Multi-Camera Activity Correlation Analysis

Chen Change Loy, Tao Xiang and Shaogang Gong
School of Electronic Engineering and Computer Science
Queen Mary University of London, London E1 4NS, UK
{ccloy,txiang,sgg}@dcs.qmul.ac.uk

Abstract

We propose a novel approach for modelling correlations between activities in a busy public space captured by multiple non-overlapping and uncalibrated cameras. In our approach, each camera view is automatically decomposed into semantic regions, across which different spatio-temporal activity patterns are observed. A novel Cross Canonical Correlation Analysis (xCCA) framework is formulated to detect and quantify temporal and causal relationships between regional activities within and across camera views. The approach accomplishes three tasks: (1) estimate the spatial and temporal topology of the camera network; (2) facilitate more robust and accurate person re-identification; (3) perform global activity modelling and video temporal segmentation by linking visual evidence collected across camera views. Our approach differs from the state of the art in that it does not rely on either intra or inter camera tracking. It therefore can be applied to even the most challenging video surveillance settings featured with severe occlusions and extremely low spatial and temporal resolutions. Its effectiveness is demonstrated using 153 hours of videos from 8 cameras installed in a busy underground station.

1. Introduction

A typical public space video surveillance system deploys a network of cameras to monitor a wide-area scene, *e.g.* underground station, airport, and shopping complex. For global activity monitoring and situation awareness, it is crucial to detect and model correlations among object activities observed across camera views. Specifically, discovering multi-camera activity correlations will lead to understanding of both the spatial topology (*i.e.* between-camera spatial relationships) and more importantly the temporal topology of a camera network, that is, we wish to discover if an activity takes place in one camera view, what other activities it may cause in different camera views after what time delay. Discovering and modelling such activity correlations among multiple camera views from data directly can facilitate person re-identification across disjoint camera views [2, 3, 5–7, 14] and global activity analysis [9, 18, 21].

Previous multi-camera activity analysis methods [6, 12, 16, 18, 21] rely on either intra-camera (within camera) tracking to detect exit and entry events for modelling transition time distribution, or both intra-camera tracking and inter-camera (between cameras) object association to solve the trajectory correspondence problem. To achieve intra-camera tracking, these approaches assume reliable object localisation and detection as well as smooth object movement. However, both assumptions are largely invalid for activities captured by CCTV cameras in public spaces typically with *crowded scenes* and *low spatial and temporal resolution* (Figure 1) [1]. In a crowded environment, the sheer number of objects with complex activities causes severe inter-object occlusions continuously, making intra-camera tracking difficult, if not impossible. The problem is compounded by the low temporal resolution of surveillance video, where large spatial displacement is observed in moving objects between consecutive frames. Without reliable intra-camera tracking, one cannot detect exit and entry events, required by existing techniques for both transition time distribution modelling and inter-camera tracking.

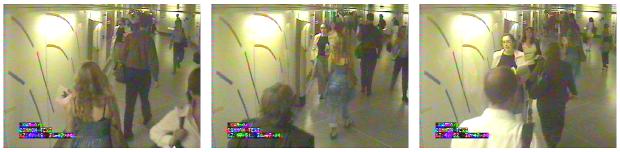


Figure 1. Three consecutive frames from a typical public space CCTV video where at a frame rate of 0.7 fps, an object can pass through the whole view in three frames.

In this work, we propose a novel approach for modelling correlations between multi-camera activities without either intra-camera tracking or inter-camera object correspondence. In our approach, each camera view is automatically decomposed into semantic regions, across which different spatio-temporal activity patterns are observed. A novel Cross Canonical Correlation Analysis (xCCA) framework is formulated to discover and quantify temporal and causal relationships of *arbitrary order* among these multi-camera regional activities. The framework addresses three fundamental problems in distributed multi-camera networks: (1)

[1] Current surveillance systems rarely record videos of more than 5 fps due to limited data bandwidth and storage space.

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

estimate the spatial and temporal topology of a network of disjoint and uncalibrated cameras; (2) facilitate more robust and accurate person re-identification among different camera views; (3) infer and model global activity, and perform video temporal segmentation by correlating regional activities from different camera views.

To our knowledge, this study is the first attempt to infer multi-camera activity correlation in a crowded scene with low-frame rate video. This is made possible by analysing the underlying spatial and temporal correlation of regional activities, as opposed to object centred activities, without relying on tracking. Moreover, the proposed approach is novel in its ability to discover and quantify the temporal and causal relationships of arbitrary order among local activities across different camera views, and to provide global activity inference from disjoint views. This approach is completely unsupervised. Although we focus on disjoint and uncalibrated cameras in this study, the proposed approach can be readily used for camera views with any degrees of overlapping. We demonstrate the effectiveness of the proposed approach using 153 hours of videos captured at 0.7 fps from a busy underground station with eight camera views, all of which feature crowded scene and complex activities.

2. Related Work

Most previous work on multi-camera activity analysis with non-overlapping views focuses on two related issues: (1) object re-identification (inter-camera object tracking) [2, 6, 7, 14] and (2) camera network topology inference [6, 12, 16]. To infer the topology of a network of non-overlapping cameras, previous work generally follows two approaches: (1) solving the inter-camera correspondence problem by matching object visual appearance, velocity, and transition time between cameras; (2) exploiting the distribution of transition times of entry and exit events without inter-camera object correspondence.

Javed *et al.* [6] took the first approach which relies on the availability of reliable visual features and motion trends from targets to achieve inter-camera trajectory association. Their method suffers from drastic feature variations across camera views due to illumination changes, camera orientation, and dynamic appearance of clothing. Although various strategies have been proposed [2, 7] to adapt and to rectify the feature variations, the object correspondence/ trajectory association problem remain a notoriously difficult problem. The second approach was adopted by [12, 14, 16] to avoid solving the correspondence problem by modelling the transition time between entry and exit events detected in different camera views.

All existing topology inference methods are based on the assumption that individual exit and entry events can be detected given reliable intra-camera tracking. However, object tracking in a busy public scene is far from being reliable especially when video's spatial and temporal resolutions are low. Our approach overcomes this problem by modelling temporal and causal relationships among activities without relying on either intra or inter camera tracking. This approach is scalable to multi-camera activity analysis even under the most challenging public scene viewing conditions.

Beyond object re-identification and topology inference, work on global activity analysis by linking visual evidence from multi-camera views has also been reported more recently [18, 21]. Zelniker *et al.* [21] stitch trajectories of the same objects observed in different views to form so called "global trajectories", which is then followed by the same trajectory clustering method developed for single view activity analysis [17]. In contrast, Wang *et al.*'s method [18] relies only on intra-camera tracking. Trajectories in different camera views are grouped into global activities using a topic model based on Latent Dirichlet Analysis (LDA), with a restriction that only co-occurrence relationships between activities can be modelled. The xCCA framework proposed in this work is capable of capturing temporal and causal relationships of *arbitrary order*. Moreover, unlike Wang *et al.*'s method [18], our approach is able to cope with coexistence of large number of objects both within and across camera views.

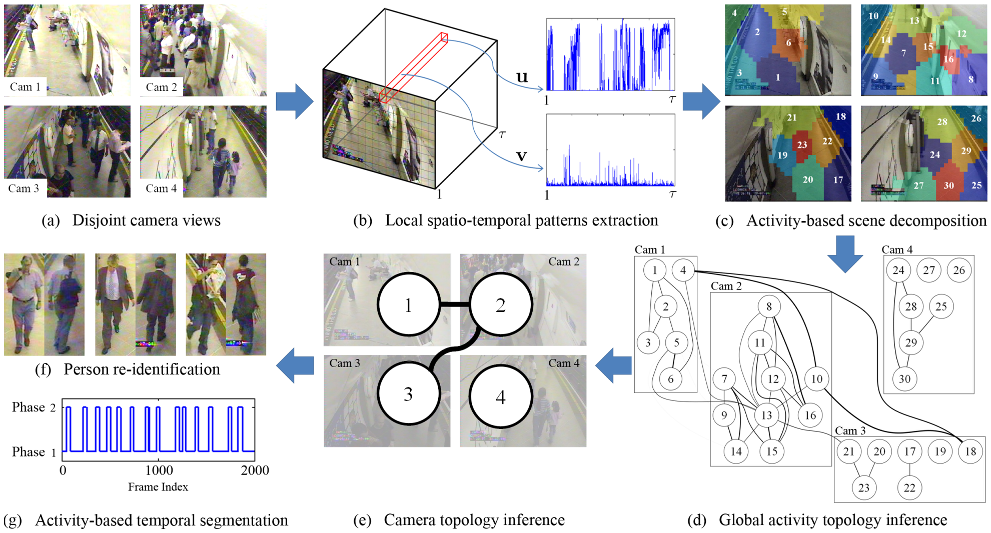
3. Multi-Camera Activity Correlation

The proposed approach consists of two main components: activity-based semantic scene decomposition and Cross Canonical Correlation Analysis (xCCA) for global activity topology inference. The overall framework is illustrated in Figure 2.

3.1. Semantic Scene Decomposition

In a complex public space scenario, each camera view naturally consists of multiple semantic regions. Activity patterns observed within each region are similar to each other whilst being dissimilar to those occurring in other regions. For instance, in a train platform as shown in Figure 2(a), object activities differ significantly at the track area, sitting areas, platform areas near the track and far away from the track. It is therefore necessary to decompose each camera view into semantic regions before correlation between activities across different regions and different camera views can be established.

We aim to decomposing the scene observed by K cameras into L semantic regions with each region being assigned a unique number ranging from 1 to L (Figure 2(c)). Consequently, the kth camera view in the network contains L_k semantic regions with L_k being determined automatically and $\sum_{k=1}^{K} L_k = L$. To this end, we first divide each camera view into blocks of 10×10 pixels (Figure 2(b)). Foreground pixels are detected using a background subtraction method [15]. To cope with illumination changes,



(a) Disjoint camera views (b) Local spatio-temporal patterns extraction (c) Activity-based scene decomposition

(f) Person re-identification

(g) Activity-based temporal segmentation (e) Camera topology inference (d) Global activity topology inference

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

we update the background model every fixed time interval (5000 frames), and reduce the chroma noise by performing colour correction in YUV colour space. Each foreground pixel is classified 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). A local pixel block activity pattern is then represented as a time series signal composed of two components: $\mathbf{u} = \{u_t : t \in \tau\}$, where u_t is the percentage of static foreground pixels within the block at time t and τ is the total number of frames used in the inference process; and $\mathbf{v} = \{v_t : t \in \tau\}$, where v_t is the percentage of pixels within the block that are classified as moving foreground. Note that more sophisticated features such as optical flow can be considered. However, given low spatial and temporal resolution, u_t and v_t are the only reliable features extractable.

After feature extraction, we group blocks into semantic regions according to the similarity of local spatio-temporal activity patterns represented as $\mathbf{u}$ and $\mathbf{v}$. 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 [11], given as $\overline{r} = 1 - |r|$. In particular, $\overline{r} = 0$ if two blocks have strongly correlated local activity patterns, or $\overline{r} = 1$ otherwise. Subsequently, we construct an affinity matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where N is the total number blocks in the camera view, defined as:

$$A_{ij} = \begin{cases} \exp\left(-\frac{(\overline{r}^{\mathbf{u}}_{ij})^2}{2\sigma^{\mathbf{u}}_i\sigma^{\mathbf{u}}_j}\right) \exp\left(-\frac{(\overline{r}^{\mathbf{v}}_{ij})^2}{2\sigma^{\mathbf{v}}_i\sigma^{\mathbf{v}}_j}\right) \exp\left(-\frac{\|\mathbf{b_i},\mathbf{b_j}\|^2}{2\sigma_b^2}\right) \\ \quad \text{if } \|\mathbf{b_i} - \mathbf{b_j}\| \leq R \text{ and } i \neq j \\ 0 \quad \text{otherwise} \end{cases}, \tag{1}$$

The correlation distances of $\mathbf{u}$ and $\mathbf{v}$ between block i and block j are given by $\overline{r}^{\mathbf{u}}_{ij}$ and $\overline{r}^{\mathbf{v}}_{ij}$ respectively. $[\sigma^{\mathbf{u}}_i, \sigma^{\mathbf{u}}_j]$ and $[\sigma^{\mathbf{v}}_i, \sigma^{\mathbf{v}}_j]$ are the correlation scaling factors for $\overline{r}^{\mathbf{u}}_{ij}$ and $\overline{r}^{\mathbf{v}}_{ij}$ respectively. The correlation scaling factors are defined as the mean correlation distance between the current block and all blocks within a radius R. The 2-D 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 σ_b is defined as the mean spatial distance between the current block and all blocks within the radius R. Note that we only compare similarity within a fixed radius R to avoid under-fitting problem during segmentation [10]. The affinity matrix is then normalised according to $\overline{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is a diagonal matrix and $D_{ii} = \sum_{j=1}^{N} A_{ij}$. A spectral clustering algorithm [20] is then employed to segment each camera view into semantic regions with the optimal number of regions being determined automatically.

To facilitate a more precise topology inference, we remove any region that is populated by more than *TH* percentage of zero-activity (background) blocks. The percentage *TH* is set at 90% in this study. We note that the clustering result is governed by the choice of R. From our experiments, we found that consistent cluster formation is obtained with R set between 20-30. 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.* [10] but with a noticeable modification on how local activities are represented. Specifically, in our method local activities are represented as time series and correlation between them are used as the similarity measure. In comparison, a Bag of Words representation is adopted in [10], which ignores the temporal order information of a local activity and is thus less discriminative than our representation.

Given scene decomposition, a *regional spatio-temporal activity* observed over time is represented as a time-series of two components $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$, obtained by averaging $\mathbf{u}$ and $\mathbf{v}$ computed from each pixel block within a region.

3.2. Cross Canonical Correlation Analysis

For any pair of semantic regions in the entire camera network, two questions are to be answered: (1) are activities in these regions correlated? (2) if yes, what are the temporal and causal relationships among them? Correlations between regional activities across disjoint camera views are complex in that there is often an arbitrary temporal gap/delay between the times when a causing activity in one region taking place and the correlated/caused activity in another region being observed. This temporal gap is modelled as a temporal dependency of arbitrary order between two time-series representing any two regional activities. To this end, we formulate a new Cross Canonical Correlation Analysis (xCCA) to measure the correlation of regional activities as a function of an unknown time-lag applied to one of the two regional activity time-series.

xCCA is an extension of Canonical Correlation Analysis (CCA) proposed by Hotelling [4], with additional steps similar in nature to the standard cross-correlation analysis (xCA) [8]. This principally involves the shifting of one time-series and computes its canonical correlation with the other. Specifically, given two regional activity time series denoted as $\mathbf{x}$ and $\mathbf{y}$, CCA finds two sets of optimal basis vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ for $\mathbf{x}$ and $\mathbf{y}$ such that the correlation of the projections of them onto the basis vectors are mutually maximised. Let the linear combinations of canonical variates be $x = \mathbf{w}_x^\intercal \mathbf{x}$ and $y = \mathbf{w}_y^\intercal \mathbf{y}$, The canonical correlation ρ is defined as:

$$\begin{aligned}\rho &= \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\mathbf{w}_x^\intercal \mathbf{x}\mathbf{y}^\intercal \mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^\intercal \mathbf{x}\mathbf{x}^\intercal \mathbf{w}_x]E[\mathbf{w}_y^\intercal \mathbf{y}\mathbf{y}^\intercal \mathbf{w}_y]}} \\ &= \frac{\mathbf{w}_x^\intercal \mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^\intercal \mathbf{C}_{xx}\mathbf{w}_x \mathbf{w}_y^\intercal \mathbf{C}_{yy}\mathbf{w}_y}}\end{aligned}, \quad (2)$$

where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within-set covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively, whilst $\mathbf{C}_{xy}$ represents their between-set covariance matrix. The maximisation can be solved by setting the derivatives in Eqn. (2) to zero, yielding the following eigenvalue equations:

$$\left\{ \begin{array}{l} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x = \rho^2 \mathbf{w}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{w}_y = \rho^2 \mathbf{w}_y \end{array} \right. , \quad (3)$$

where the eigenvalues ρ^2 are the square canonical correlations and the eigen vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ are the basis vectors. We only need to solve one of the eigenvalue equations since the equations are related by:

$$\left\{ \begin{array}{l} \mathbf{C}_{xy}\mathbf{w}_y = \rho\lambda_x \mathbf{C}_{xx}\mathbf{w}_x \\ \mathbf{C}_{yx}\mathbf{w}_x = \rho\lambda_y \mathbf{C}_{yy}\mathbf{w}_y \end{array} \right. , \quad (4)$$

where

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\mathbf{w}_y^\intercal \mathbf{C}_{yy}\mathbf{w}_y}{\mathbf{w}_x^\intercal \mathbf{C}_{xx}\mathbf{w}_x}}. \quad (5)$$

The canonical correlation ρ measures how strong $\mathbf{x}$ and $\mathbf{y}$ are correlated in a co-current or zero-order sense. To measure correlations beyond zero-order, cross canonical correlation $\mathbf{z}_{xy}(t)$ is computed as a function of time-lag t:

$$\mathbf{z}_{xy}(t) = \frac{E[\mathbf{w}_x^\intercal \mathbf{x}\mathbf{y}(t)^\intercal \mathbf{w}_{y(t)}]}{\sqrt{E[\mathbf{w}_x^\intercal \mathbf{x}\mathbf{x}^\intercal \mathbf{w}_x]E[\mathbf{w}_{y(t)}^\intercal \mathbf{y}(t)\mathbf{y}(t)^\intercal \mathbf{w}_{y(t)}]}}, \quad (6)$$

where $\mathbf{y}(t)$ is obtained by shifting $\mathbf{y}$ by t and $t \in [1-\tau, \tau-1]$, *i.e.* $\mathbf{y}$ can be shifted either forward or backward. The maximum cross canonical correlation $z_{xy}^{\max}$ is then obtained by locating the peak value in function $\mathbf{z}_{xy}(t)$, whilst the temporal delay of the two regional activities can be computed as

$$T_{xy}^{\text{delay}} = \underset{t}{\operatorname{argmax}}\, \mathbf{z}_{xy}(t). \quad (7)$$

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 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 (MCMC) Bayesian network structure learning [13]. 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 a multi-camera scene. 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 view angles and the temporal delays between correlated activities across camera views.

3.3. Topology Inference

Given the ability to correlate regional activities, we wish to infer a global activity topology, which reflects correlations between all the semantic regions discovered in multiple camera views. For now we do not consider time delays of the correlated regional activities explicitly. That will be addressed further in Sec. 3.5 on global activity modelling. Instead we focus solely on the strength of the correlations. Specifically, each region is denoted as a node in the global activity topology. The thickness of an edge between any two nodes is determined by $z_{xy}^{\max}$ and reflects the strength of the correlation. An example of global activity topology is given in Figure 2(d). Once we have estimated the global activity topology, the camera topology can be inferred. In the camera topology, the edges between any two camera views are computed by averaging the inter-camera regional activity correlations. To reduce the influence of noise, we only consider the correlations greater than the mean inter-camera regional activity correlation. An example of inferred camera topology is given in Figure 2(e). Again, the thickness of the edges encodes the strength of the correlations.

3.4. Context-aware Person Re-identification

A more robust person re-identification can be achieved by incorporating the inferred global activity topology as visual context. A simple colour histogram feature is employed here for person re-identification. Note that more sophisticated features can be used [2, 3], but our focus here is to demonstrate the effectiveness of using the learned temporal and causal relationships between regional activities to reduce the search space and resolve ambiguities arisen from similar visual features presented by different objects. The similarity between two colour histograms H_a and H_b is computed using Bhattacharyya score [1] as follows:

$$S_{\text{bha}} = \sum_{i=1}^{N_{\text{bin}}} \sqrt{H_a^i \, H_b^i}. \tag{8}$$

The number of bins N_{bin} used in a histogram is set to 256. Each histogram bin is normalised using the total number of pixels in the colour image. Note that in order to compare two colour objects a and b, we must compute the Bhattacharyya score for three RGB channels. Thus the overall Bhattacharyya score $\overline{S}_{\text{bha}}$ is the average of similarity scores computed in all three channels. To incorporate the output of our Cross Canonical Correlation Analysis into the score computation, we compute the overall score as follows:

$$S_{ab} = \begin{cases} \overline{S}_{\text{bha}} z_{ab}^{\max} & \text{if } |t_{a,b}| \in [|T_{ab}^{\text{delay}}| \pm 0.5 T_{ab}^{\text{delay}}] \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where $z_{ab}^{\max}$ is the maximum cross-canonical correlation of the two semantic regions occupied by object a and object b respectively, $t_{a,b}$ is the time gap between the two objects being observed (could be negative corresponding to b being observed before a), and T_{ab}^{delay} is computed using Eqn. (7).

3.5. Global Activity Modelling

Correlated activities across multiple camera views should be modelled collectively. This is because by utilising visual evidence 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 global activity topology (Fig. 2(d)) is only concerned with the correlations of regional activities. It does not reveal either the contribution of these regional activities to the global activities or the temporal dynamics of the global activities. We therefore need to discover these global activities and build models for them.

A complex camera network can capture multiple global activities occurring simultaneously. These global activities are discovered and modelled by taking the following steps. (1) A regional activity affinity matrix $\mathcal{R} \in \mathbb{R}^{L \times L}$ is constructed, where L is the total number of regions in the camera views, and:

$$\mathcal{R}_{ij} = z_{ij}^{\max} \tag{10}$$

where $z_{ij}^{\max}$ is the cross canonical correlation between the ith and jth regions (see Sec. 3.2). (2) The same spectral clustering algorithm used in Sec. 3.1 is employed using $\mathcal{R}$ as input to discover global activities defined by correlated regional activities across camera views. Specifically, the formed clusters with the highest mean cross canonical correlations correspond to global activities composed of strongly correlated regional activities. (3) One regional activity in each cluster is set as the reference point to temporally align the activity patterns of other regions in the cluster in accordance to the respective T_{ij}^{delay} computed using xCCA. (4) The aligned regional activity patterns in every cluster, each represented as a 2-D time series (*i.e.* $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$), are used as input to a Multi-Observation Hidden Markov Model (MOHMM) [19] to model the temporal dynamic of each global activity.

The learned MOHMM for each global activity can be used for activity-based temporal segmentation. The objective is to segment a continuous video stream into phases based on 'what is happening' not only in this particular view but also in other correlated views [19]. To that end, for each discovered global activity, the optimal number of hidden states of the MOHMM is determined using Bayesian Information Criterion (BIC). When applying the learned model to unseen video streams, global activity phases are inferred for real-time temporal segmentation using online filtering.

4. Experimental Results

4.1. Dataset

Our dataset contains synchronised and static views, from eight uncalibrated and disjoint cameras installed in a busy underground station. The video from each camera lasts over 19 hours from 5:28am to 12:58am the next day, giving a total of 153 hours of video footage (or 384,000 frames) at a frame rate of 0.7 fps. Each image frame has a size of 320 × 230. Two train platforms are covered by 3 cameras each. The rest two cameras monitor a connected concourse, which is far away from the two platforms. The samples of each camera view and the topology are given in Figure 3.

All the scenes are crowded, especially during the peak

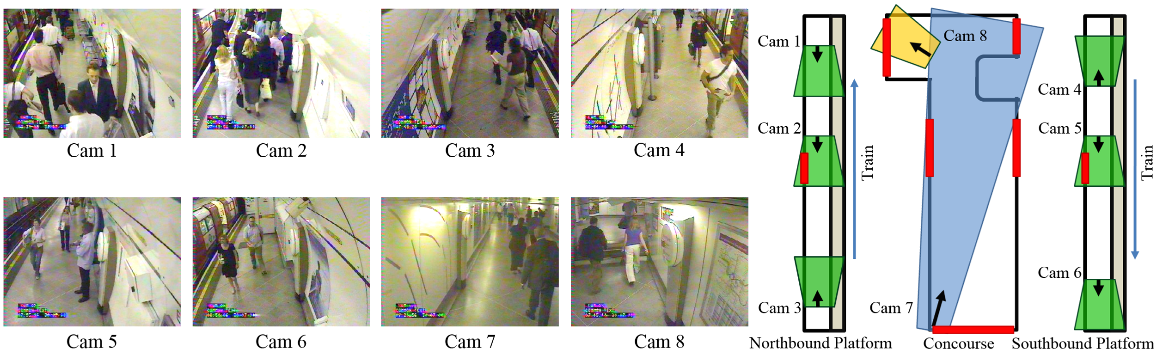
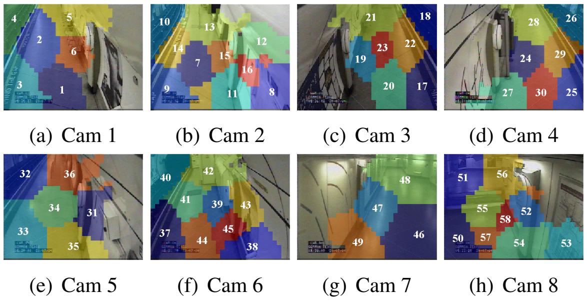


Figure 3. The layout of the underground station with the camera locations. Entry and exit points are highlighted in red colour.

hours. This dataset is thus challenging in terms of having enormous number of objects and low video frame rate. In addition, the complex activities in the scene make global activity analysis difficult. For example, there are trains to different destinations using the same platform. Passengers waiting on the platform thus may choose not to get on an arrived train. In addition, the eight cameras only cover a small section of this large station, which introduces additional uncertainty to the captured global activities.

4.2. Semantic Scene Decomposition


(a) Cam 1 (b) Cam 2 (c) Cam 3 (d) Cam 4

(e) Cam 5 (f) Cam 6 (g) Cam 7 (h) Cam 8

Figure 4. Semantic scene decomposition results.

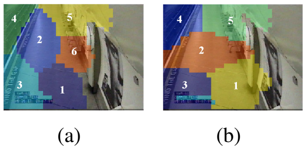
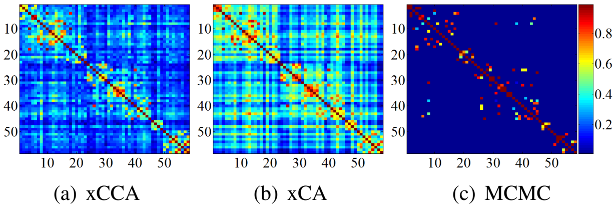

(a) (b)

Figure 5. Better scene decomposition result (a) was obtained using our correlation based distance metric, as compared to the result obtained using histogram-based technique [10] (b).

We used 5000 frames (or 2 hours) from each camera for activity correlation analysis. Figure 4 shows the results of scene decomposition. Each camera view was automatically segmented into semantic regions based on the local pixel block activities, despite the heavy occlusion and low temporal resolution. For example, the areas corresponding to the train tracks and platforms were well isolated. The sitting areas (regions 5 and 21) were also segmented from areas where people standing or walking. We performed comparison with the scene decomposition method introduced in [10] and found that our method yielded more accurate region boundary (Figure 5). As we explained in Sec. 3.1, this is because our local activity representation captures the temporal dynamics of activity while the Bag of Words based representation in [10] ignores the temporal order of the activity occurrences.

4.3. Multi-Camera Topology Inference

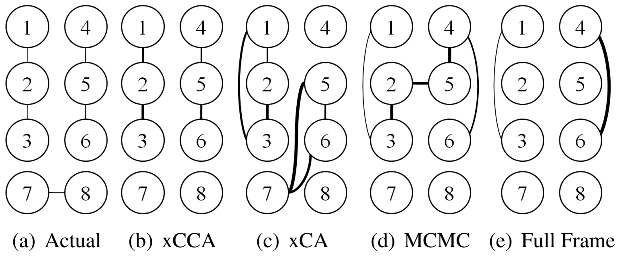


(a) xCCA (b) xCA (c) MCMC

Figure 6. Regional activity affinity matrices obtained using xCCA, xCA and MCMC Bayesian network structure learning.

In the experiments on topology inference, we first compared xCCA with xCA and MCMC Bayesian network structure learning for learning regional activity correlation. The regional activity affinity matrices (see Eqn. (10)) are shown in Figure 6. From the xCCA affinity matrix shown in Figure 6(a), it is evident that regions within the same camera view exhibited high correlations, as expected. More importantly, xCCA also correctly discovered correlation between regions across camera views. In comparison, xCA tended to 'over-correlate' regions causing the detection of correlations that do not exist. In contrast, MCMC Bayesian network structure learning revealed few and also incorrect correlations with a lot of missing detections.

The camera topologies yielded by different methods are shown in Figure 7. From the results, we observe that xCCA yielded the closest topology to the actual one. It is not surprising to see that our xCCA outperformed MCMC Bayesian network structure learning significantly. As we discussed earlier, the learned structure using Bayesian network structure learning can only reveal zero-order temporal dependencies, *i.e.* co-occurrence relationships, between activities. Thus it cannot cope with more complex correlations that are common in a multi-camera scene. xCCA



(a) Actual (b) xCCA (c) xCA (d) MCMC (e) Full Frame

Figure 7. xCCA yielded the closest topology to the actual one as compared to other methods. Figure (e) shows the topology inferred using xCCA without scene decomposition. Only edges with normalised correlation value larger than 0.8 are shown.

outperformed xCA 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.

To demonstrate the importance of semantic 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 result is shown in Figure 7(e) which suggests that without scene decomposition, even the proposed xCCA failed to learned the correct camera topology.

All methods failed to infer the connection between camera views 7 and 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 the 8 cameras (see Figure 3). This weakened the correlation between these two camera views.

4.4. Context-aware Person Re-identification

In this experiment, we compared the recognition performance of people across camera views using colour histogram (CH) alone, CH + xCA, and CH + xCCA. Note that MCMC Bayesian network structure learning is not able to quantify the temporal and causal relationship between two correlated regions. It is thus not suitable for context-aware person re-identification. Score returned by CH was computed by Eqn. (8), whilst score returned by CH + other methods were computed by Eqn. (9). In this evaluation, we performed re-identification on 16 individuals against 298 persons with their blobs manually segmented.

The results are shown in Figure 8 and example matches are given in Figure 9. It can be seen that the result yielded by CH + xCCA was significantly better than that obtained using CH alone and CH + xCA. In particular, CH + xCCA yielded the best performance with approximately 68.75% of the queries generated a true match in the top 20 rank, compared to 43.75% and 25% using CH + xCA and CH alone. Without considering the camera topology, each person has to be compared against all possible candidates. On the contrary, with the inferred topological information, the search space and ambiguity were greatly reduced which has resulted in better recognition rate. Our results also show that CH + xCCA outperforms CH + xCA. This is because, as demonstrated in Sec. 4.3, xCCA is able to learn the regional activity correlations more accurately than xCA.

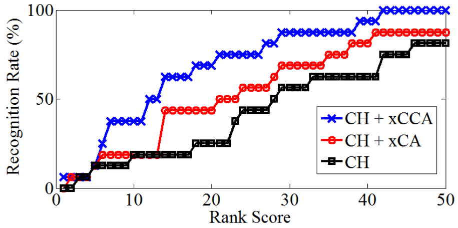


Figure 8. Performance comparison on person re-identification using colour histogram (CH) alone, CH + xCA, and CH + xCCA.

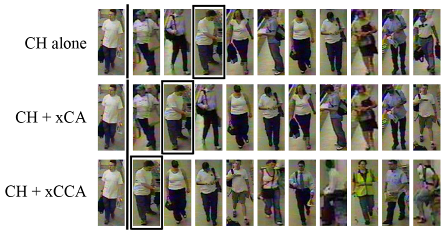


Figure 9. Example of matches using colour histogram (CH) alone, CH + xCA and CH + xCCA. The left column is the query image, and the remaining columns are the top matches ordered from left to right. Matches that correspond to the queries are highlighted.

4.5. Activity-based Temporal Segmentation

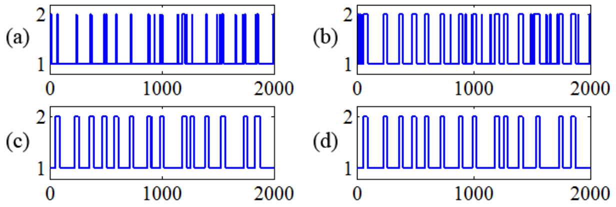
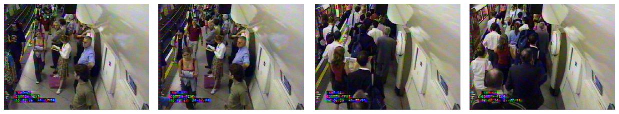


Figure 10. Example of phases inferred using (a) single view activity analysis without semantic scene decomposition, (b) single view activity analysis with semantic scene decomposition, and (c) global activity analysis. The ground truth is shown in (d). Y-axis represents the inferred phases and X-axis represents the frame index. Only 2000 frames from the test set are shown.

Global activities were discovered by performing spectral clustering on the regional activity affinity matrix (see Figure 6). Two global activities are learned, corresponding to the platform activities observed by Cam 1,2,3 and Cam 4,5,6 respectively. Due to space constraint, we only report the temporal segmentation result on the platform activity



(a) Phase 1 - train is absent (b) Phase 2 - train is present

Figure 11. Example frames from the phases inferred using global activity analysis.

monitored by Cam 1,2,3. The segmentation result was compared with those from individual single camera view without semantic scene decomposition and from single camera view with semantic scene decomposition.

We employed 5000 frames per camera to train a MOHMM following the steps described in Sec. 3.5. The test set consists of the rest of the videos (43000 frames per camera). For all three methods, it turned out that two hidden states gave the best BIC score. The two phases have clear semantic meaning: one phase corresponds to the period when train is absent, whilst the other phase is the period when train is present. We then compared the inferred phases obtained using the three methods with the ground truth. The accuracy yielded by single view analysis without scene decomposition was 68.17%. The accuracy increased to 86.71% after we employed scene decomposition on the single view analysis, whilst the proposed method based on global activity analysis gave $\mathbf{94.33}\%$. Examples of the inferred phases by different methods and some example frames from the segmented phases are shown in Figure 10 and Figure 11 respectively. The results 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 of activities during the occlusions and low frame rate. The ambiguities were greatly reduced by exploiting semantic scene decomposition. As compared to single view activity analysis, our global activity modelling utilises evidences collected from multiple correlated regions across camera view. It has therefore further reduced visual ambiguities, resulting in a more accurate segmentation result.

5. Conclusions

We presented a novel approach for multi-camera activity correlation analysis and global activity inference over a distributed camera network of non-overlapping views. We introduced a Cross Canonical Correlation Analysis framework to detect and quantify temporal and causal relationships between local semantic regions within and across camera views. The approach addressed three fundamental problems in multi-camera activity analysis: (1) estimate the spatial and temporal topology of the camera network; (2) facilitate more robust and accurate person re-identification through context-awareness; (3) perform global activity modelling and video temporal segmentation.

References

[1] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by probability distributions. *Bull. Calcutta Math. Soc.*, 35(99-109), 1943.

[2] N. Gheissari, T. B. Sebastian, J. Rittscher, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, pages 1528–1535, 2006.

[3] D. Gray and H. Tao. Viewpoint invariant pedestrain recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008.

[4] H. Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[5] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *TPAMI*, 28(4):663–671, 2006.

[6] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *ICCV*, pages 952–957, 2003.

[7] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *CVPR*, pages 26–33, 2005.

[8] M. Kendall and J. K. Ord. *Time Series*. Edward Arnold, 1990.

[9] L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: establishing a common coordinate frame. *TPAMI*, 22(8):758–768, 2000.

[10] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *ECCV*, pages 383–395, 2008.

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

[12] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *CVPR*, pages 205–210, 2004.

[13] R. E. Neapolitan. *Learning Bayesian Network*. Prentice Hall, 2003.

[14] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, 2008.

[15] D. Russell and S. Gong. Minimum cuts of a time-varying background. In *BMVC*, pages 809–818, 2006.

[16] K. Tieu, G. Dalley, and W. E. L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *ICCV*, pages 1842–1849, 2005.

[17] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *ECCV*, pages 110–123, 2006.

[18] X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *TPAMI*, Preprint, 2009.

[19] T. Xiang and S. Gong. Activity based surveillance video content modelling. *Pattern Recognition*, 41(7):2309–2326, 2008.

[20] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.

[21] E. E. Zelniker, S. Gong, and T. Xiang. Global abnormal behaviour detection using a network of CCTV cameras. In *IEEE Intl. Workshop on Visual Surveillance*, 2008.