fig. 2(e)). We refer the process of activity-based scene decomposition and xCCA as training in this study. Finally the inferred regional activity correlations are exploited for camera topology inference, and used as contextual information for person re-identification and global activity temporal segmentation (Fig. 2(f–h)).

### 3.1 Scene Decomposition and Activity Representation

A complex public scene naturally consists of multiple local regions, each of which encapsulates a unique set of activity patterns correlated with each other either explicitly or implicitly. Given a set of training video sequences, our goal is to decompose $M$ camera views into $N$ regions $\boldsymbol{\mathcal{R}}$ according to the spatial-temporal distribution of activity patterns, where $\boldsymbol{\mathcal{R}}$ is given as

$$\boldsymbol{\mathcal{R}} = \{\mathcal{R}_n | n = 1, \ldots, N\}. \quad (1)$$

Consequently, the $m$-th camera view in the network contains $N_m$ regions with $N_m$ being determined automatically and $N = \sum_{m=1}^{M} N_m$.



(a) Disjoint camera views (b) Local spatio-temporal patterns representation (c) Activity-based scene decomposition (d) Cross Canonical Correlation Analysis (e) Regional Activity Correlation (f) Camera topology inference (g) Person re-identification (h) Activity-based temporal segmentation

**Fig. 2** A diagram illustrating our multi-camera activity correlation approach

### *3.1.1 Robust Background Modelling*

Sudden and frequent intensity changes in real-world surveillance videos may introduce noise that affects the accuracy of time-delayed correlation analysis. To address this problem, a robust background modelling method is formulated.[2] The changes in intensity level are caused either by lighting condition changes (e.g. moving clouds outdoor or flashing advertising boards indoor) or camera response to different crowdedness in the scene. In the latter case, auto gain and white balancing functions of cameras yield different global intensity level on a particular video frame when crowd or large objects are present in the video. An example of the latter can be seen in Fig. 3 where drastically different global intensity level are observed when an underground train platform changes from crowding (Fig. 3(a)) to empty (Fig. 3(b)).

The key idea of our method is to adapt the background image to the intensity level of current frame prior to background subtraction. In particular, a static background image is first constructed by employing a method proposed by Russell and Gong (2006). We then compute pixel-level intensity ratios **g** between all pixels of the stored background image and the current frame. Subsequently, the background image is adjusted by multiplying all its pixels with **g**. However, not all the ratios reflect the true intensity level change as some of them belong to foreground regions. To eliminate the effect

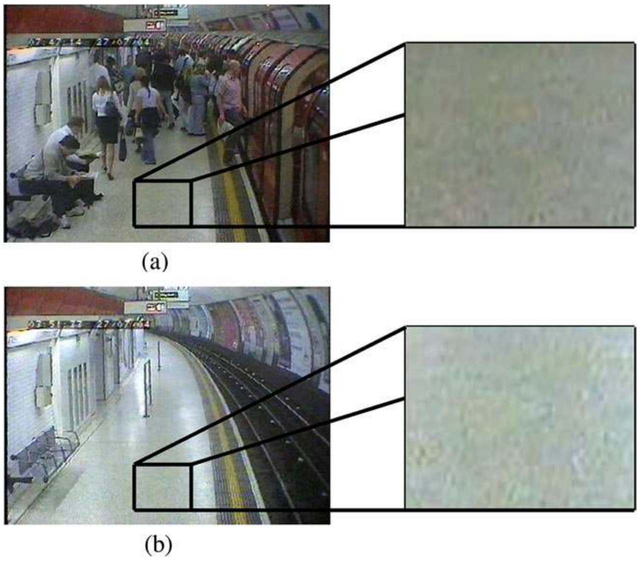


**Fig. 3** The figure depicts a frame with abrupt intensity level change (**a**) compared to its background model (**b**). The ratios of RGB channels between the extracted regions of (**a**) and (**b**) are 1.2380, 1.2829 and 1.3428 respectively

[2] Matlab implementation of the proposed background subtraction method is available at http://www.eecs.qmul.ac.uk/~ccloy/files/substractBackgroundMS.zip.

of incorrect ratios caused by foreground regions, a mean-shift procedure (Fukunaga and Hostetler 1975) is performed to find the stationary point of the distribution of $\mathbf{g}$. In particular, the centre of a Gaussian kernel denoted by $\{\mathbf{c}_j\}_{j=1,2,\ldots}$ is iteratively moved from the current point to the new point according to:

$$\mathbf{c}_{j+1} = \frac{\sum_{i=1}^{N_{\text{pixel}}} \mathbf{g}_i \exp(-\frac{\|\mathbf{c}_j - \mathbf{g}_i\|^2}{2h^2})}{\sum_{i=1}^{N_{\text{pixel}}} \exp(-\frac{\|\mathbf{c}_j - \mathbf{g}_i\|^2}{2h^2})}, \quad j = 1, 2, \ldots, \tag{2}$$

where $N_{\text{pixel}}$ is the number of pixels and the size of the Gaussian kernel $h$ is set to 1 in this study. To obtain the initial point $\mathbf{c}_1$ of the kernel, we first perform a coarse background subtraction between the stored background image and the input image. The initial point is then computed as the mean of the intensity ratios between all the non-foreground pixels. The mean shift procedure terminates when the maximum iteration allowed is reached or

$$\|\mathbf{c}_{j+1} - \mathbf{c}_j\| < \epsilon, \tag{3}$$

where $\epsilon$ is set to a small value. The final centre of the kernel gives the most likely intensity ratios that account for the change of intensity level. These ratios are used to adjust the original background image in an online manner and a fine background subtraction is performed to obtain a foreground mask that is least affected by abrupt changes of intensity level.

Apart from implementing robust background modelling, we also perform colour correction in YUV colour space by blending the chrominancy components of previous and current frames to reduce the chroma noise commonly found in surveillance videos.

#### *3.1.2 Discussion*

Many previous work has been done in robust background modelling. There are quite a few methods that can handle gradual lighting changes but are still vulnerable to sudden lighting changes (Stauffer and Grimson 2000; Zivkovic and van der Heijden 2006; Friedman and Russell 1997). In particular, they are based on statistical background modelling which are slow in model update, thus being less effective in handling rapid lighting changes. Recently, a number of methods have been proposed to cope with sudden lighting changes (Pilet et al. 2008; Sung et al. 2008; Xie et al. 2004). Our background subtraction method is similar to Sung et al. (2008) but with a key difference on the intensity ratio estimation. In their approach, a set of recent frames are kept to estimate a background model; therefore $\mathbf{g}$ is estimated between the current frame and previous frame. However, given surveillance videos featured with crowded scenes, it is hard to maintain a reliable background model with limited number of recent frames. Therefore, we choose to generate a single background image, and adjust it based on $\mathbf{g}$ estimated between the current frame and the background image itself.

#### *3.1.3 Local Block Activity Pattern Representation*

First, we divide the image space of a camera view into equal-sized blocks with 10 × 10 pixels each (Fig. 2(b)). Foreground pixels are then detected using the aforementioned background subtraction method. The foreground pixels are categorised as either static or moving via frame differencing (e.g. sitting people are detected as static foreground whilst passing-by people are detected as moving foreground). Activity patterns of a block are then represented as a bivariate time-series

$$\begin{aligned} \mathbf{u_b} &= (u_{\mathbf{b},1}, \ldots, u_{\mathbf{b},t}, \ldots, u_{\mathbf{b},T}), \\ \mathbf{v_b} &= (v_{\mathbf{b},1}, \ldots, v_{\mathbf{b},t}, \ldots, v_{\mathbf{b},T}), \end{aligned} \tag{4}$$

where $\mathbf{b}$ representing the two-dimensional coordinates of a block in the image space and $T$ is the total number of frames used in training, $u_{\mathbf{b},t}$ and $v_{\mathbf{b},t}$ are the percentage of static and moving foreground pixels within the block at frame $t$ respectively. Note that $T$ needs to be sufficiently large to cover enough repetitions of activity patterns, depending on the complexity of a scene.

The low spatial and temporal resolution of surveillance footages has imposed great challenges to the selection of appropriate features for local block activity pattern representation. As explained in Sect. 1, trajectory features (Hu et al. 2006b; Saleemi et al. 2009) are extremely unreliable under these restrictions. More sophisticated features such as optical flow (Wang et al. 2009; Yang et al. 2009) are found to be unstable too. Importantly, optical flow computation assumes small object displacement and constant brightness for the computation of velocity field; both assumptions are invalid for videos with very low frame rate and poor image quality. Similarly, spatio-temporal gradients proposed by Kratz and Nishino (2009) would fail due to motion discontinuities in low-frame rate videos.

Consequently $\mathbf{u_b}$ and $\mathbf{v_b}$ are chosen as they are the only features that can be extracted reliably given videos of low spatial and temporal resolution such as those used in our experiments (Sect. 4). Despite their simplicity as time-series features $\mathbf{u_b}$ and $\mathbf{v_b}$ are found to be effective in capturing the temporal characteristics of activity patterns including temporal persistence of different patterns and their temporal order.

#### *3.1.4 Activity-based Scene Decomposition*

After feature extraction, we group blocks into regions according to the similarity of local spatio-temporal activity patterns represented as $\mathbf{u}_\mathbf{b}$ and $\mathbf{v}_\mathbf{b}$. Specifically, two blocks are considered similar and grouped together if they are closed to each other spatially and exhibit high correlations in both static and moving foreground activities over time. The grouping process begins with computing correlation distances among local activity patterns of each pair of blocks. A correlation distance is defined as a dissimilarity metric derived from Pearson's correlation coefficient (Liao 2005), given as

$$\bar{r} = 1 - |r|. \tag{5}$$

In particular, $\bar{r} = 0$ if two blocks have strongly correlated local activity patterns, or $\bar{r} = 1$ otherwise. Subsequently, we construct an affinity matrix $\mathbf{A} = \{A_{ij}\} \in \mathbb{R}^{B \times B}$, where $B$ is the total number of blocks in the camera view and $A_{ij}$ is defined as:

$$A_{ij} = \begin{cases} \exp\bigl(-\frac{(\bar{r}^{\mathbf{u}}_{ij})^2}{2\sigma^{\mathbf{u}}_i \sigma^{\mathbf{u}}_j}\bigr) \exp\bigl(-\frac{(\bar{r}^{\mathbf{v}}_{ij})^2}{2\sigma^{\mathbf{v}}_i \sigma^{\mathbf{v}}_j}\bigr) \exp\bigl(-\frac{\|\mathbf{b}_i - \mathbf{b}_j\|^2}{2\sigma^2_{\mathbf{b}}}\bigr), \\ \quad \text{if } \|\mathbf{b}_i - \mathbf{b}_j\| \le R \text{ and } i \neq j, \\ 0 \quad \text{otherwise,} \end{cases} \tag{6}$$

where the correlation distances of $\mathbf{u}_\mathbf{b}$ and $\mathbf{v}_\mathbf{b}$ between block $i$ and block $j$ are given by $\bar{r}^{\mathbf{u}}_{ij}$ and $\bar{r}^{\mathbf{v}}_{ij}$ respectively, whilst $[\sigma^{\mathbf{u}}_i, \sigma^{\mathbf{u}}_j]$ and $[\sigma^{\mathbf{v}}_i, \sigma^{\mathbf{v}}_j]$ are the respective correlation scaling factors for $\bar{r}^{\mathbf{u}}_{ij}$ and $\bar{r}^{\mathbf{v}}_{ij}$. The correlation scaling factors are defined as the mean correlation distance between the current block and all blocks within a radius $R$. The coordinates of the two blocks are denoted as $\mathbf{b}_i$ and $\mathbf{b}_j$. Similar to the correlation scaling factors, the spatial scaling factor $\sigma_\mathbf{b}$ is defined as the mean spatial distance between the current block and all blocks within the radius $R$. The affinity matrix is then normalised according to

$$\overline{\mathbf{A}} = \mathbf{L}^{-\frac{1}{2}} \mathbf{A} \mathbf{L}^{-\frac{1}{2}}, \tag{7}$$

where $\mathbf{L}$ is a diagonal matrix and $L_{ii} = \sum_{j=1}^{B} A_{ij}$. Upon obtaining the normalised affinity matrix $\overline{\mathbf{A}}$, we employed spectral clustering method proposed by Zelnik-Manor and Perona (2004) to decompose each camera view into regions with the optimal number of regions being determined automatically.

In the computation of the affinity matrix, we follow Li et al. (2008) to compute similarity within a fixed radius $R$. This strategy was shown to prevent under-fitting problem during decomposition in comparison to the local scaling strategy proposed by Zelnik-Manor and Perona (2004). Note that similarity in the Gaussian kernel affinity matrix is governed by the selection of scaling factors (Zelnik-Manor and Perona 2004; Ng et al. 2001). In this study, the scaling factors are functions of the radius $R$; the scene decomposition results are therefore governed by the selection of $R$. From our experiments, we observed that the cluster formations are generally stable when $R$ is set within the range of 20–30. Consequently, we selected $R = 20$ in this study. Figure 2(c) shows some examples of scene decomposition. It is evident that each camera view is decomposed into semantically meaningful regions such as train track areas and people sitting areas.

Our scene decomposition method is similar to that of Li et al. (2008) but with a noticeable modification on how local activities are represented. Specifically, in our method local activities are represented as time series and correlation distance between them are used as the dissimilarity measure. In comparison, a Bag of Words representation is adopted by Li et al. (2008), which ignores the temporal order information of a local activity and is thus less discriminative than our representation.

#### 3.1.5 *Regional Activity Representation*

Given the scene decomposition, regional activity patterns of a camera view are formed based on the local block activity patterns. In particular, the regional activity patterns at region $\mathcal{R}_n$ is represented as

$$\hat{\mathbf{u}}_n = \frac{1}{|\mathcal{R}_n|} \sum_{\mathbf{b} \in \mathcal{R}_n} \mathbf{u}_\mathbf{b}, \qquad \hat{\mathbf{v}}_n = \frac{1}{|\mathcal{R}_n|} \sum_{\mathbf{b} \in \mathcal{R}_n} \mathbf{v}_\mathbf{b}, \tag{8}$$

where $|\mathcal{R}_n|$ is the number of blocks belong to region $\mathcal{R}_n$. To facilitate a more accurate time delayed correlation analysis, we remove any region with half of its blocks exhibiting low activity. In this study, a low-activity block is defined as a block with activity patterns having a standard deviation that is less than three.

### 3.2 Cross Canonical Correlation Analysis

For any pair of regions in a camera network, two questions are to be answered: (i) are activities in these regions correlated? (ii) if yes, how strong are the correlations and what are the temporal relationships among them? It is non-trivial to discover and quantify correlations and temporal relationships between cameras. Different viewing angles of cameras may introduce pattern variations across camera views. Importantly, correlations between regional activities across disjoint camera views are complex in that there is often an unknown temporal gap/delay between the times when a causing activity in one region taking place and the correlated/caused activity in the other region being observed.

To this end, we wish to search for linear combinations[3] of the regional activities (represented as time-series) having maximal correlation and model the temporal gap as a temporal dependency of arbitrary order between the two time-series. Consequently, we formulate a new Cross Canonical Correlation Analysis (xCCA) to measure the correlation of two regional activities as a function of an unknown time lag $\tau$ applied to one of the two regional activity time-series.

Our approach differs from Canonical Correlation Analysis (CCA) (Hotelling 1936) in that CCA can only measure how strong two vector variables are correlated in a concurrent or zero-order sense. The proposed xCCA extends CCA to measure correlations beyond zero order by including additional steps similar in nature to the standard Cross Correlation Analysis (xCA) (Kendall and Ord 1990). This principally involves shifting of one time series and computes its canonical correlation with the other. An example is shown in Fig. 2(d).

Formally, let $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ denote the two regional activity time series observed in the $i$th and $j$th regions respectively. Note that $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ are time-series of $N_f$-dimensional variables. In our case, $N_f = 2$ since we extract two features $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ from each region. For clarity in the following equations, we denote $\mathbf{y}(t) = \mathbf{x}_j(t+\tau)$. We also omit the symbol $t$ for conciseness, e.g. time series $\mathbf{x}_i(t)$ becomes $\mathbf{x}_i$ and $\mathbf{y}(t)$ becomes $\mathbf{y}$.

At each time delay index $\tau$ (or each shifting step), xCCA finds two sets of optimal basis vectors $\mathbf{w}_{x_i}$ and $\mathbf{w}_y$ for $\mathbf{x}_i$ and $\mathbf{y}$ such that correlation of the projections of them onto the basis vectors are mutually maximised. Let linear combinations of canonical variates be $x_i = \mathbf{w}_{x_i}^{\mathsf{T}}\mathbf{x}_i$ and $y = \mathbf{w}_y^{\mathsf{T}}\mathbf{y}$, canonical correlation $\boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau)$ is defined as:

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau) &= \frac{\mathrm{E}[x_i y]}{\sqrt{\mathrm{E}[x_i^2]\mathrm{E}[y^2]}} \\ &= \frac{\mathrm{E}[\mathbf{w}_{x_i}^{\mathsf{T}}\mathbf{x}_i\mathbf{y}^{\mathsf{T}}\mathbf{w}_y]}{\sqrt{\mathrm{E}[\mathbf{w}_{x_i}^{\mathsf{T}}\mathbf{x}_i\mathbf{x}_i{}^{\mathsf{T}}\mathbf{w}_{x_i}]}\sqrt{\mathrm{E}[\mathbf{w}_y^{\mathsf{T}}\mathbf{y}\mathbf{y}^{\mathsf{T}}\mathbf{w}_y]}} \\ &= \frac{\mathbf{w}_{x_i}^{\mathsf{T}}\mathbf{C}_{x_i y}\mathbf{w}_y}{\sqrt{\mathbf{w}_{x_i}^{\mathsf{T}}\mathbf{C}_{x_i x_i}\mathbf{w}_{x_i}}\sqrt{\mathbf{w}_y^{\mathsf{T}}\mathbf{C}_{yy}\mathbf{w}_y}},\end{aligned} \tag{9}$$

where $\mathbf{C}_{x_i x_i}$ and $\mathbf{C}_{yy}$ are within-set covariance matrices of $\mathbf{x}_i$ and $\mathbf{y}$, respectively, whilst $\mathbf{C}_{x_i y}$ is between-set covariance matrix.

The maximisation of $\boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau)$ at each time delay index $\tau$ can be solved by setting the derivatives in (9) to zero, yielding the following eigenvalue equations:

$$\begin{cases}\mathbf{C}_{x_i x_i}^{-1}\mathbf{C}_{x_i y}\mathbf{C}_{yy}^{-1}\mathbf{C}_{y x_i}\mathbf{w}_{x_i} = \boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}^2(\tau)\mathbf{w}_{x_i}, \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{y x_i}\mathbf{C}_{x_i x_i}^{-1}\mathbf{C}_{x_i y}\mathbf{w}_y = \boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}^2(\tau)\mathbf{w}_y,\end{cases} \tag{10}$$

where the eigenvalues $\boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}^2(\tau)$ are the square canonical correlations and the eigen vectors $\mathbf{w}_{x_i}$ and $\mathbf{w}_y$ are the basis vectors. We only need to solve one of the eigenvalue equations since the equations are related by:

$$\begin{cases}\mathbf{C}_{x_i y}\mathbf{w}_y = \boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau)\lambda_{x_i}\mathbf{C}_{x_i x_i}\mathbf{w}_{x_i}, \\ \mathbf{C}_{y x_i}\mathbf{w}_{x_i} = \boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau)\lambda_y\mathbf{C}_{yy}\mathbf{w}_y,\end{cases} \tag{11}$$

where

$$\lambda_{x_i} = \lambda_y^{-1} = \sqrt{\frac{\mathbf{w}_y^{\mathsf{T}}\mathbf{C}_{yy}\mathbf{w}_y}{\mathbf{w}_{x_i}^{\mathsf{T}}\mathbf{C}_{x_i x_i}\mathbf{w}_{x_i}}}. \tag{12}$$

The time delay that maximises the canonical correlation between $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ is computed as:

$$\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j} = \arg\max_{\tau} \frac{\sum^{\Gamma}\boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\tau)}{\Gamma}, \tag{13}$$

where $\Gamma = \min(\operatorname{rank}(\mathbf{x}_i), \operatorname{rank}(\mathbf{x}_j))$. Note that we average the canonical correlation function with $\Gamma$ to obtain a single correlation value at each time delay index. The associated maximum canonical correlation is then obtained by locating the peak value in the averaged canonical correlation function as:

$$\hat{\rho}_{\mathbf{x}_i,\mathbf{x}_j} = \frac{\sum^{\Gamma}\boldsymbol{\rho}_{\mathbf{x}_i,\mathbf{x}_j}(\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j})}{\Gamma}. \tag{14}$$

We compute the maximum canonical correlation and the associated time delay for each pair of regional activity patterns to construct a regional activity affinity matrix

$$\mathbf{P} = \{P_{ij}\} \in \mathbb{R}^{N\times N}, \quad P_{ij} = \hat{\rho}_{\mathbf{x}_i\mathbf{x}_j}, \tag{15}$$

and a time delay matrix

$$\mathbf{D} = \left\{D_{ij}\right\} \in \mathbb{R}^{N\times N}, \quad D_{ij} = \hat{\tau}_{\mathbf{x}_i\mathbf{x}_j}. \tag{16}$$

Note that $0 \le \hat{\rho}_{\mathbf{x}_i,\mathbf{x}_j} \le 1$ with equality to 1 if, and only if the two regional time series are identical. If $\tau = 0$, xCCA is equivalent to performing CCA on $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$.

#### *3.2.1 Discussion*

Our xCCA compares favourably to alternative correlation analysis methods. One alternative approach is to represent each region as a node in a Bayesian network and

[3] Recall that activity patterns over all regions are represented by the percentage of static and moving foreground pixels (8), which reflect the crowd densities in the regions. If two regions are connected with correlated activity patterns, we expect to observe similar changes in crowd density across them after a certain time delay. It is therefore reasonable to assume that the relationships of the features extracted from the two regions to be linear.

learn the optimal structure of the network. This can be achieved by performing search over the space of candidate network structures, using methods such as Markov Chain Monte Carlo Bayesian network structure learning (MCMC-BNSL) (Neapolitan 2003). The strength of dependency between two regions can then be represented by the frequency of an edge being selected from the sampled structures. However, the learned structure can only reveal zero-order temporal dependency, and thus cannot cope with more complex (and higher order) correlations that are common in multi-camera scenes. Another alternative is the standard Cross Correlation Analysis (xCA). Compared to xCA, xCCA is more capable of capturing the underlying mutual patterns of two regional activity time series. This is because by projecting them into an optimal subspace, it minimises the effect of pattern variations introduced by different camera viewing angles and the temporal delays between correlated activities across camera views.

### 3.3 Topology Inference

With the regional activity correlation of arbitrary order being discovered and quantified ((15) and (16)), we wish to infer a camera topology. It is observed that considerable high correlations can be found between some region pairs even though they are not close to each other both spatially and temporally. This could be caused by noise or constant crowdedness in both regions. As a result, when shifting is performed to compute the correlation function of two region time-series, the algorithm may locate a 'spurious' peak that does not reflect the true correlation among their activity patterns. Fortunately, those 'spurious' peaks are normally found at a point where the time delay has a large value. Therefore, we exploit both time delay and correlation strength for topology inference. Specifically, two cameras will be connected in the inferred topology if they contain connected regions which are defined as those with high correlation value (14) and short time delay (13).

First, we compute a region connectivity matrix $\mathbf{\Psi} = \{\Psi_{ij}\} \in \mathbb{R}^{N\times N}$, which represents how likely each pair of regions in the camera network are connected, or the strength of their connectivity. More specifically, each element in the region connectivity matrix is computed as

$$\Psi_{ij} = \overline{\hat{\rho}_{\mathbf{x}_i,\mathbf{x}_j}}(1 - \overline{|\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j}|}), \tag{17}$$

where $\overline{\hat{\rho}_{\mathbf{x}_i,\mathbf{x}_j}}$ is obtained from normalised regional activity affinity matrix $\mathbf{P}$, so that it has a value range of [0, 1]. Whilst $\overline{|\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j}|}$ is obtained by normalising the absolute values of the elements of the time delay matrix $\mathbf{D}$. These two normalisations ensure that we have $0 \leq \Psi_{ij} \leq 1$. The higher the value of $\Psi_{ij}$, the stronger the connectivity between a region pair.

Once we have obtained the region connectivity matrix, the camera topology, represented as a camera connectivity matrix $\mathbf{\Phi} = \{\Phi_{ij}\} \in \mathbb{R}^{M\times M}$, can be inferred. Specifically the strength of the connectivity between the $i$th and $j$th camera nodes is obtained by averaging the regional activity connectivity strength (17) between each pair of regions across the two camera views. In this study, in order to reduce the influence of possible noise and redundant connectivities in $\mathbf{\Psi}$, we first search for the strongest connectivity between a region in the $i$th camera view with all regions in the $j$th camera view. This searching step is repeated for all regions in the $i$th camera view. Subsequently, the top $N_e$ connectivities are averaged to obtain $\Phi_{ij}$, with $N_e$ being set to half of the number of regions in the $i$th camera view. Finally $\mathbf{\Phi}$ is normalised so that its elements have a value range of [0, 1]. Two cameras are then deemed as being connected if the corresponding $\Phi_{ij}$ value is greater than the mean value of all the elements of $\mathbf{\Phi}$.

### 3.4 Context-aware Person Re-identification

The goal of person re-identification is to search for a given individual who disappeared in one camera view over other camera views. Here we describe how the learned time-delayed activity correlations can be employed as contextual information to reduce the search space as well as to resolve ambiguities arising from:

(i) Similar visual features presented by different people.
(ii) Feature variations caused by different poses, camera viewing angles and illumination changes.

Simple colour histogram feature is used for discriminating an individual against others. Though more sophisticated features are available (Gheissari et al. 2006; Gray and Tao 2008), the use of simple features provides a baseline for evaluating to what extent the time delayed correlations could improve the person re-identification accuracy. Specifically, given the bounding boxes of two people $a$ and $b$ observed in different camera views, we first normalise the bounding boxes to equal size. We then segment each normalised boxes into $N_{\text{h}}$ horizontal strips of equal height, from which colour histograms are computed and concatenated for representing the visual appearances of $a$ and $b$.

The similarity between the two concatenated colour histograms $H^a$ and $H^b$ of $a$ and $b$ is measured using Bhattacharyya score (Bhattacharyya 1943; Comaniciu et al. 2000) as follows:

$$S^{a,b}_{\text{bha}} = \sum_{i=1}^{N_{\text{bin}}} \sqrt{H^a_i H^b_i}, \tag{18}$$

where $N_{\text{bin}}$ represents the number of bins. Each histogram bin is normalised using the total number of pixels in the normalised image, so that $\sum_{i=1}^{N_{\text{bin}}} H^a_i = 1$ and $\sum_{i=1}^{N_{\text{bin}}} H^b_i = 1$. Note that the Bhattacharyya score is close to zero (minimum value is 0) if $H^a$ and $H^b$ are very different, or have a maximum value of 1 if two histograms are identical. The Bhattacharyya score is first computed for each colour channel separately. The overall Bhattacharyya score $\overline{S}^{a,b}_{\text{bha}}$ is then obtained by multiplying the scores $S^{a,b}_{\text{bha}}$ computed over all channels.

To incorporate the learned activity correlations and time delays into the final score for person re-identification, we first determine the regions (see Sect. 3.1) occupied by person $a$ and $b$ and the associated inter-region correlation and time delay. In particular, if a person's bounding box overlaps $N_{\text{r}}$ regions in the image space, the occupancy fractions of individual regions within the bounding box are computed and represented as a set of weights:

$$\boldsymbol{\mu} = \{\mu_i | i = 1, \dots, N_{\text{r}}\}, \tag{19}$$

where $\sum_{i=1}^{N_r} \mu_i = 1$. The weights are used to calculate the correlation between regions occupied by person $a$ and $b$ as follows:

$$\hat{\rho}^{a,b} = \sum_{i=1}^{N_{\text{r}}^a} \mu_i^a \left( \sum_{j=1}^{N_{\text{r}}^b} \mu_j^b \, \hat{\rho}_{\mathbf{x}_i,\mathbf{x}_j} \right), \tag{20}$$

where $\hat{\rho}_{\mathbf{x}_i,\mathbf{x}_j}$ is the maximum cross canonical correlation computed using (14). The corresponding time delay is given as:

$$\hat{\tau}^{a,b} = \sum_{i=1}^{N_{\text{r}}^a} \mu_i^a \left( \sum_{j=1}^{N_{\text{r}}^b} \mu_j^b \, \hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j} \right), \tag{21}$$

where $\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j}$ is computed using (13). The overall score is computed as follows:

$$S^{a,b}_{\text{overall}} = \begin{cases} \overline{S}^{a,b}_{\text{bha}} \hat{\rho}^{a,b} & \text{if } 0 < t^{a,b}_{\text{gap}} < \alpha \hat{\tau}^{a,b}, \\ 0 & \text{otherwise}, \end{cases} \tag{22}$$

where $t^{a,b}_{\text{gap}}$ is the time gap of observing the two people in the two camera views, whilst $\alpha$ is a factor that determines the maximum allowable transition time between cameras during person matching.

### 3.5 Global Activity Interpretation

Global activities defined by correlated activities across multiple camera views should be modelled collectively. This is because by utilising visual evidences collected from different views, global activity modelling is more robust to noise and visual ambiguities than modelling activities separately within individual camera views. Note that the regional activity affinity matrix $\mathbf{P}$ (15) is only concerned with the correlations of regional activities. It does not reveal either the contributions of these regional activities to the global activities or the temporal dynamics of the global activities.

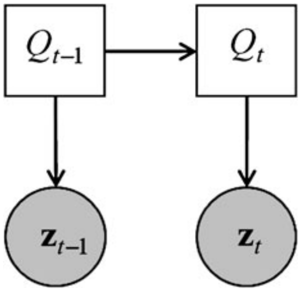


**Fig. 4** Hidden Markov Model with two time slices unrolled. Observation nodes are shown as *shaded circles* and hidden nodes are shown as *clear squares*

A complex camera network can capture many activities occurring simultaneously. However, not all the activities are correlated and they should be excluded during global activity modelling. In this study, the underlying global activities are discovered and modelled by taking the following steps:

(i) The same spectral clustering algorithm used in Sect. 3.1 is employed to group regional activities using the regional activity affinity matrix $\mathbf{P}$ (15) as an input affinity matrix.

(ii) The clusters returned by the spectral clustering algorithm are examined for discovering highly-correlated global activities. Specifically, those clusters that consist of cross-camera regions with the highest mean cross canonical correlations are selected.

(iii) Activity patterns in one of the $\chi$ selected regions are set as a reference point to temporally align activity patterns of other regions in accordance to the respective temporal offsets $\hat{\tau}_{\mathbf{x}_i,\mathbf{x}_j}$ computed using (13).

(iv) The aligned regional activity patterns, each represented as a two-dimensional time series (i.e. $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$), are concatenated together to form global activity patterns, $\mathbf{z}_t = \hat{\mathbf{u}}_{1,t} \| \hat{\mathbf{v}}_{1,t} \| \cdots \| \hat{\mathbf{u}}'_{\chi,t} \| \hat{\mathbf{v}}'_{\chi,t}$, with the prime symbol indicating an aligned time-series according to the temporal offset. The global activity patterns, $\mathbf{z}_t$ is then used as inputs to train an Hidden Markov Model (HMM) to model the temporal dynamics of the global activity.

(v) The HMM structure is shown in Fig. 4. It is an ergodic (fully-connected) model with $Q_t$ being discrete random variable, $Q_t \in \{q^i | i = 1, \dots, K\}$. We assume that the model is first-order Markov, i.e. $p(Q_t | Q_{1:t-1}) = p(Q_t | Q_{t-1})$. We also assume that the observations are conditionally first-order Markov, i.e. $p(\mathbf{z}_t | Q_t, \mathbf{z}_{1:t-1}) = p(\mathbf{z}_t | Q_t)$.

(vi) Automatic model selection is performed based on the Bayesian Information Criterion (BIC) score to find the number of hidden states $K$ in the model. For model parameter estimation, we first group the aligned regional activity patterns at different time instances using $K$-means clustering algorithm into $K$ groups. The clustering results, i.e. the means and covariances of individual groups are used to initialise a $K$-hidden states HMM. The model parameters are then estimated using the Baum-Welch algorithm (Baum et al. 1970).

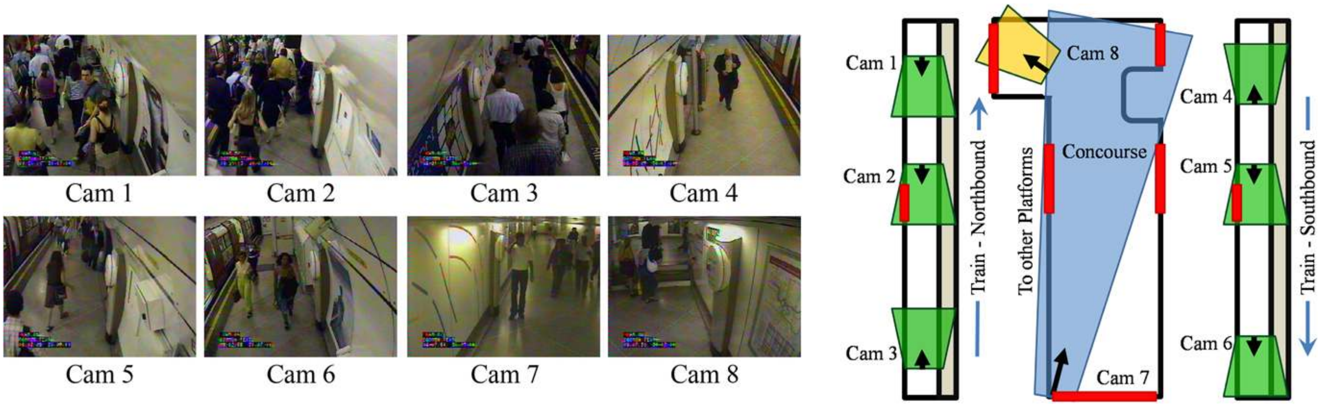


**Fig. 5** (Color online) The station layout and camera topology of Station A dataset. Entry and exit points are highlighted in *red bars*

The learned HMM for each global activity can be used for real-time activity-based temporal segmentation. The objective is to segment unseen video streams into activity phases based on 'what is happening' not only in a particular view but also in other views with highly-correlated activities. These activity phases are obtained by inferring the hidden states $Q_t$ at each time instance using online filtering method (Murphy 2002). In particular, given $\mathbf{z}_t$ observed as a continuous data stream, the probability of a particular hidden state $p(Q_t|\mathbf{z}_{1:t})$ is computed as a function of current input $\mathbf{z}_t$ and prior belief state $p(Q_{t-1}|\mathbf{z}_{1:t-1})$:

$$p(Q_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|Q_t, \mathbf{z}_{1:t-1})p(Q_t|\mathbf{z}_{1:t-1})$$
$$= p(\mathbf{z}_t|Q_t)\left[\sum_{Q_{t-1}} p(Q_t|Q_{t-1})p(Q_{t-1}|\mathbf{z}_{1:t-1})\right]. \quad (23)$$

Based on the Markovian assumption, we can use $p(\mathbf{z}_t|Q_t)$ to replace $p(\mathbf{z}_t|Q_t, \mathbf{z}_{1:t-1})$. Similarly, $p(Q_t|\mathbf{z}_{1:t-1})$ can be computed from the prior belief state under the Markovian assumption. To infer the activity phase $Q_t^*$, the probabilities $p(Q_t = q^i|\mathbf{z}_{1:t})$ are first computed using (23). The most likely hidden state is then determined by choosing the hidden state that yields the highest probability:

$$Q_t^* = \arg\max_{q^i} p(Q_t = q^i|\mathbf{z}_{1:t}). \quad (24)$$

## 4 Experimental Results

### 4.1 Datasets

The two datasets employed in our experiments contain synchronised and static views, captured at a frame rate of 0.7 fps from uncalibrated and disjoint cameras installed at two busy underground stations. Each image frame has a size of 320 × 230 pixels.

### *4.1.1 Station A Dataset*

A snapshot of each of the 8 camera views and the camera topology of this station are depicted in Fig. 5. The two train platforms of this station are covered by three cameras each (Cam 1–6). The rest two cameras (Cam 7–8) monitor a connected concourse, which is far away from the two platforms. The video from each camera lasts over 19 hours from 5:28 am to 12:38 am the next day, giving a total of 153 hours of video footage. Typically, when a train arrives at one of the platform, passengers on the train get off and leave the platform whilst passengers waiting on the platform get into the train. Nonetheless, it is also common that some passengers remain staying at the platform to wait for a later train to a different destination.

### *4.1.2 Station B Dataset*

The camera topology of this station is shown in Fig. 6, alongside with sample images of 9 camera views. The station has a ticket hall and a concourse leading to two train platforms via escalators. Three cameras are placed in a ticket hall and two cameras are positioned to monitor the escalator areas. Both train platforms are covered by two cameras each. The video from each camera lasts around 20 hours from 5:42 am to 01:19 am the next day, giving a total of 177 hours of video footage. Typically, passengers enter from the main entrance, walk through the ticket hall or queue up for tickets (Cam 1), enter the concourse through the ticket barriers (Cam 2, 3), take the escalators (Cam 4, 5), and enter one of the platforms. The opposite route is taken if they are leaving the station. Apart from the two platforms in Cam 6–7 and Cam 8–9, the passengers may also proceed from the concourse to other platforms (not visible in the camera views) without taking the escalators. In addition, after getting off a train they may also go to a different platform without leaving the station.

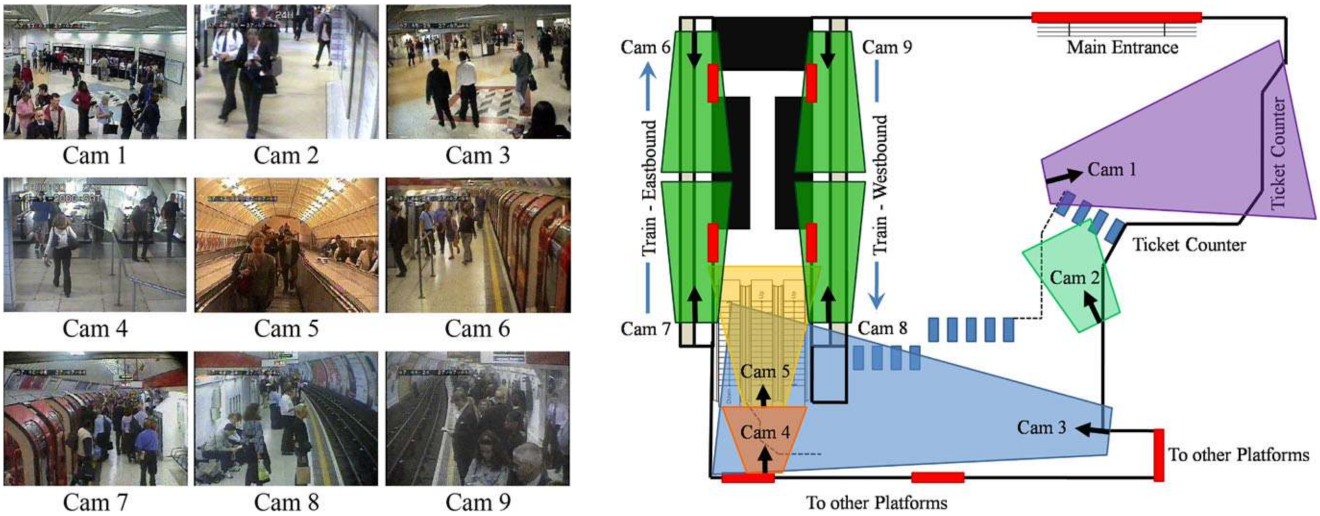


**Fig. 6** (Color online) The station layout and camera topology of Station B dataset. Entry and exit points are highlighted in *red bars*

The two datasets employed in our experiments are different in that Station A dataset has a larger time gaps between cameras, thus it is more challenging for the person re-identification task. Whilst Station B dataset features more diverse scenes and complex activities, hence it is more ideal for experiments in topology inference. In general, both datasets are difficult in several aspects:

(i) Complexity and diversity of the scenes. Activities observed in the scenes take place at ticket hall, concourse, train platforms and escalators, which are thus very different in nature.
(ii) Low video temporal and spatial resolution.
(iii) The lighting conditions are very different across camera views.
(iv) Heavy inter-object occlusions due to enormous number of objects in the scene especially during peak hours.
(v) Complex crowd dynamics, e.g. passengers may appear in a group or individually, remain stationary at any point of the scenes, or not get on an arrived train.
(vi) Only limited areas of the two large underground stations are covered by the cameras. In particular, there are multiple entry and exit points that are not visible in the camera views. This increases the uncertainties in the interpretation of the observed activities.

### 4.2 Background Subtraction

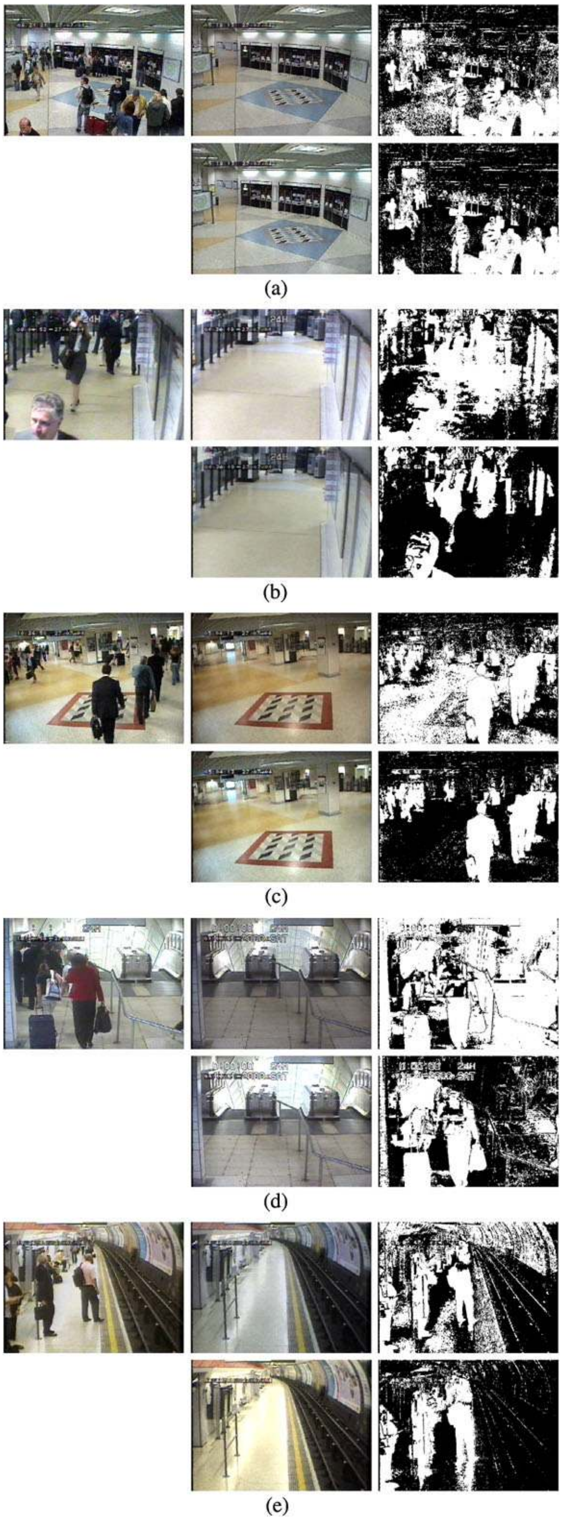
A comparison between the proposed mean-shift based background subtraction method and the frame differencing based method in Loy et al. (2009) was carried out. Some qualitative results are shown in Fig. 7. As can be seen, foreground masks yielded by the proposed method is noticeably better. Apart from qualitative evaluation, we also performed quantitative evaluation on both methods on topology inference task. The results are reported in Sect. 4.4.

### 4.3 Activity-based Scene Decomposition

We used 5000 frames ($\approx$2-hour in length) from each camera view for activity-based scene decomposition. In particular, the eight camera views from Station A dataset were automatically decomposed into 62 regions (Fig. 8). Whilst the nine camera views from Station B dataset were decomposed into 96 regions (Fig. 9). As can be seen from Fig. 8 and Fig. 9, the camera views were decomposed automatically into semantically meaningful regions in spite of the heavy inter-object occlusions and low temporal resolution. For instance, the areas corresponding to the train tracks and platforms formed distinctive regions. The sitting areas (e.g. regions 3 and 7 of Station A dataset, regions 80 and 86 of Station B dataset) were also segmented from areas where people standing or walking. Another example is the different escalators exits (regions 40, 43, 46) in Station B dataset, which were clearly decomposed into different regions in accordance to the object dynamics.

We performed both qualitative and quantitative comparisons between scene decomposition method introduced by Li et al. (2008) and our method. The two methods differ mainly in their feature representations, i.e. time-series representation in our method and Bag of Words representation in that of Li et al. (2008).

(i) Qualitative result: we found that our method yielded more meaningful region boundaries. Some results are shown in Fig. 10. As can be seen from most of the camera views depicted in column (a) of Fig. 10, the train track regions were clearly separated from the platform regions using the time-series representation. In contrast, some train track areas and platforms were segmented as a single region using Bag of Words representation (column (b) in Fig. 10).


(a)

(b)

(c)

(d)

(e)

**Fig. 7** The figure shows background models (the *second column*) and foreground masks (the *third column*) yielded by frame differencing method without background adjustment (the *first row*) and our approach (the *second row*) on frames with abrupt global intensity level change (the *first column*)

(ii) Quantitative result: It is difficult to provide quantitative result on activity-based scene decomposition as the correct region segmentation is subjective, especially when the segmentation is not based on visual information but activity patterns observed over time. Therefore, we performed quantitative evaluation on a synthetic dataset, in which the segmentation ground truth is known. In the dataset, all blocks ($10 \times 10$ pixels each) within the same region encompassed similar time-series patterns, whilst blocks located in different regions contained different time-series patterns. All time series had a length of 5000 and were corrupted by Gaussian noise. We measured the decomposition accuracy based on the agreement between our segmentation and the segmentation ground truth. As can be seen from Fig. 11, our time-series representation yielded higher accuracy, 99.73% compared to 83.83% obtained by using Bag of Words representation proposed by Li et al. (2008).

As we explained in Sect. 3.1, better performance is obtained because our time-series activity representation captures the temporal dynamics of activity while the Bag of Words representation utilised by Li et al. (2008) ignores the temporal order of the activity occurrences.

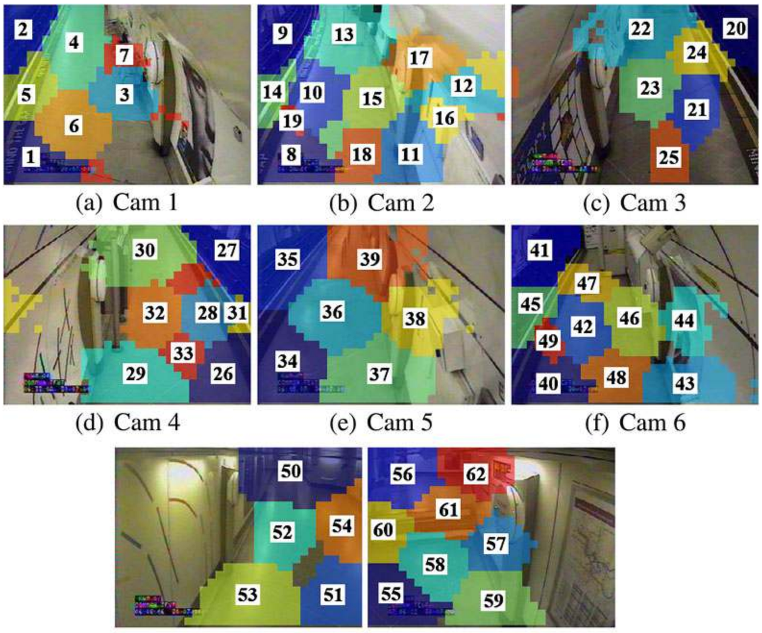
## 4.4 Activity Correlation Analysis

### *4.4.1 Discovering and Quantifying Regional Activity Correlation*

The proposed xCCA was compared with xCA and MCMC-BNSL for learning regional activity correlations. The regional activity affinity matrices **P** (15) (normalised to have a value range of [0, 1]) and the time delay matrices **D** (16) yielded by different methods for Station A and Station B datasets are shown in Figs. 12 and 13 respectively. Note that time delay matrix is not available for MCMC-BNSL since it can only discover zero-order temporal dependency between regional activities.

It can be seen from Figs. 12 and 13 that all methods are able to discover high correlations and relatively shorter time delays (except MCMC-BNSL) between regions from the same camera views (see the block structure along the diagonals of the **P** matrices). Importantly, a number of interesting cross-camera correlations were discovered and quantified accurately by xCCA. For instance, in Station B dataset, high correlation value (see Fig. 13(a)) with a time delay of 9 frames or 13 seconds (see Fig. 13(b)) are discovered by xCCA between region 46 (Cam 4) and region 51 (Cam 5). This corresponds to the frequently occurred inter-camera activity of passengers taking the upward escalator (with part of the escalator invisible from the view), and leaving from the escalator exit (see Fig. 1(a)). In comparison, although xCA can also learn these correlations, it tends to 'over-correlate'

**Fig. 8** Station A dataset: activity-based scene decomposition results

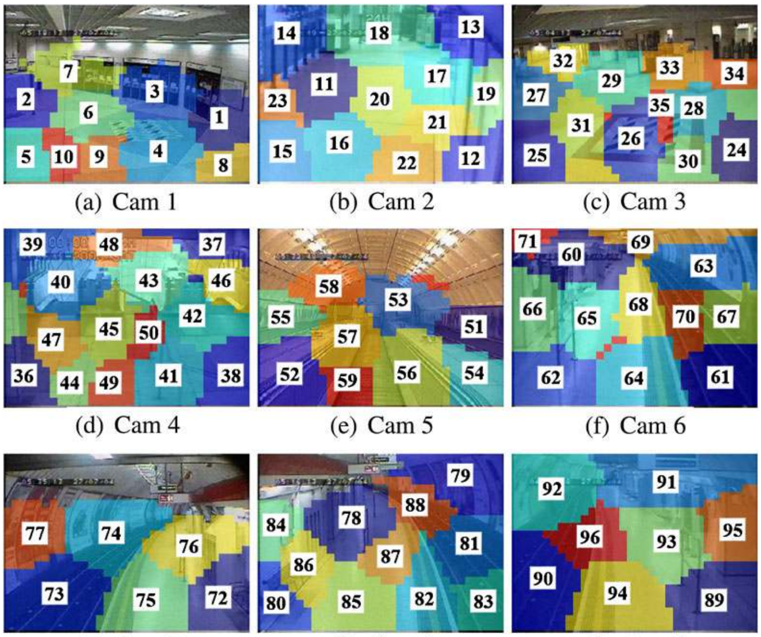


(a) Cam 1 (b) Cam 2 (c) Cam 3

(d) Cam 4 (e) Cam 5 (f) Cam 6

(g) Cam 7 (h) Cam 8

**Fig. 9** Station B dataset: activity-based scene decomposition results



(a) Cam 1 (b) Cam 2 (c) Cam 3

(d) Cam 4 (e) Cam 5 (f) Cam 6

(g) Cam 7 (h) Cam 8 (i) Cam 9

regions, i.e. detect correlations that do not exist (e.g. region pairs 3–91 and 18–48). In contrast, MCMC-BNSL revealed few and also incorrect correlations (e.g. region pairs 13–48 and 9–91) with a lot of miss-detections (e.g. region pairs 46–51 and 40–55 which correspond to passenger getting upward and downwards using the escalators respectively).

### 4.4.2 *Camera Topology Inference*

Given the regional activity affinity matrices and time delay matrices yielded by different methods, we generated the camera topologies $\boldsymbol{\Phi}$ by following the steps described in Sect. 3.3. The camera topologies for both Station A and Station B dataset are shown in Figs. 14 and 15 respectively. The inferred topologies are compared with the actual topology obtained manually and the numbers of missing edges (M) and redundant edges (R) are also shown in the two figures.

For both datasets, we observe that xCCA yielded the closest topology to the actual one based on the M and R metrics. It is not surprising to see that our xCCA outperformed MCMC-BNSL significantly. As we discussed earlier, the learned structure using MCMC-BNSL can only reveal zero-order temporal dependencies, i.e. co-occurrence relationships, between activities. Thus it cannot cope with more complex time delayed correlations that are common in multi-camera scenes. As expected, xCA yielded a number of redundant edges in both datasets. The better performance of xCCA compared to xCA is due to its ability in capturing the underlying mutual patterns of two regional activity time series by projecting them onto an optimal subspace. This is critical for analysing a busy public space such as an underground station where significant variations exist for correlated activities in different views caused by different camera view angles and uncertainties on activity time delays between views.

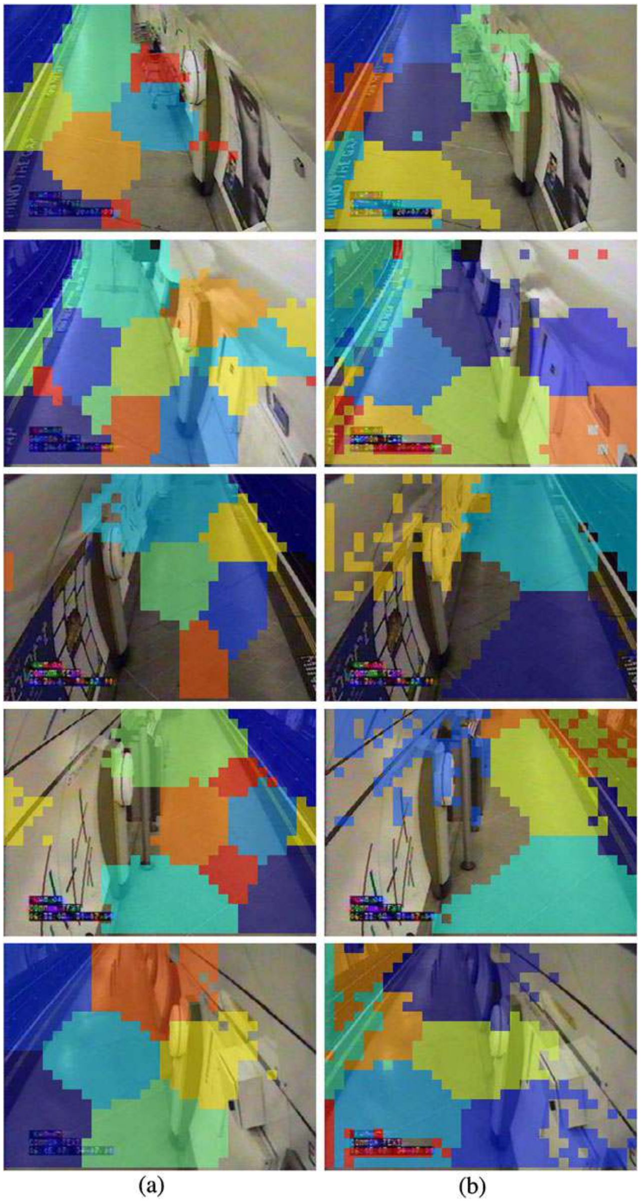


**Fig. 10** Better scene decomposition result (**a**) was obtained using our time-series activity representation and correlation based distance metric, as compared to the result (**b**) obtained using Bag of Words representation (Li et al. 2008)

Let us discuss those missing and redundant edges in the topologies. For Station A dataset, all methods except xCA (which tends to 'over-correlate') failed to infer the connection between Cam 7 and Cam 8 because the area in Cam 8 adjacent to Cam 7 is too far away from the camera (at the end of the concourse). In addition, there are four entry/exit points in the field of view of Cam 7 leading to spaces not covered by Cam 8 (see Fig. 5). This weakened the correlation between these two camera views and explains why the edge was miss-detected. Similarly, all methods except xCA failed to infer the connection between Cam 3 and 4 in Station B data as the connection point is too far away from the field of view as well as the existence of multiple entry/exit points. All methods inferred additional edges for camera pairs 1–3 and 4–6 in Station A dataset. Again this is not unexpected. Specifically, although they are not directly adjacent to each other (e.g. as shown in Fig. 5, Cam 1 is adjacent to Cam 2 which is then next to Cam 3), they cover the same platforms therefore sharing a number of common activities which are highly correlated, e.g. the arrival/departure of trains, passenger getting on/off trains.

To demonstrate the importance of activity-based scene decomposition on topology inference, we also performed xCCA without scene decomposition, i.e. the activities within each camera view as a whole are correlated with those in other camera views to infer the camera topology. The results are shown in Figs. 14(e) and 15(e) which suggest that without scene decomposition, even the proposed xCCA would not be able to learned the correct camera topology.

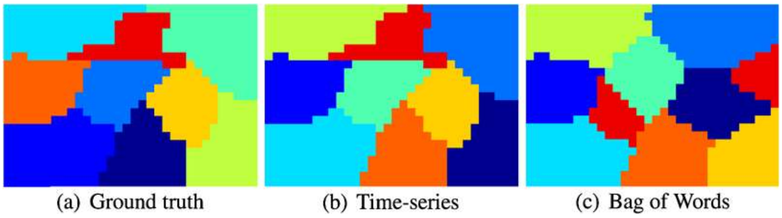
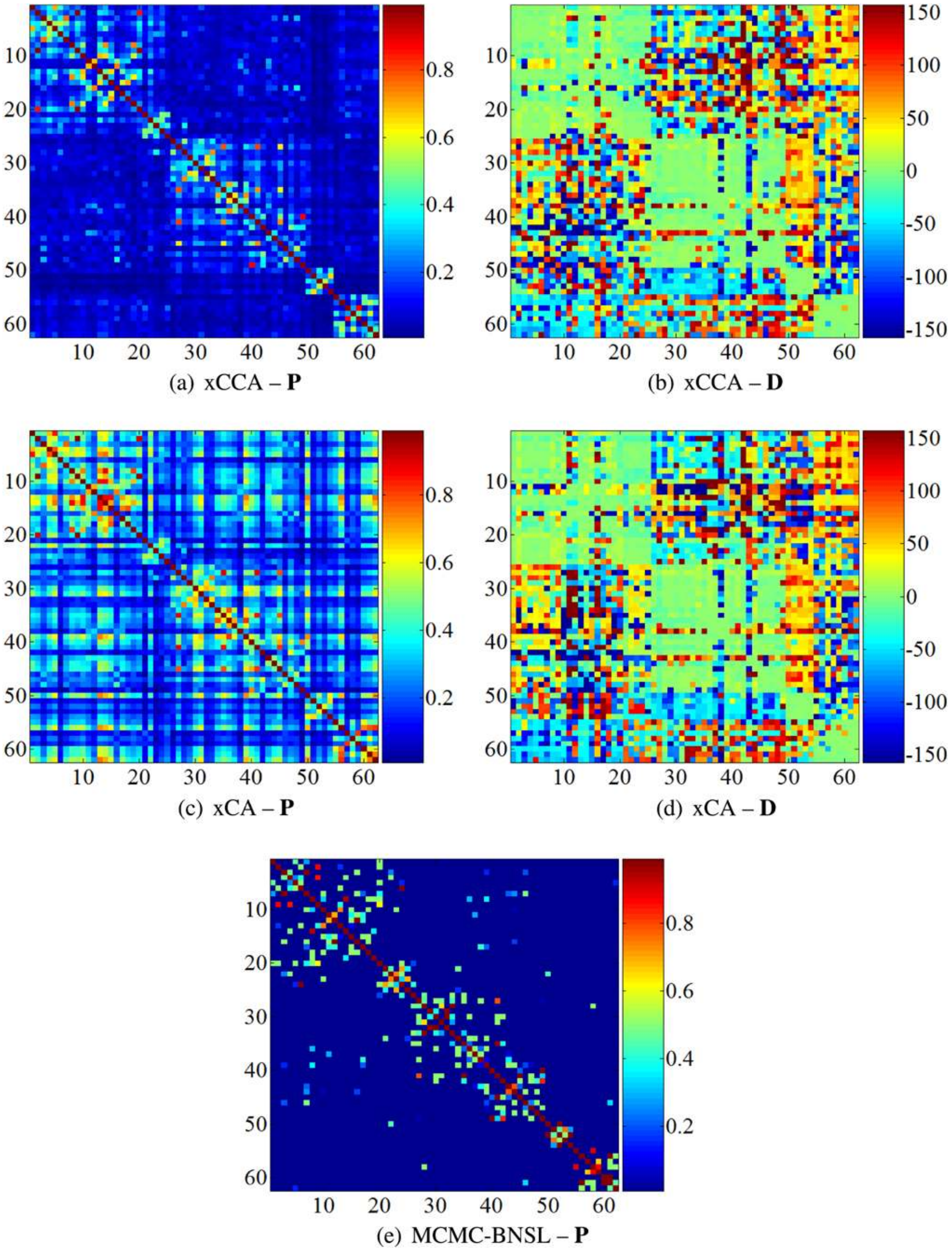


**Fig. 11** Quantitative comparison between our time-series representation (decomposition accuracy = 99.73%) against Bag of Words representation (Li et al. 2008) (decomposition accuracy = 83.83%) on a synthetic dataset



(a) xCCA – **P**

(b) xCCA – **D**

(c) xCA – **P**

(d) xCA – **D**

(e) MCMC-BNSL – **P**

**Fig. 12** Station A dataset: regional activity affinity matrices **P** (normalised to have a value range of [0, 1]) and the associated time delay matrices **D** obtained using xCCA, xCA and MCMC Bayesian network structure learning

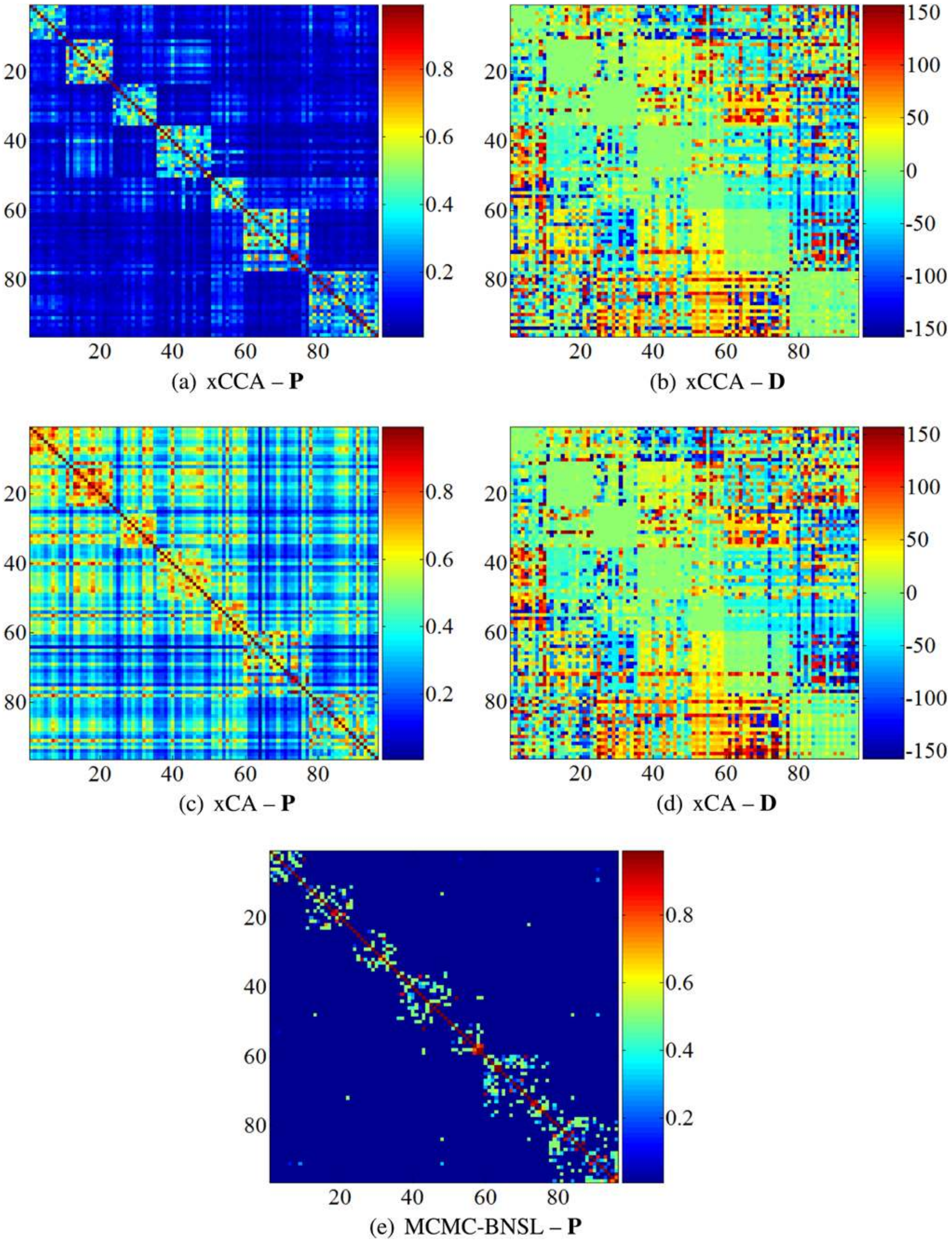
### 4.4.3 *Comparison with Method Proposed in Loy et al.* (2009)

In our previous work (Loy et al. 2009), a naive background subtraction method based on frame differencing is employed. In addition, camera topology is inferred solely based on correlation strength. Experiments were conducted on Station B dataset to compare the approach reported in Loy et al. (2009) and the new method proposed in this study on topology inference. The results are shown in Fig. 16.

As can be seen from Fig. 16(a), poor topology was inferred with naive background subtraction method and based only on correlation strength (method in Loy et al. 2009). In contrast, as shown in Fig. 15(b), camera topology inferred using our method produces fewer missing and redundant edges.

Even if we exploited both correlation and time delay, topology inferred with naive background subtraction method still consists of a number of missing and redundant edges (Fig. 16(b)). On the other hand, if we employed robust background subtraction method, poor topology was still obtained when the topology was inferred based solely on correlation strength (Fig. 16(c)). These results demonstrate that robust background modelling as well as the use of both correlation strength and time delay play important

**Fig. 13** Station B dataset: regional activity affinity matrices **P** (normalised to have a value range of [0, 1]) and the associated time delay matrices **D** obtained using xCCA, xCA and MCMC Bayesian network structure learning



(a) xCCA – **P**

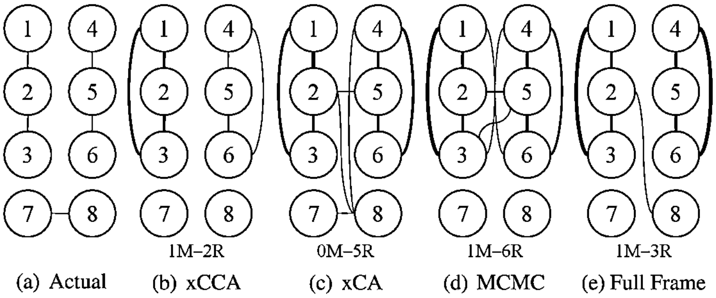
(b) xCCA – **D**

(c) xCA – **P**

(d) xCA – **D**

(e) MCMC-BNSL – **P**

**Fig. 14** Station A dataset: xCCA yielded the closest topology to the actual one as compared to other methods. M = missing edges, R = redundant edges



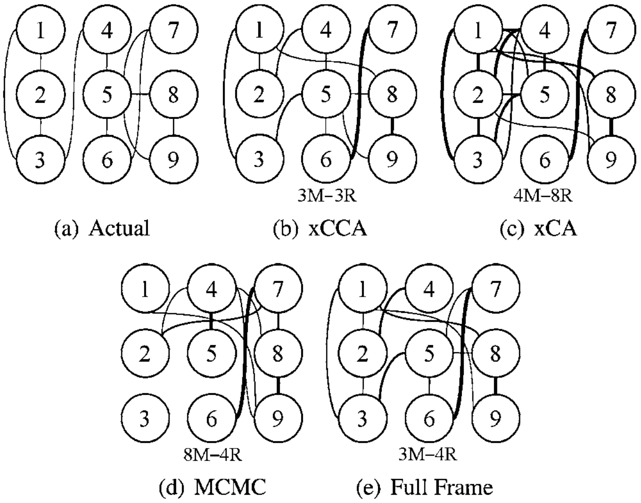
(a) Actual

(b) xCCA

(c) xCA

(d) MCMC

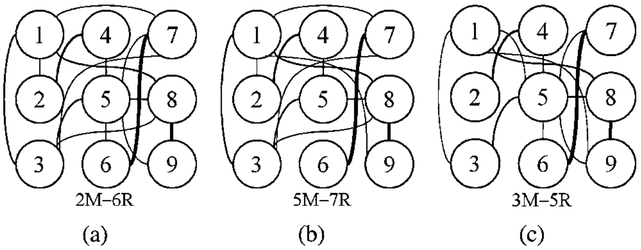
(e) Full Frame



(a) Actual (b) xCCA (c) xCA

(d) MCMC (e) Full Frame

**Fig. 15** Station B dataset: xCCA yielded the closest topology to the actual one as compared to other methods. M = missing edges, R = redundant edges



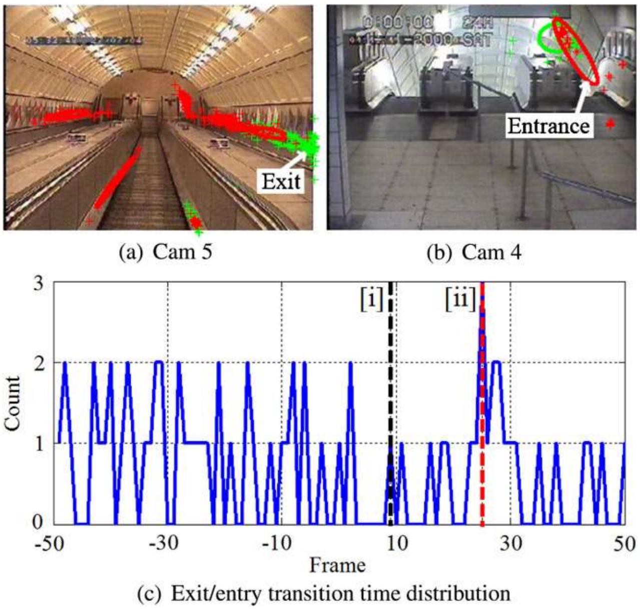
(a) (b) (c)

**Fig. 16** Camera topology inferred on station B dataset (**a**) without robust background subtraction + using correlation alone, (**b**) without robust background subtraction + using both correlation and time delay and (**c**) with robust background subtraction + using correlation alone

roles in making the proposed method scalable to challenging and complicated multi-camera scenes such as the one in the Station B dataset.

### 4.4.4 *Comparison with Tracking-based Method*

To highlight the inadequacy of tracking-based topology inference approach, we compared our results with a method proposed by Makris et al. (2004). In particular, tracking was performed on Cam 4 and Cam 5 of Station A dataset using a state of the art multi-object tracker (Chen et al. 2005). The starting and ending points of individual trajectories were clustered using Gaussian Mixture Model (GMM) to automatically locate the entry and exit zones, as shown in Fig. 17(a–b). Two entry and exit zones that correspond to the upward escalator and exit were selected and the corresponding exit/entry transition time distribution was plotted in Fig. 17(c). A peak in the transition time distribution at 25 frames suggests the existence of inter-zone connection. To verify this result, we manually recorded the amount of time taken by 50 objects passing across Cam 4 and Cam 5.



(a) Cam 5 (b) Cam 4

(c) Exit/entry transition time distribution

**Fig. 17** (**a**) Passengers leave the field of view of Cam 5 from a zone marked with 'Exit' and (**b**) enter Cam 4 from a zone marked with 'Entrance'. (**c**) The exit/entry transition time distribution for selected pairs of zones obtained using tracking-based method proposed by (Makris et al. 2004). *Dotted lines* labelled as [i] at 9 frames and [ii] at 25 frames represent the time delays between the selected pairs of zones estimated using our method and the tracking-based method respectively. The average time delay obtained from manual observations is 9.12 frames

It is observed that on average, an object took 9.12 frames to pass through the two zones. This demonstrates that the tracking-based method failed to estimate the correct transition time. The failure of the tracking-based method is mainly due to the difficulty in performing object tracking in low-frame rate video featured with heavy occlusion. The resultant fragmentation of object trajectories produced unreliable trajectory starting points and ending points, leading to inaccurate estimation of the entry/exit zones and transition time distribution. In comparison, using our method region 46 and 51 were automatically segmented which correspond the two entry and exit zones. The time delay between the two regions was estimated using xCCA as 9 frames, which is very close to the manual observation.

## 4.5 Context-aware Person Re-identification

In this experiment, we compared the performance of matching people across camera views using colour histogram (CH) alone, CH+xCA, and CH+xCCA. Note that MCMC-BNSL was excluded from this experiment since it is not able to quantify the time delayed relationships between regional activities.

The probe set consists of 250 individuals which were matched against a gallery set of 1800 people extracted from Station A dataset. The image of each person was manually

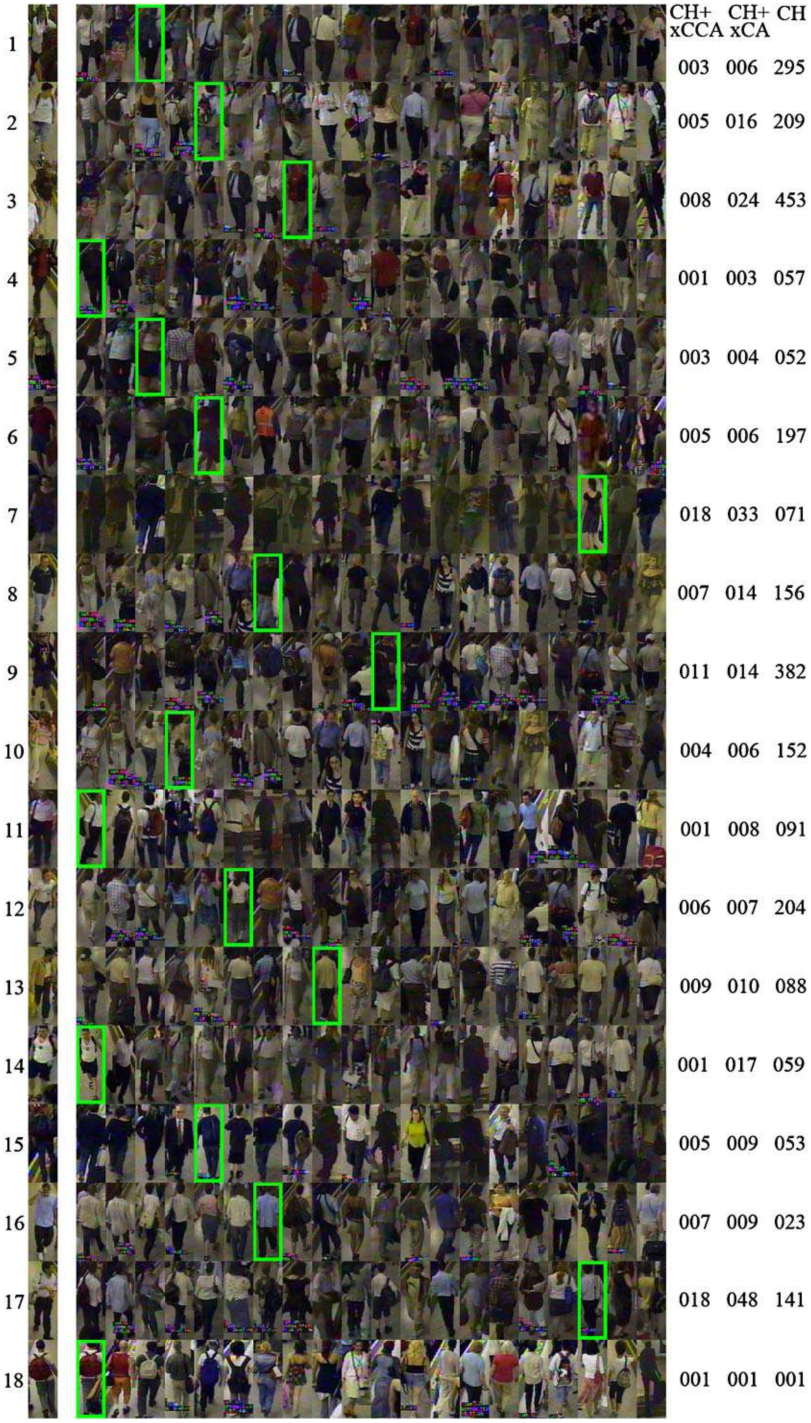


**Fig. 18** (Color online) Example queries selected from the person re-identification experiment. The first image in each row is a probe image. It is followed by top 20 results, sorted *from left to right* according to the ranking obtained using CH+xCCA, with the correct match highlighted using a *green bounding box*. The ranks returned by the evaluated methods are included at the *rightmost columns* for comparison. Note the visual ambiguity in the search space due to variations of pose, colours, lighting changes; as well as poor image quality caused by low spatial resolution

segmented and normalised to 48 × 128 pixels. Each image was then divided into $N_\mathrm{h} = 8$ horizontal strips equally, from which we extracted the concatenated colour histograms. To select the best setting for CH, we varied the number of bins $N_\mathrm{bin}$ from $\{8, 16, 32, 64, 128, 256\}$ and attempted both RGB and YUV colour spaces. It turned out that RGB colour space with 256 bins yielded the best result. Score returned by CH was computed as $\overline{S}^{a,b}_{\mathrm{bha}}$ (see Sect. 3.4), whilst score returned by CH + other methods were computed by (22). The factor $\alpha$ that defines the size of search window was set to 10.

Given a probe image, we computed the matching scores over all 1800 people and ranked them from the most likely match to least likely one. To examine the recognition rate at different ranks, a Cumulative Matching Characteristic (CMC) curve (Gray and Tao 2008) with a cut-off rank of 30 was plotted (Fig. 19). Example matches are given in Fig. 18. It can be seen that despite the poor image quality

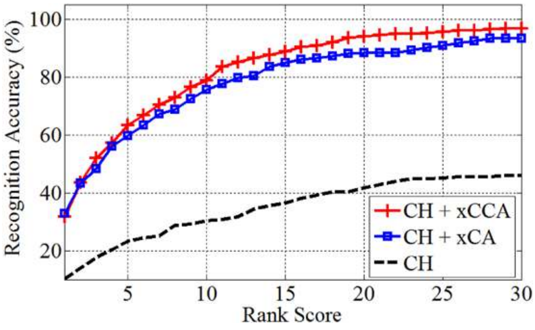


**Fig. 19** Cumulative Matching Characteristic (CMC) curve for CH+xCCA, CH+xCA and CH

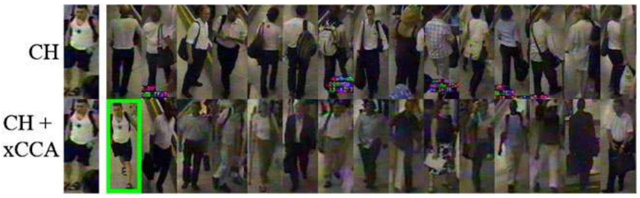


**Fig. 20** Comparing person re-identification result obtained using CH+xCCA and CH alone. Given the probe image at the leftmost column, CH+xCCA found the same person in another camera view at rank 1, whilst CH can only find the true match at rank 59. Ambiguities due to similar visual features presented by multiple objects are greatly reduced by introducing time delayed activity correlation as contextual information

and drastic feature variations across camera views, good results were obtained using both CH+xCCA and CH+xCA with CH+xCCA yielding better result. In comparison the result of using CH, i.e. visual appearance alone, was significantly worse. In particular, CH+xCCA yielded the best performance with approximately **94.00%** of the queries generated a true match in the top 20 rank, compared to 88.40% and 41.60% using CH+xCA and CH alone.

Without considering the activity correlation and time delay factor (CH alone), each person has to be compared against all possible candidates. However, as shown in Fig. 20, passengers in the underground stations tend to wear clothes with similar colours (e.g. white shirt with black trousers). It is thus difficult to match the same person over a large camera network by considering the colour information or any visual appearance information alone. On the contrary, with the inferred time delayed activity correlations employed as contextual information (CH+xCCA or CH+xCA), the search space and ambiguities were greatly reduced which has resulted in significantly better recognition rate. Note that one can also employ the time delays estimated using tracking-based methods (Makris et al. 2004; Tieu et al. 2005) as contextual information to reduce the search space. However, given low-frame rate videos with crowded scene, the estimation becomes inaccurate due to unreliable tracking (see Fig. 17(c)). Incorporating the time delays estimated thus will harm instead of improving the person re-identification performance.

## 4.6 Global Activity Modelling

Global activities were discovered by performing spectral clustering on the regional activity affinity matrix **P**. For each global activity, we employed 5000 frames to train an HMM following the steps described in Sect. 3.5. The test set which consists of the rest of the videos was used to evaluate the performance of a model in temporal segmentation. The segmentation result obtained using the proposed multi-view global activity analysis was compared with those from (i) individual single camera view without activity-based scene decomposition and (ii) single camera view with activity-based scene decomposition.

For Station A dataset, two global activities were learned by clustering **P** (see Fig. 12(a)), corresponding to the platform activities observed by Cam 1, 2, 3 and Cam 4, 5, 6 respectively. For Cam 1, 2, 3, it turned out that an HMM with two hidden states gave the best BIC score in the model selection process. The two phases have clear semantic meaning: phase one corresponds to the period when train is absent, whilst phase two is the period when train is present. We compared the phases inferred using the three methods with the ground truth. The accuracy yielded by single view analysis (Cam 3) without scene decomposition was 73.40%. The accuracy increased to 83.78% after we employed scene decomposition on the single view analysis, whilst the proposed method based on global activity analysis gave **97.90%**. Examples of the inferred phases by different methods and some example frames from the segmented phases are shown in Figs. 21 and 22 respectively. We obtained similar results on Cam 4, 5, 6—the accuracies were 86.59%, 91.70% and **94.09%** for methods without scene decomposition, with scene decomposition and scene decomposition + single view analysis.

We repeated the same procedures on Station B dataset. Several global activities were discovered, which include the platform activities monitored by Cam 6, 7 and Cam 8, 9, as well as escalator activities captured by Cam 4, 5. Here, we report the global activities that occurred at the escalator area, from which regions 46 and 51 were automatically detected as highly-correlated regions. A two-phase HMM was selected in the automatic model selection process. The two phases contain clear semantic meaning: phase one occurs when passengers on the escalator track approach the escalator exit, whilst phase two takes place when passengers move clear of the escalator exit area. In general, the results achieved on Station B dataset were relatively poorer compared to those obtained on Station A dataset as occlusion problem was more severe. In particular, single view analysis (Cam 4) without and with scene decomposition yielded similar results on this dataset, giving accuracies of 67.32% and 67.10% respectively. Scene decomposition failed to improve the result on single view analysis because region 46 in Cam 4 (see Fig. 9(d)) only occupies a small portion of many regions in the whole view. The activity model was thus dominated by activity patterns learned from other regions. Differing from the platform scene in Station A dataset, there are multiple global activities in the scene captured by Cam 4 and 5, namely passengers getting upwards and downwards. Using all regions blindly thus would not help improve the segmentation accuracy. Overall, the proposed method based on global activity analysis gave the best accuracy of **79.08**%. Examples of the inferred phases by different methods and some example frames from the segmented phases are shown in Figs. 23 and 24 respectively.

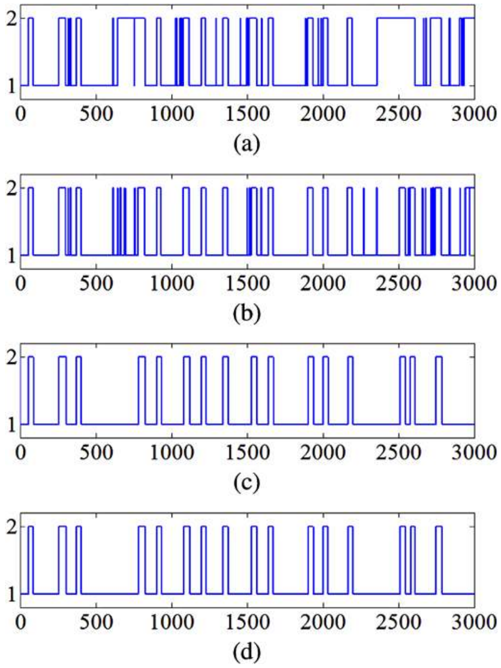


**Fig. 21** Station A dataset: example of phases inferred using (**a**) single view activity analysis without activity-based scene decomposition, (**b**) single view activity analysis with activity-based scene decomposition, and (**c**) multi-view global activity analysis. The ground truth is shown in (**d**). $Y$-axis represents the inferred phases and $X$-axis represents the frame index. Only 3000 frames from the test set are shown

The results on both datasets demonstrate the effectiveness of our global activity modelling based on the learning of regional activity correlations. In particular, single view activity analysis was susceptible to noise and visual ambiguities due to heavy occlusions and low frame rate. As compared to single view activity analysis, our global activity modelling utilises evidences collected from multiple correlated regions across camera view. It therefore reduced visual ambiguities, resulting in a more accurate segmentation result.

## 4.7 Computational Cost

The computational cost of each component of the proposed approach is analysed below:

- Activity-based scene decomposition: A total of $B(B-1)/2$ computations are required to obtain the pairwise correlation distances among local activity patterns of each block pair. The spectral clustering involves computation of eigenvectors of affinity matrix $\mathbf{A}$ (6) with computational complexity of $\mathrm{O}(B^3)$.
- Cross canonical correlation analysis: To obtain each element in a regional activity affinity matrix (15) and a time delay matrix (16), a regional time-series is shifted against another regional time-series and canonical correlation is performed at each shifting step. A total of $2T-1$ shifting steps are required where $T$ is the total of training frames. However, if we bound the maximum time delay

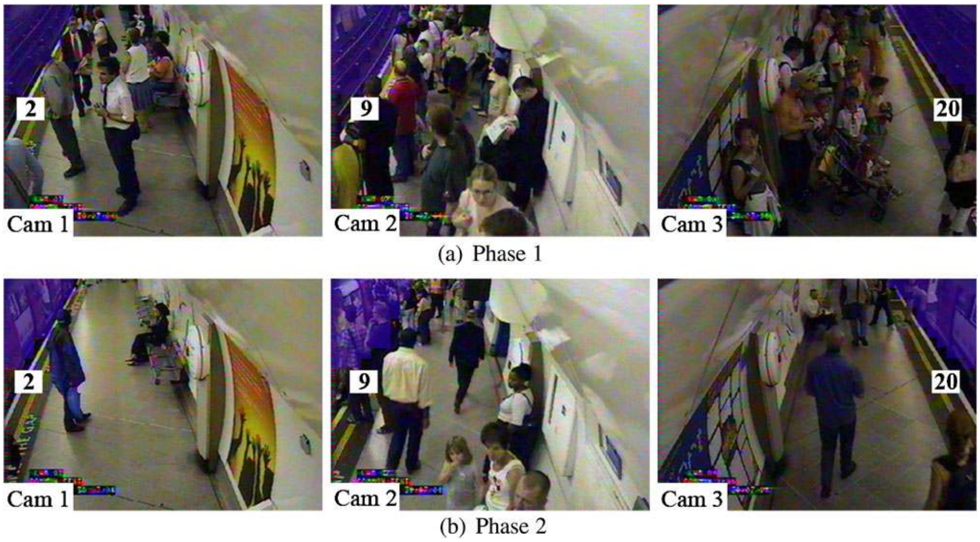


**Fig. 22** Station A dataset: example frames from the phases inferred using our global activity analysis. Phase 1: train is absent and passengers are waiting for train on the platform. Phase 2: train arrives and passengers get on/off the train

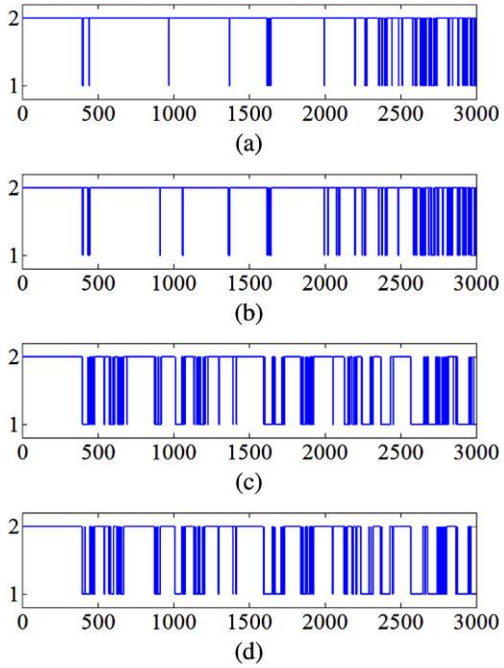
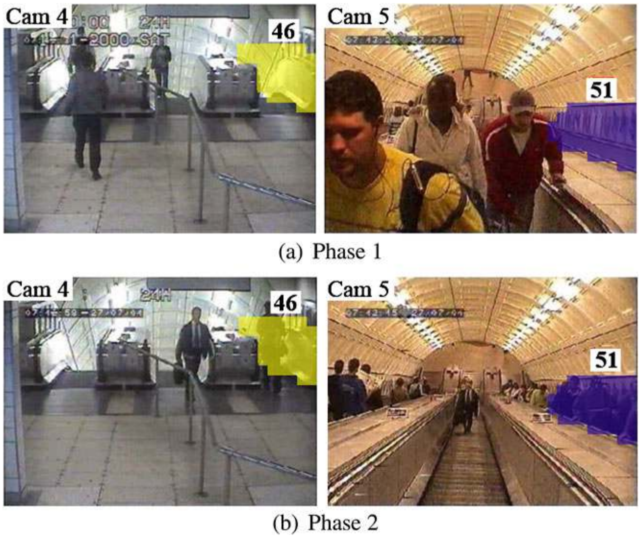


**Fig. 23** Station B dataset: example of phases inferred using (**a**) single view activity analysis without activity-based scene decomposition, (**b**) single view activity analysis with activity-based scene decomposition, and (**c**) multi-view global activity analysis. The ground truth is shown in (**d**). $Y$-axis represents the inferred phases and $X$-axis represents the frame index. Only 3000 frames from the test set are shown



(a) Phase 1

(b) Phase 2

**Fig. 24** Station B dataset: example frames from the phases inferred using global activity analysis. Phase 1: passengers on the escalator track are approaching the escalator exit; Phase 2: passengers move clear of the escalator exit area

as $\tau_{\max}$, the total number of shifting steps can be reduced to $\tau_{\max} - 1$. The computational cost of canonical correlation analysis is dominated by singular value decomposition (SVD). However, since $\Gamma$ in (13) is small, the complexity of SVD, $O(\Gamma^3)$ is low.

In practise, on a 2.8 GHz single-core machine, the computation of correlation distances in Matlab takes approximately 6 minutes on each camera view, whilst the spectral clustering implemented in C code requires 7 seconds. For xCCA, Matlab implementation takes approximately 12 minutes on Station A dataset and 30 minutes on Station B datasets.

### 4.8 Failure Modes, Limitations and Possible Extensions

There are a number of areas to improve on:

(i) As discussed in Sect. 4.4, if a region is located far away from camera, our method may fail to infer its connection with other regions in the camera network due to lack of visual information.

(ii) The activity correlations in the current framework are assumed to be static once learned. It is desirable to formulate a computationally tractable incremental learning framework to address the dynamic changes of regions and their correlations. This is a non-trivial problem and is part of our ongoing work.

(iii) The discovered activity correlations are limited to pairwise correlations and multiple dependencies in a global context are not considered. Thus, some redundant correlations caused by noise may affect the accuracy in activity understanding. This problem could be addressed by formulating a structure learning method for global optimisation on activity correlations.

(iv) It is assumed in our approach that there is only one delay time between two regions. In most cases, this assumption is valid because objects with different speeds (such as cars and pedestrians) appear in different regions, and their activities will thus be separated into different regions using our scene decomposition method. However, there are still cases where objects with different speeds and directions appear in the same location (e.g. middle of a traffic intersection). We did not consider modelling multiple delay modes because the features we used do not capture motion speed and direction due to videos with low temporal and spatial resolution. Nonetheless, if object speed and direction can be measured given videos with higher frame rate, it is straightforward to extend our xCCA framework to capture multiple delay and correlation modes. Specifically, we could decompose different directions into different bins (e.g. direction 001 = 0°–15°, direction 010 = 15°–30°, etc.). For each decomposed direction, we could then perform xCCA and model the correlation surface of different speeds and time delays.

There are several possible extensions to the proposed approach. Firstly, topology inference is performed in an unsupervised manner. Nevertheless, if coarse or partial information on a camera topology is available, it can be integrated into the proposed framework. For example, one can incorporate this information into training stage in a form of regularisation, e.g. increasing the correlation value between a region pair if their connection is known and penalising the correlation value if otherwise. Secondly, global anomaly detection, i.e. detection of abnormal events across camera views is not attempted in this study, which is an important application of multi-camera activity understanding. Finally, while we have demonstrated the effectiveness of our simple feature representation on challenging surveillance videos, including more sophisticated features is expected to improve the time-delayed correlation analysis when they can be computed reliably.

## 5 Conclusions

In this work we have presented a novel approach to multi-camera activity understanding by discovering and modelling the correlations with unknown time delays between activities observed within and across non-overlapping camera views. In particular, we introduced Cross Canonical Correlation Analysis to detect and quantify correlation and temporal relationships between partial observations across local regions. Experimental results have shown that the time delayed activity correlations are not only useful for inferring the spatial and temporal topology of a camera network, but also important as contextual information to facilitate more robust and accurate person re-identification, global activity interpretation, and video temporal segmentation. The proposed framework does not rely on either inter-camera or intra-camera tracking. Consequently, as demonstrated through our experiments, it can be applied to the most challenging surveillance videos, featured with heavy occlusions due to enormous number of objects in the scenes, as well as poor image quality caused by low video frame rate and image resolution.

## References

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99–109.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chen, T. P., Haussecker, H., Bovyrin, A., Belenov, R., Rodyushkin, K., Kuranov, A., & Eruhimov, V. (2005). Computer vision workload analysis: Case study of video surveillance systems. *Intel Technology Journal*, *9*(2), 109–118.

Cohen, N., Gatusso, J., & MacLennan-Brown, K. (2006). *CCTV operational requirements manual—is your CCTV system fit for purpose*? Home Office Scientific Development Branch, version 4 (55/06) edition.

Comaniciu, D., Ramesh, V., & Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *IEEE international conference on computer vision and pattern recognition*, pp. 142–149.

Du, Y., Chen, F., & Xu, W. (2007). Human interaction representation and recognition through motion decomposition. *IEEE Signal Processing Letters*, *14*(12), 952–955.

Friedman, N., & Russell, S. (1997). Image segmentation in video sequences: a probabilistic approach. In *Uncertainty in artificial intelligence*, pp. 175–181.

Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions of Information Theory*, *21*, 32–40.

Gheissari, N., Sebastian, T. B., Rittscher, J., & Hartley, R. (2006). Person reidentification using spatiotemporal appearance. In *IEEE international conference on computer vision and pattern recognition*, pp. 1528–1535.

Gong, S., & Xiang, T. (2003). Recognition of group activities using dynamic probabilistic networks. In *IEEE international conference on computer vision*, pp. 742–749.

Gray, D., & Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pp. 262–275.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, pp. 321–377.

Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., & Maybank, S. (2006a). Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 663–671.

Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., & Maybank, S. (2006b). A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(9), 1450–1464.

Javed, O., Rasheed, Z., Shafique, K., & Shah, M. (2003). Tracking across multiple cameras with disjoint views. In *IEEE international conference on computer vision*, pp. 952–957.

Javed, O., Shafique, K., & Shah, M. (2005). Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE international conference on computer vision and pattern recognition*, pp. 26–33.

Kendall, M., & Ord, J. K. (1990). *Time series*. Sevenoaks: Edward Arnold.

Kratz, L., & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *IEEE international conference on computer vision and pattern recognition*, pp. 1446–1453.

Kruegle, H. (2006). *CCTV surveillance: video practices and technology*. Stoneham: Butterworth-Heinemann.

Lee, L., Romano, R., & Stein, G. (2000). Monitoring activities from multiple video streams: establishing a common coordinate frame. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 758–768.

Li, J., Gong, S., & Xiang, T. (2008). Scene segmentation for behaviour correlation. In *European conference on computer vision*, pp. 383–395.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern Recognition*, *38*(11), 1857–1874.

Loy, C. C., Xiang, T., & Gong, S. (2009). Multi-camera activity correlation analysis. In *IEEE international conference on computer vision and pattern recognition*, pp. 1988–1995.

Makris, D., Ellis, T., & Black, J. (2004). Bridging the gaps between cameras. In *IEEE international conference on computer vision and pattern recognition*, pp. 205–210.

Murphy, K. P. (2002). *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California at Berkeley, Computer Science Division.

Neapolitan, R. E. (2003). *Learning Bayesian network*. New York: Prentice Hall.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. In *Advances in neural information processing systems*, pp. 849–856.

Oliver, N., Rosario, B., & Pentland, A. (2000). A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 831–843.

Pilet, J., Strecha, C., & Fua, P. (2008). Making background subtraction robust to sudden illumination changes. In *European conference on computer vision*, pp. 567–580.

Prosser, B., Gong, S., & Xiang, T. (2008). Multi-camera matching using bi-directional cumulative brightness transfer functions. In *British machine vision conference*.

Russell, D., & Gong, S. (2006). Minimum cuts of a time-varying background. In *British machine vision conference*, pp. 809–818.

Saleemi, I., Shafique, K., & Shah, M. (2009). Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(8), 1472–1485.

Stauffer, C., & Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 747–757.

Sung, K., Hwang, Y., & Kweon, I. (2008). Robust background maintenance for dynamic scenes with global intensity level changes. In *International conference on ubiquitous robots and ambient intelligence*, pp. 759–762.

Tieu, K., Dalley, G., & Grimson, W. E. L. (2005). Inference of non-overlapping camera network topology by measuring statistical dependence. In *IEEE international conference on computer vision*, pp. 1842–1849.

van den Hengel, A., Dick, A., & Hill, R. (2006). Activity topology estimation for large networks of cameras. In *IEEE conference on advanced video and signal based surveillance*.

Wang, X., Ma, X., & Grimson, W. E. L. (2009). Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(3), 539–555.

Wang, X., Tieu, K., & Grimson, W. E. L. (2010). Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 56–71.

Xie, B., Ramesh, V., & Boult, T. (2004). Sudden illumination change detection using order consistency. *Image and Vision Computing*, *22*(2), 117–125.

Yang, Y., Liu, J., & Shah, M. (2009). Video scene understanding using multi-scale analysis. In *International conference of computer vision*.

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. In *Advances in neural information processing systems*, pp. 1601–1608.

Zelniker, E. E., Gong, S., & Xiang, T. (2008). Global abnormal behaviour detection using a network of CCTV cameras. In *IEEE international workshop on visual surveillance*.

Zheng, W., Gong, S., & Xiang, T. (2009). Associating groups of people. In *British machine vision conference*.

Zhou, H., & Kimber, D. (2006). Unusual event detection via multi-camera video mining. In *IEEE international conference on pattern recognition*, pp. 1161–1166.

Zivkovic, Z., & van der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, *27*(7), 773–780.