Graph-based Correlated Topic Model for Motion Patterns Analysis in Crowded Scenes from Tracklets

Manal Al Ghamdi
maalghamdi@uqu.edu.sa

Umm Al-Qura University
Department of Computer Science
Saudi Arabia

Yoshihiko Gotoh
y.gotoh@sheffield.ac.uk

University of Sheffield
Department of Computer Science
United Kingdom

Abstract

This paper presents a graph-based correlated topic model (GCTM) to model the different motion patterns at highly cluttered and crowded environment. Unlike the existing methods that address trajectory clustering and crowd topic modelling using local motion features such as optical flow, it builds on portions of trajectory known as tracklets extracted from crowded scenes. It extends the correlated topic model (CTM) in the text processing field, by integrating the spatio-temporal graph (STG) as a prior to capture the spatial and temporal coherence between tracklets during the learning process. Two types of correlation are defined over the tracklets. Firstly intra-correlation of the extracted tracklets is encoded by the locality-constrained linear coding (LLC) with the geodesic distance and the shortest path graph. Secondly inter-correlation is derived between the tracklets by constructing a shortest path graph with k-nearest neighbourhood (kNN) from both the spatial and temporal domains. Experiments and comparisons show that the GCTM outperforms state-of-the-art methods both on qualitative results of learning motion patterns and on quantitative results of clustering tracklets.

1 Introduction

Trajectory clustering and analysis of crowd movements have been vital components of various applications in public surveillance, such as flow estimation. The goal is to analyze individual movements by a trajectory associated with a cluster label, thus representing individuals' paths. A highly crowded scene is particularly challenging because of the density, heavy occlusions and variations in the view. Additionally interaction between individuals can lead to misdetection of body parts [15]. The presence of such challenges makes it difficult to analyze movements using conventional techniques such as background subtraction and motion segmentation, although they may work effectively with less-crowded scenes.

To overcome the shortcomings of conventional techniques, motion patterns have been introduced for processing crowded scenes. Examples of motion pattern techniques include scene structure-based force models [7] and the Bayesian framework with spatio-temporal

motion models [10]. These models are based on the assumption that the objects move coherently in one direction throughout a video. This is a major shortcoming, as it fails to represent the complex crowded scenes with multiple dominant crowd behaviours in each location. Thus, trajectory clustering have been presented for various applications including crowd analysis and video surveillance. In many applications, a vast amount of trajectories and motion patterns are extracted and clustered into groups without manually labeled of the data. Lin *et al*. [11] detected motion trajectories in crowd scenes by processing the flow fields followed by a two-step clustering process to define semantic regions. Lu *et al*. [13] extracted the motion trajectories to investigate the characteristics of pedestrians in unstructured scenes. Trajectories clustered using fuzzy c-means (FCM) algorithm to form the motion patterns. Zhao *et al*. [20] detected crowd groups and learned the semantic regions using a hierarchical clustering framework with three priors based on the Gestalt laws.

Many works have been proposed for trajectory clustering based on mid-level features learning. These features are usually observed as paths defined by individuals' movements, which aim to map the segments of trajectories from low-level feature space to their clusters [21]. Trajectory mid-level features can be learnt with hierarchical latent variable Bayesian models, such as latent Dirichlet allocation (LDA) [9] and the correlated topic models (CTM) [3]. These models are known as 'topic models', adopted from the text processing field. Using these models, documents are represented by trajectories and visual words are given by observations of object trajectories. The CTM was adopted by Rodriguez *et al*. [15] as a mid-level feature to represent multiple motion behaviours in one scene. Zhou *et al*. [21] proposed a random field topic (RFT) model to perform trajectory clustering in a crowd scene. Zou *et al*. [23] extended the CTM to a scene prior belief based correlated topic model (BCTM). Therefore, it could only be used with situations where scene priors were available.

This paper presents a graph-based correlated topic model (GCTM) for learning crowd behaviour from tracklets and to cluster tracklets. A tracklet is a fragment of a trajectory and is obtained by a tracker in a short period [21]. Hence it is possible to estimate more stable tracklets than longer trajectories. We use a Kanade-Lucas-Tomasi (KLT) tracker [16] to extract tracklets from highly crowded scenes. GCTM advances the CTM by integrating the spatio-temporal graph (STG) as prior to enforce the spatial and temporal coherence between tracklets during the learning process. We make the following contributions: 1) learn a representations of multi-modal crowd behavior using the spatio-temporal correlations; 2) extend the CTM with graph-based representation to solve an existing problem in a crowd analysis framework. 3) present a motion pattern clustering framework without any priors information about the scene.

Different from the exiting trajectory clustering methods which assumed that trajectories were independent given their cluster labels, GCTM defines two types of spatio-temporal correlations over tracklets. Firstly, in order to encode visual vocabulary from a video sequence, the locality-constrained linear coding (LLC) technique [18] is applied between the tracklets. It captures the intra-correlation by translating the extracted tracklets into their local codes with fewer codebook basis using the geodesic distance and the shortest path graph. Secondly the inter-correlation is derived over the tracklets by constructing a spatio-temporal shortest path graph with k-nearest neighbourhood (kNN) to model their spatial and temporal connections. Finally GCTM learns the topics from visual vocabulary and the STG neighbours to create clusters for the tracklets. Experiments on two different video datasets – one collected at the crowded Grand Central station in New York [21] and the other collected from two different locations at Al-Masjid Al-Haram [19], both of which are well known for crowded and busy scenes (Figure 1) – show the effectiveness of the presented approach.

Figure 1: Sample frames from indoor scenes at (1) (2) Al-Masjid Al-Haram [19] and (3) New York's Grand Central Station [21].

2 Related Work

When analysing crowded scenes it is important to consider the characteristics of the extracted features and their correlation. There exist two directions for crowd analysis [9, 22]. Firstly microscopic approaches aim to identify individuals, their body parts or objects. On the other hand, with macroscopic approaches, the crowd is analysed as a whole with no explicit information about individuals. Macroscopic methods deal with problems that require information about the scene to be holistically considered, *e.g.*, estimation of traffic flows [5]. Such events can be detected by analysing the variations with the motion models, and moves of individuals can be defined as outliers with respect to the entire crowd. Zhou *et al.* [22] defined abnormal behaviours by statistically analysing extracted trajectories from a crowd. They applied the KLT feature tracker [16] to define the motion pattern, which was then clustered to form representative trajectories. The multi-observation hidden Markov model (MOHMM) was applied to determine the abnormality of the motion. Using the same concept of global and holistic representation, Khokher *et al.* [9] represented a dynamic motion by applying multiple spatial scales to extract dense features from a crowded scene. They employed a median filter as a tracker, a histogram of oriented gradients (HOG) [12] as a descriptor and the support vector machine (SVM) as a classifier. Recently, Huang *et al.* [8] presented an unsupervised deep learning framework to detect anomaly events in crowded scenes. Multiple features were extracted and used to train three convolutional restricted Boltzmann machines. A multimodal fusion scheme was then used to learn the deep representation of crowd patterns and a one-class support vector machine model was utilized to detect anomaly events.

Trajectory clustering approaches based on mid-level feature learning have attracted attention. Zhou *et al.* [21] proposed a Random Field Topic (RFT) model to perform trajectory clustering in a crowd scene. It extended the LDA models by integrating scene priors and using a Markov random field (MRF) algorithm. They significantly improve the clustering performance over LDA models; however, it can drop in crowded scenes with correlated topics where topics are shared with multiple clusters, and clusters are also shared with multiple topics. To address this problem, Rodriguez *et al.* [15] extended the CTM to define a weighted tracker that predict a rough displacement using a codebook generated from all moving pixels in the scene along with the learned high-level behaviour. Although CTM is an effective model, it ignores the distribution of data, thus its performance depends on the low-level features used as a combination with the mid-level features. Zou *et al.* [23] extended the CTM to a scene prior belief based correlated topic model (BCTM). They performed a parameter estimation using a scene prior based on a joint Gaussian distribution to uncover the relations between trajectory clusters and the mid-level features.

Despite the effectiveness of the above models, however, most of them ignore the temporal

relationship within the crowded scenes and also the distribution of data. Therefore, they require a complex parameter estimation and variable inference procedure. In contrast, the presented approach considers the spatio-temporal information represented by the motion pattern of a crowd and the structure of the scene. The goal is to address the problem of trajectory clustering and motion pattern analysis in high-density crowds without using any prior knowledge of the scene such as exists and entrances. To the best of our knowledge this has not been previously attempted in crowd analysis for complex scenes such as at Al-Masjid Al-Haram.

3 Our Approach

This section outlines how the mid-level features (topics) are learnt as motion patterns (paths) by GCTM parameters estimation. To make the paper self-contained, we start by reviewing the conventional CTM (Section 3.1) followed by the proposed GCTM (Section 3.2).

3.1 Correlated Topic Model

Figure 2(a) shows the graphical representation of the CTM that was originally developed in the text-processing field [3]. Let I, N and K denote the number of documents, the number of words in a document and the number of hidden variables (or 'topics') in the model, respectively. The circles in the figure are random variables or model parameters, and the edges specify the probabilistic dependencies (or the conditional independences) among them; boxes, with I, N and K, are compact notations for multiple instances of the variables or parameters. Shaded variables represent the observed variables, while unshaded variables indicate the latent variables. The CTM assumes that each document is a mixture of words based on a set of hidden topics, and in turn each topic is determined by a distribution over the entire vocabulary. In the figure, π is a K-dimensional vector, specifying the topic priors for each document; z is a hidden variable, following a parameterized multinomial distribution $Mult$; x is the random variable whose value is the observed word (*i.e.*, 'feature'); and β is a hyper-parameter, corresponding to the mid-level features. Finally μ and Σ are the mean and the covariance matrix of the multivariate Gaussian process. The generative process of the CTM is outlined as follows:

- Draw $\pi|\{\mu,\Sigma\} \sim N(\mu,\Sigma)$
- Draw the document-specific topic proportions θ as $\theta = \frac{exp(\pi)}{\sum_i^K \pi_i}$
- For each visual word $x_n, n \in \{1,\ldots,N\}$:
  1. Choose a topic assignment $z_n|\pi$ from $Mult(\theta)$;
  2. Choose a word $x_n|\{z_n,\beta_{1:K}\}$ according to $p(x_n|z_n,\beta)$.

  According to this model, the document probability given topic variable θ, word x and individual topic assignment z is:

$$p(\pi,z,x|\mu,\Sigma,\beta) = p(\pi|\mu,\Sigma)\prod_{n=1}^{N} p(z_n|\pi)p(x_n|z_n,\beta) \tag{1}$$

  Notice that the topic-level information given by π and z is hidden, while the word-level representation is observed.

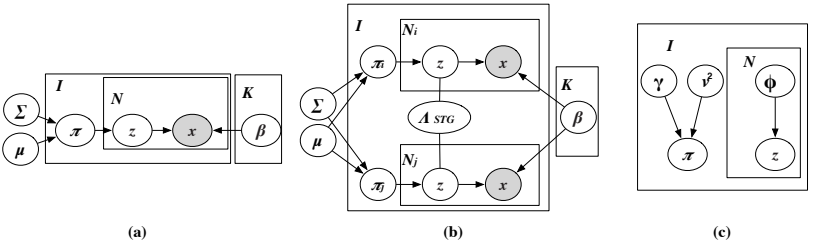


Figure 2: CTM and GCTM models. (a) Graphical representation of CTM [3]. (b) Graphical representation of GCTM. (c) Graphical representation of approximate distribution of GCTM.

An approximate method (variational approximation) has been used to estimate the likelihood of performing training and to estimate the most likely topic proportions π and topic assignments z. Further details can be found in [3].

3.2 Graph-based Correlated Topic Model

The graphical representation of GCTM is presented in Figure 2 (b). Corpus, document, topic and words (for text data) in CTM are replaced with path, tracklet, motion pattern and observation visual words that are quantized from the observation points (for video data) in GCTM. The topic mixture of a document corresponds to a set of different motion patterns in a tracklet. GCTM learns crowd movements by clustering tracklets. Observed visual codes and the spatio-temporal graph are the inputs for GCTM. Section 4.1 describes the generation of the visual codes, while Section 4.2 describes the construction spatio-temporal graph.

Suppose that a corpus has I documents, each of which is modelled as a mixture of K topics. Each document i is encoded with N visual codes (motion words) $X = \{x_1, x_2, \ldots, x_n\}$. Each topic k is a distribution over a word vocabulary given by the hyper-parameter $\beta = \{\beta_k\}$ to be optimised. It is assumed that π_i is a continuous variable sampled from a Gaussian distribution $p(\pi_i|\mu,\Sigma)$ with the mean μ and the covariance Σ. For each motion word n in document i, a topic z_{in} is drawn with a probability π_{nk}. z_{in} is a hidden variable assigned to a spatio-temporal word x_{in} drawn from a multinomial distribution $\beta_{z_{in}}$. The joint distribution is [3]:

$$P(\pi_i, z_{in}|x_{in}, \beta_{1:K}, \mu, \Sigma) = \frac{p(\pi_i|\mu,\Sigma)\prod_{n=1}^{N} p(z_{in}|\pi_i)p(x_{in}|z_{in},\beta_{1:K})}{\int p(\pi_i|\mu,\Sigma)\prod_{n=1}^{N}\sum_{z_{n=1}}^{K} p(z_{in}|\pi_i)p(x_{in}|z_{in},\beta_{1:K})d\pi} \quad (2)$$

i, n and k are indices of documents, words and topics respectively. $p(z_{in}|\pi_i)$ is specified by the STG as:

$$p(Z|\pi) \propto exp(\sum_{i} \log \pi_i + \sum_{j\in\varepsilon(i)} \sum_{n_1 n_2} \Lambda(z_{in_1}, z_{jn_2}) \quad (3)$$

where $Z = \{z_i n\}$ and $\pi = \{\pi_i\}$. Λ weights the correlation between tracklets based on the spatio-temporal graph, and $\varepsilon(i)$ is the set of spatio-temporal neighbourhood tracklets for tracklet i and both are defined by the STG in Section 4.2. The intuition behind our model is interpreted as follows. According to the techniques of topic modelling, words often appear in the same documents will be under one topic. Therefore, if two regions share many tracklets,

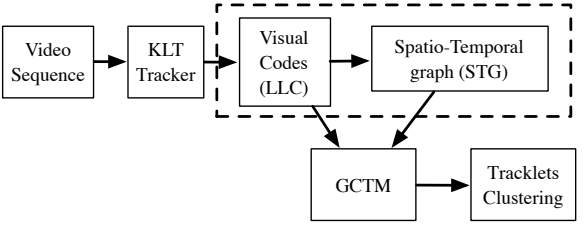


Figure 3: Flowchart of the crowd behaviours modeling framework with GCTM.

they tend to be interpreted by the same behaviour (topic) . The STG term Λ encourages tracklets which are spatially and temporally close to have similar distributions over topic.

Given a set of tracklets with visual codebook, the objective is to find the maximum likelihood estimation for model parameters $\{\pi, \beta, \mu, \Sigma\}$. In order to estimate parameters for GCTM, we used parts of video sequences as training data and adopt the variational expectation maximization (EM) algorithm to do variable inference and parameter estimation [3]. Figure 2(c) is the graphical representation of the approximate distribution of the GCTM where $\gamma_{I\times K}$, $v_{I\times K}$ and Φ are variational parameters. Therefore, the log-likelihood for a document i is given by:

$$log\, p(x|\mu,\Sigma,\beta) = L(\gamma,v,\phi;\mu,\Sigma,\beta) + KL(q(\pi,z|\gamma,v,\phi)||p(\pi,z|x,\mu,\Sigma,\beta)) \quad (4)$$

We iteratively maximize the term $L(.)$ instead of $p(x|\mu,\Sigma,\beta)$, which results in the minimum of difference between the distribution in Figure 2(b) and Figure 2(c). For details of computation, please refer to [3].

4 Tracklets Clustering

For tracklets clustering, the first step is to extract the tracklets and represent them with a collection of visual codes (Section 4.1). The second step is to apply a spatio-temporal graph on the tracklets to uncover spatio-temporal relations among them and to be used later for the learning process (Section 4.2). Given the STG and the set of visual codes, the final step is to learn the mid-level features by GCTM (Section 3.2) and produce the final clustering. The framework is shown by a flow chart in Figure 3.

4.1 Visual Vocabulary

In order to apply the GCTM model, we firstly need to represent the video sequence by a set of spatio-temporal visual words. We use a KLT tracker [16] to generate the tracklets and the motion vectors of objects. All the features are then quantized into spatio-temporal codes according to a visual codebook B using LLC algorithm [18]. The LLC is a coding scheme proposed by Wang *et al.* [18] and extended to space-time domain in [10] to project individual descriptors onto their respective local coordinate systems. For each tracklet x_{in}, the algorithm works by firstly constructing a spatio-temporal graph between the tracklet features and a codebook B, computing the shortest path, performing a kNN search, and finally solving a constrained least-square fitting problem.

4.2 Spatio-temporal Graph

For the STG, $\Lambda()$ is defined as spatio-temporal neighbourhood graph [6] as follow. A distance matrix $D = \{d_{ij}\}$ between two tracklet x_i and x_j is calculated by: $d_{ij} = \left(\sum_{d=1}^{D} \|x_i - x_j\|^2\right)^{\frac{1}{2}}$ where $\|\cdot\|^2$ is the ℓ^2 norm. Then, for each instance x_i $(i = 1,\ldots,I, j)$:

1. L tracklets, whose distance is the closest to x_i, are connected. They are referred to as spatial neighbours (sn): $sn_{x_i} = \left\{x_{j1},\ldots,x_{jL} \mid \text{argmin}_j{}^{L}(d_{ij})\right\}$ where $\underset{j}{\text{argmin}}^{L}$ implies L node indexes j with the shortest distances;

2. Another L tracklets, chronologically ordered around x_i, are set as temporal neighbours (tn): $tn_{x_i} = \left\{x_{j-\frac{L}{2}},\ldots,x_{j-1},x_{j+1},\ldots,x_{j+\frac{L}{2}}\right\}$

3. Optimally, (tn_{sn}) is selected from temporal neighbours of spatial neighbours as: $tn_{sn_{x_i}} = \left\{tn_{x_{j1}} \cup tn_{x_{j2}} \cup \ldots \cup tn_{x_{jL}}\right\} \cap tn_{x_i}$

4. Spatial and temporal neighbours are then integrated, producing spatio-temporal neighbours (ε) for tracklet x_i as: $\varepsilon_i = sn_{x_i} \cup tn_{sn_{x_i}}$

   The above formulation of ε_i effectively selects x_i's temporal neighbours that are similar, with a good chance, to its spatial neighbours.

Given the spatio-temporal neighbourhood, Dijkstra's shortest path algorithm is applied to the nodes [17]. This forms a new correlation matrix $G = \{X, Y, E\}$ of pairwise geodesic distances with $V = X \cup Y$ as the vertex set and $E = \{\omega_{ij}\}$ as the edge set, where ω_{ij} presents the shortest path distance between two tracklets x_i and x_j. If the tracklet j is a spatio-temporal neighbour of i and $j \in \varepsilon_i$, then $\Lambda(z_{in_1}, z_{jn_2}) = \omega_{ij}$, otherwise, $\Lambda(z_{in_1}, z_{jn_2}) = 0$.

4.3 Tracklet Prediction

After the mid-level features are learnt and the topic probabilities of the training tracklets are computed, each tracklet has a set of K topics to choose from. A topic label with the highest probability is assigned to the tracklet. Given a new tracklet m with an unknown path, the algorithm first correlates the given tracklet with the tracklets from the training set by generating the spatio-temporal neighbours ε. The spatio-temporal neighbours with the minimum entropy on z is then chosen for the given tracklet to infer its topic label.

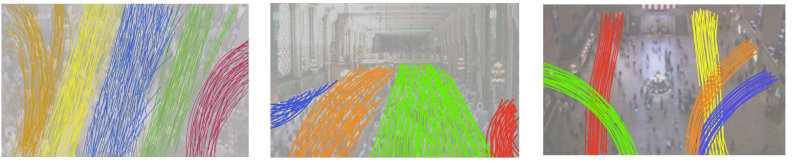
5 Experiments

We evaluated the GCTM using motion patterns (or paths) clustering task in crowded videos. Once the GCTM model is learnt, tracklets are clustered based on the motion pattern they belong to. For each tracklet, the decision of the topic is made to the cluster that gives the highest likelihood probability.

Experiments were conducted on two different datasets. The first one is the New York's Grand Central Station dataset [21], collected from the inside of the Grand Central railway station in New York, USA. It contains multiple entrances and exits where individuals have different paths to follow. Therefore, the crowd presents multiple behaviours (or paths) in

Dataset	Resolution	Duration	Codebook size	Tracklets
Al-Masjid (S1)[19]	960×540	5,600 sec	$96 \times 54 \times 4$	87,321
Al-Masjid (S2)[19]	960×540	3,400 sec	$96 \times 54 \times 4$	61,760
Station [21]	720×480	1,800 sec	$72 \times 48 \times 4$	47,866

Table 1: Resolution, duration, codebook size and No. of extracted tracklets for each dataset.



(a) (b) (c)

Figure 4: Tracklet clustering based on motion patterns learnt GCTM for (a) Al-Masjid (S1), (b) (S2) and (c) Station (Seen better in colour.)

various moving directions. The second one is Al-Masjid Al-Haram dataset[19], collected from indoor scenes at the holy mosque of Mecca, Saudi Arabia. This dataset involved a number of difficult problems, such as lighting changes, occlusions, a variety of objects, changes of views and environmental effects. Al-Masjid videos were collected from two scenes. The first one was at one of the Tawaf area stairs used to enter or leave the Tawaf. The second scene was recorded at the second and the third floors of SAFA and MARWA area, which is a long walkway with two different directions. For simplicity, we denote the first dataset as 'Station' and the second one as 'Al-Masjid (S1)' and 'Al-Masjid (S2)'. The details of both datasets are presented in Table 1.

For the low-level feature step, the initial codebook B used for the LLC codes was learnt from a random half of the tracklets. In both datasets, the size of the codebook was designed as follows: the $W \times H$ scene was divided into 10×10 cells and the velocities of key-points were quantized into four directions. For the STG, the similarity matrix was computed using the Euclidean distance and the KNN graph was constructed with $L = 25$.

5.1 Results

Tracklet clusters, generated by the GCTM, are identified by different colours and presented in Figure 4. Tracklets were assigned to the cluster whose corresponding topic's probability is the highest. In both datasets most tracklets were broken; however, spatially distant tracklets could be clustered in one group when they were found to have the same path. For example, the leftmost cluster from Al-Masjid (S1) shown in Figure 4(a) contained tracklets for pedestrians walking towards the left side of the scene. It was not easy to obtain this cluster because occlusion caused by the people sitting on the marble pillar resulted in tracklets observed mostly either at the start or the end of the path. In Figure 4(b), movements were clustered into four groups; one of them was up the left side with an exit and another one was down the right side with another exit. Tracklets were mixed with adjacent paths and occluded by the heavy traffic; however, GCTM was able to identify these paths and their exit positions. Similarly, in Figure 4(c), tracklets were clustered into five different paths; two of them were on the right side to exit the station. Tracklets were shared between these two exits, but the GCTM was able to distinguish between their paths.

For comparison, Figure 5 presents tracklet clusters from Al-Masjid (S1) by various ap-

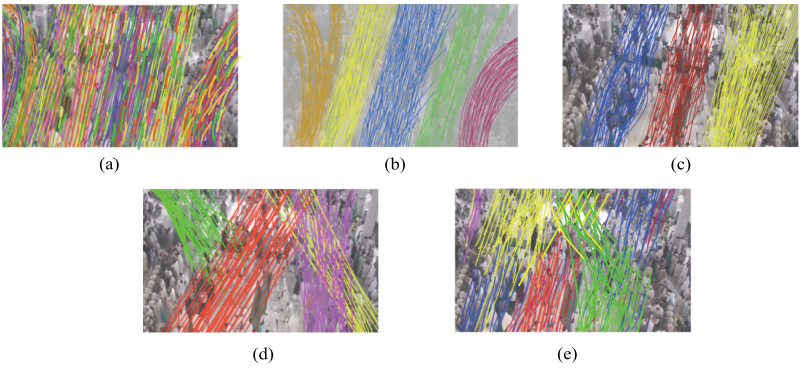


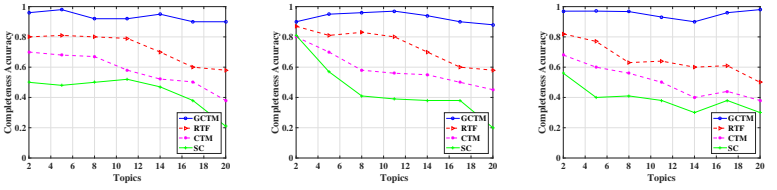
Figure 5: Comparison of tracklet clustering approaches: (a) original tracklet set, (b) GCTM, (c) RFT, (d) CTM and (e) SC.

proaches, including GCTM, random field topic[1] (RFT) [21], CTM [15] and spectral clustering (SC) [7]. We implemented the SC using a linear interpolation and the Euclidean distance to measure the similarities. Different colours in the figure represent different clusters (paths). It can be observed that GCTM was able to produce the cleanest tracklet paths and clusters. The other three approaches failed to perform tracklet clustering, which was particularly evident with the side paths towards the exits because of their heavy occlusion. RFT achieved better results for the central paths in comparison to CTM and SC. SC was the worst. It was only able to cluster the tracklets at one end of the movements (the starting or ending positions) as one path and the other end as a different path.

For further quantitative evaluation of the clustering performance, we adopted correctness and completeness introduced by [14]. Correctness is the accuracy with which a pair of tracklets from different pathways (with the groundtruth) are clustered into different groups. Completeness is the accuracy with which a pair of tracklets from the same path are clustered into the same group. In an extreme case, a 100% completeness and 0% correctness may be achieved when all the tracklets are clustered into a single group. Another extreme is 0% completeness and 100% correctness, achieved when each tracklet is clustered into a different group. A good clustering algorithm should achieve high percentages in both correctness and completeness. As a groundtruth we manually labelled 2,500 tracklets for correctness and 1,700 for completeness with Al-Masjid (S1), 2,000 tracklets for correctness and 1,500 for completeness with Al-Masjid (S2) and 2,000 tracklets for correctness and 1,500 for completeness with Station.

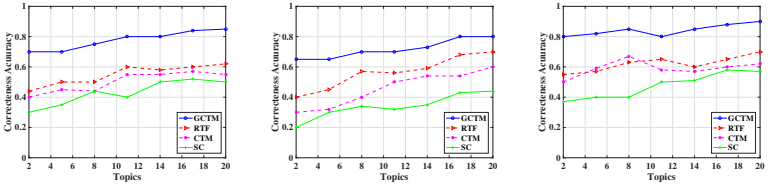
Correctness and completeness for GCTM, RFT, CTM and SC are reported in Figures 6 and 7. The correctness and completeness results show that GCTM outperformed the other three approaches in both datasets with a clear margin. The margin was even wider for completeness when the number of topics was larger. The GCTM with the STG is able to learn discriminative mid-level features better, even with a large number of topics to share the clusters. The other three approaches did not cluster tracklets well because most of these tracklets were short and mixed and difficult to be clustered. RFT has advanced the LDA [4] by considering belief priors based on the position and the spatial correlation of tracklets along

[1] We used the publicly available code from the authors' websites.



(a) Al-Masjid (S1) (b) Al-Masjid (S2) (c) Station

Figure 6: Completeness accuracies of tracklet clustering approaches.



(a) Al-Masjid (S1) (b) Al-Masjid (S2) (c) Station

Figure 7: Correctness accuracies of tracklet clustering approaches.

the video sequence. However, the spatio-temporal correlation between tracklets was disregarded. CTM considered four motion directions at each spatial location, but it ignored the temporal relation between sequential local motions in crowded scenes. SC was adversely affected by the outliers because it relied on the linear distance for clustering and did not consider ordering of points or the direction of moves. All three methods process low-level features of the tracklets in the high-dimensional feature space, which is very sparse, making it difficult to directly perform clustering.

Figure 7 shows that GCTM had better correctness accuracies compared to the others. RFT, with its priors information achieved the second best performance apart from the Station videos, where CTM with five and eight topics in the Station dataset outperforms RFT. This is because the CTM approach could perform well where scenes were not too crowded (*e.g.*, Station, as opposed to Al-Masjid), and thus full and complete tracklets could be generated with its object-tracking algorithm. They were clustered well by the CTM; however, the accuracy dropped as the number of topics increased. Finally, including the preprocessing time of feature detection, codebook generation and the topic learning, the GCTM model takes a few minutes (less than 10) on a 2.6 Ghz machine, which is faster than RFT and CTM.

6 Conclusions

We have proposed a graph-based correlated topic model (GCTM) for learning the motion patterns in crowded scenes from tracklets. By constructing a scene prior based spatio-temporal correlations over the extracted tracklets, GCTM could effectively reflect the relations between tracklets, and learn discriminative crowed features. The learned topics capture the global structures of the scenes in long range with clear behaviour interpretation. It is also able to separate different paths at fine scales with good accuracy. Experiments and comparisons with recent methods have shown that GCTM is faster and more able to learn a crowd topic model and to cluster tracklets.

References

[1] M. Al Ghamdi, N. Al Harbi, and Y. Gotoh. Spatio-temporal video representation with locality-constrained linear coding. In *The 12th European Conference on Computer Vision*, pages 101–110, 2012.

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*, pages 1–14, 2008.

[3] D. Blei and J. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022, 2003.

[5] A. Chan, M. Morrow, and N. Vasconcelos. Analysis of crowded scenes using holistic properties. *IEEE Transactions on Signal Processing*, pages 174–188, 2009.

[6] M. A. Ghamdi and Y. Gotoh. Alignment of nearly-repetitive contents in a video stream with manifold embedding. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1255–1259, 2014.

[7] Weiming Hu, Dan Xie, Zhouyu Fu, Wenrong Zeng, Steve Maybank, and Senior Member. Semantic-based surveillance video retrieval. *IEEE Transactions on Image Processing*, page 1168âĂŞ1181, 2007.

[8] Shaonian Huang, Dongjun Huang, and Xinmin Zhou. Learning multimodal deep representations for crowd anomaly event detection. *Mathematical Problems in Engineering*, 2018:13, 2018.

[9] M. Khokher, A. Bouzerdoum, and S. Phung. Crowd behavior recognition using dense trajectories. In *International Conference on Digital Image Computing: Techniques and Applications*, pages 1–7, 2014.

[10] L. Kratz and K. Nishino. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *IEEE Transactions on PAMI*, 34(5):987–1002, 2012.

[11] W. Lin, Y. Mi, W. Wang, J. Wu, J. Wang, and T. Mei. A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. *IEEE Transactions on Image Processing*, 25(4):1674–1687, 2016.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[13] Wei Lu, Xiang Wei, Weiwei Xing, and Weibin Liu. Trajectory-based motion pattern analysis of crowds. *Neurocomputing*, 247:213 – 223, 2017. ISSN 0925-2312.

[14] B. Moberts, A. Vilanova, and J. J. van Wijk. Evaluation of fiber clustering methods for diffusion tensor imaging. In *VIS 05. IEEE Visualization, 2005.*, pages 65–72, 2005.

[15] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *ICCV*, pages 1389–1396, 2009.

[16] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University, 1991.

[17] L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality reduction: a comparative review, 2008.

[18] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[19] Basim Zafar Yasir Ali and Mohammed Simsim. Estimation of density levels in the holy mosque from a network of cameras. In *Traffic and Granular FlowâĂŹ15*, 2015.

[20] Weiqi Zhao, Zhang Zhang, and Kaiqi Huang. Gestalt laws based tracklets analysis for human crowd understanding. *Pattern Recognition*, 75:112 – 127, 2018. ISSN 0031-3203.

[21] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *CVPR*, pages 3441–3448, 2011.

[22] S. Zhou, W. Shen, D. Zeng, and Z. Zhang. Unusual event detection in crowded scenes by trajectory analysis. In *ICASSP*, pages 1300–1304, 2015.

[23] J. Zou, Q. Ye, Y. Cui, D. Doermann, and J. Jiao. A belief based correlated topic model for trajectory clustering in crowded video scenes. In *ICPR*, pages 2543–2548, 2014.