Computer Vision and Image Understanding 191 (2020) 102710

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Identifying motion pathways in highly crowded scenes: A non-parametric tracklet clustering approach

Allam S. Hassanein [a,*], Mohamed E. Hussein [b,d,1], Walid Gomaa [a,d], Yasushi Makihara [c], Yasushi Yagi [c]

[a] Egypt-Japan University of Science and Technology, Alexandria, 21934, Egypt
[b] Information Sciences Institute, 3811 Fairfax Dr # 200, Arlington, VA 22203, USA
[c] Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan
[d] Faculty of Engineering, Alexandria University, Alexandria, Egypt

ARTICLE INFO

Communicated by Amit Roy-Chowdhury

Keywords:
Manuscript
Tracklet similarity
DD-CRP
Semantic prior
Tracklet cluster likelihood
Anomaly detection

ABSTRACT

Many approaches that address the analysis of crowded scenes rely on using short trajectory fragments, also known as *tracklets*, of moving objects to identify motion pathways. Typically, such approaches aim at defining meaningful relationships among tracklets. However, defining these relationships and incorporating them in a crowded scene analysis framework is a challenge. In this article, we introduce a robust approach to identifying motion pathways based on tracklet clustering. We formulate a novel measure, inspired by line geometry, to capture the pairwise similarities between tracklets. For tracklet clustering, the recent distance dependent Chinese restaurant process (DD-CRP) model is adapted to use the estimated pairwise tracklet similarities. The motion pathways are identified based on two hierarchical levels of DD-CRP clustering such that the output clusters correspond to the pathways of moving objects in the crowded scene. Moreover, we extend our DD-CRP clustering adaptation to incorporate the source and sink gate probabilities for each tracklet as a high-level semantic prior for improving clustering performance. For qualitative evaluation, we propose a robust pathway matching metric, based on the chi-square distance, that accounts for both spatial coverage and motion orientation in the matched pathways. Our experimental evaluation on multiple crowded scene datasets, principally, the challenging Grand Central Station dataset, demonstrates the state-of-the-art performance of our approach. Finally, we demonstrate the task of motion abnormality detection, both at the tracklet and frame levels, against the normal motion patterns encountered in the motion pathways identified by our method, with competent quantitative performance on multiple datasets.

1. Introduction

Crowd activities and behaviors have significantly expanded in daily life because of the massive increase in population and technological advances. This growth has forced governments and authorities to be more alert to these types of scenes, particularly for crowd management and security assurance. As a result, the automatic analysis of crowded scenes has ignited interest in multiple research disciplines.

Vision-based crowded scene analysis is plagued by a number of daunting challenges. Specifically, object detection and tracking are much more difficult in highly complex crowded scenes with the presence of significant occlusion and (possibly dynamic) background clutter. Because of these challenges and the importance of the task, crowded scene analysis has become one of the popular topics in the computer vision and pattern recognition research community.

Motion representation is a core component of any vision-based approach for crowded scene analysis. In this regard, three levels of motion representation can be identified (Li et al., 2015): flow-based, spatio-temporal, and trajectory/tracklet-based.

The first level is flow-based representation, in which motion features are extracted at the pixel level (Saleemi et al., 2010; Mehran et al., 2010; Wang et al., 2014). Such low-level representations have the advantage of making the fewest assumptions about the scene and motion dynamics. However, they do not capture long-range temporal dependencies. They also suffer from ignoring spatial changes in scene dynamics. Moreover, the approaches that depend on dense flow representation can be remarkably time-consuming.

The second level is local spatio-temporal representation in which the crowded scene is expressed with local information extracted from two-dimensional (2D) patches or three-dimensional volumes

* Corresponding author.
E-mail address: allam.shehata@ejust.edu.eg (A.S. Hassanein).
[1] Mohamed E. Hussein mainly worked on this project while at the Egypt-Japan University of Science and Technology.

https://doi.org/10.1016/j.cviu.2018.08.004
Received 16 September 2017; Received in revised form 30 July 2018; Accepted 21 August 2018
Available online 24 August 2018

(Kratz and Nishino, 2012; Su et al., 2013; Chen and Kämäräinen, 2016). Representations at this level spontaneously capture long-range spatial and temporal dependencies. However, their applications are mainly limited to anomaly detection or activity recognition. By contrast, our aim in this paper is to understand scene dynamics, which can then be used in anomaly detection. Local spatio-temporal representations also typically suffer from high computational cost in the training and quantization stages, for example, Li et al. (2014).

The third level is trajectory/tracklet-based representation. At this level, individual tracks for a particular time duration are adopted as motion units that incorporate the necessary information for motion analysis (Zhou et al., 2011; Shao et al., 2014; Zou et al., 2015). Representations at this level robustly incorporate information about semantically meaningful moving entities (e.g., a feature point or object) for a period of time. Owing to the severe occlusions in crowded scenes, it is difficult to obtain long or complete trajectories. Hence, short trajectory fragments, called *tracklets*, have often been adopted in this family (Zhou et al., 2011). Tracklet construction is more conservative in terms of resolving tracking ambiguities than trajectories: a tracklet is stopped whenever ambiguities are detected. As a result, a tracklet is less likely to drift, and hence can be more robustly extracted, even from highly crowded scenes (Li et al., 2015).

Because of the reasons outlined above, trajectory/tracklet-based representations have been favored in many crowded scene analysis approaches. Despite this, existing trajectory/tracklet-based crowd analysis approaches still have numerous weaknesses. For instance, in Dehghan and Kalayeh (2015), after collecting the motion tracklets, a particular spatial quantization scheme is imposed by dividing the crowded scene into hexagons. A 2D orientation distribution function (ODF) is learned for each hexagon, which is time-consuming considering that the scene is divided into many such hexagons. Moreover, a post-processing step is required to smooth the identified paths.

In Zhou et al. (2011), the knowledge of scene entrances and exits is an important semantic prior for obtaining reasonable results. Conversely, such a semantic prior is not considered at all in the Bayesian approach proposed in Wang et al. (2008). For non-Bayesian approaches, such as Atev et al. (2006) and Zhang et al. (2009), it is also quite difficult to include such prior information to improve clustering.

In addition to the algorithmic limitations in the aforementioned approaches for trajectory/tracklet-based crowded scene analysis, there are other notable deficiencies in their estimated semantic regions. These estimated semantic regions, which correspond to the motion pathways in our work, are frequently observed at one of two extremes. At one extreme, they only partially cover the spatial extents of the actual semantic regions observed in the scene, that is, some of the motion dynamics that belong to a given pathway may not be accounted for. At the other extreme, one estimated pathway may cover the motion dynamics in more than one actual pathway, for example, producing a pathway that starts from or ends at more than one gate, which violates our adopted notion of a pathway that extends from one entrance gate to one exit gate. Therefore, it seems quite difficult for these approaches to provide a semantic interpretation of the actual routes of moving objects. Moreover, most of these approaches adopt off-line processing in their analysis by assuming the existence of all scene trajectories/tracklets beforehand, which makes them categorically unsuitable for real-time applications.

In this article, we introduce a robust crowded scene analysis approach that attempts to overcome most of the aforementioned limitations. Using our method, we identify motion pathways through a highly crowded scene based on tracklet clustering. After collecting the motion tracklets from a crowded scene video, the pairwise spatial/orientation similarities between them are captured using a novel measure inspired by line geometry. This measure can be flexibly tuned to account for different inter-tracklet relationships that can be encountered within a motion pathway. For tracklet clustering, we use the recent distance dependent Chinese restaurant process (DD-CRP) as a non-parametric clustering model (Blei and Frazier, 2011) and effectively adapt it to use the estimated similarities among tracklets. We also further extend this clustering model by including source/sink gate probabilities of tracklets as a high-level semantic prior. This prior is incorporated in such a manner as to ensure that each identified motion pathway starts from its associated source gate and ends at its associated sink gate with full spatial coverage. The evaluation experiments demonstrate that our model outperforms baseline methods and achieves state-of-the-art performance.

We summarize the main contributions of this work as follows: (1) a flexible tracklet similarity measure based on line geometry, (2) an adaptation of the DD-CRP model to the clustering of tracklets, (3) a new tracklet cluster likelihood (TCL) based on adopting the source/sink gate distributions of tracklets as a high-level semantic prior, (4) new comprehensive ground truth (GT) pathways for the Grand Central Station scene based on different gate annotations, provided for practical evaluation, (5) a robust quantitative/qualitative evaluation for the semantic analysis of crowded scenes based on a reliable pathway matching metric, and (6) a simple tracklet representation used for anomaly detection through the identified pathways. The first two of these contributions were presented in our previous work (Hassanein et al., 2016).

The remainder of the paper is organized as follows: In Section 2, the related work is outlined. In Section 3, the computation of pairwise tracklet similarities is explained. In Section 4, the non-parametric clustering algorithm and newly proposed cluster likelihood term are introduced. In Section 5, the overall stages of identifying motion pathways are summarized. In Section 6, the formulation of the pathway matching metric is explained. Evaluation experiments are presented in Section 7, which also includes an evaluation of anomaly detection based on the estimated motion pathways. Finally, Section 8 concludes the paper.

2. Related work

For crowded scene analysis, two main levels of analysis are considered: *macroscopic* and *microscopic* (Li et al., 2015). At the *macroscopic* level, crowd dynamics are expressed as global motion pattern(s) of a mass of objects, with no concern for the movements of individuals. By contrast, at the *microscopic* level, the movements of objects and interactions among them are considered, and the analysis occurs based on the collective information about them. To serve the aforementioned levels of analysis, many approaches have been proposed to understand and recognize the behaviors of moving objects in crowded scenes. In Section 2.1, we summarize the current modeling approaches for crowd dynamics and the level of use of each one. Then, Section 2.2 is devoted to similarity-based tracklet grouping approaches.

2.1. Modeling crowd dynamics

We can categorize recent modeling approaches for crowd dynamics into three approaches: holistic-based, agent-based, and topic models.

2.1.1. Holistic-based approaches

Holistic-based approaches consider crowd dynamics as a fluid with particles. They work better at the macroscopic level with medium and high-density crowds. Several of these analytical methods for crowd modeling have been inspired by statistical mechanics and thermodynamics (Helbing et al., 2007; Ali and Shah, 2008). A real-time crowd model was presented in Treuille et al. (2006) based on continuum dynamics. This model can yield a set of dynamic potentials and velocity fields to guide an individual's motions. In Solmaz et al. (2012), linear dynamics are assumed over the video scene, in which the linear transformation originates from the Jacobian matrix whose elements are estimated using optical flow. Stability analysis of the underlying dynamical system has been conducted to identify five types of global crowd behaviors: blocking, lanes, bottlenecks, ring/arch, and fountainheads. An alternative holistic representation of crowded scenes based on dynamic textures was proposed in Li et al. (2014).

2.1.2. Agent-based approaches

Agent-based approaches are more suitable for modeling crowd dynamics at the microscopic level for low-density crowds. Moving objects are considered as autonomous agents that actively sense the environment and make decisions according to some predefined rules. In Zhao et al. (2011), the velocity field-based social force model (SFM) is used to locate crowd behavior instability while the influence of the velocity field on the interaction force among individuals is considered. A new mixture model of dynamic agents was proposed in Zhou et al. (2015) to learn the collective dynamics of pedestrians in an unsupervised manner. The entire crowd is treated as a mixture model, and its parameters are learned from the extracted trajectories. This model can simulate crowd behaviors; however, it fails when the scene contains complex motion patterns, such as U/S-turns. in Raghavendra et al. (2011), the estimated interaction forces (driven by the SFM Helbing and Molnar, 1995) between moving objects are optimized using particle swarm optimization to guide the particle advection process over image frames.

A meta-tracking approach was proposed in Jodoin et al. (2013) to extract the dominant motion patterns and the scene gates' spatial extents from the crowded scene. The method first computes motion histograms at each pixel and then converts them into orientation distribution functions (ODFs). Given these ODFs, the meta-tracks (i.e., particle trajectories) are produced based on a particle meta-tracking procedure, which uses particles to follow the dominant flow of traffic. In a final step, a novel method is used to simultaneously identify the main entry/exit areas and recover the predominant motion patterns. However, this method still produces poor results in unstructured scenes, such as the Grand Central Station scene, and takes a great deal of time for optical flow field computation.

2.1.3. Topic model-based approaches

Recently, topic models have been borrowed from the natural language processing field and extended to capture spatial and temporal dependencies to solve computer vision problems. Particularly, these models have been recently used at both microscopic and macroscopic levels of crowd analysis (Rodriguez et al., 2009; Chen et al., 2017). For instance, the existing latent Dirichlet allocation topic model (Blei et al., 2003) was extended to learn semantic regions of crowded scenes in Zhou et al. (2011). In this approach, a topic corresponds to a semantic region, a tracklet is treated as a document, and the points on tracklets are quantized into words according to a codebook based on their locations and velocity directions. A Markov clustering topic model was proposed in Hospedales et al. (2009) to analyze dynamic scenes based on hierarchical visual activities clustering.

Furthermore, the DD-CRP model has been adopted in language modeling, computer vision problems, and clustering applications. It is expressed as a flexible class of distributions over partitions (topics) that allows for dependencies between elements (documents). It was examined for image segmentation in Ghosh et al. (2011), in which a new hierarchical extension was proposed. A tracklet clustering approach based on DD-CRP was proposed in Topkaya et al. (2015) for tracking enhancement, in which the tracklets are clustered based on their color and spatial/temporal similarities. Recently, DD-CRP has also been used in biological science, as introduced in Baldassano et al. (2015) to divide biological elements (i.e., the human brain) in a spatial map based on their connectivity properties. The identified locally connected clusters correspond to coherent semantic units in the desired part of the map.

2.2. Similarity-based tracklet grouping

Various methods have been proposed for motion path estimation based on the grouping of motion features, where the definition of a motion path varies from one method to another depending on the target application. Several of these methods depend on incorporating similarities between the motion features in the grouping process. In this section, we focus on the methods that are most closely related to ours, particularly, recent methods that incorporate tracklets as motion features.

Time-based tracklet similarity measures are common when the objective is to group tracklets that belong to the same object. For example, in Alahi et al. (2014), the objective is to group tracklets that belong to a single moving person. In this approach, a tracklet is represented as a social affinity map (SAM), which captures the relationship between that tracklet and the nearby moving tracklets. Then, the similarity between a pair of tracklets is estimated using the Hamming distance between their associated SAMs. in Chen et al. (2014), multi-person tracking is achieved by linking elementary groups of tracklets (that each consist of two tracklets), considering the time overlap, spatial distance, and motion smoothness. A coherent crowd motion detection technique was proposed in Zhou et al. (2012) for the purpose of detecting groups of objects moving together. In this approach, *coherent neighbor invariance* was introduced to characterize coherent motion in contrast to random motion. This invariance was then deployed in a clustering approach called *coherent filtering*.

Our objective is to group tracklets of objects moving in the same pathway regardless of the motion time, motion region, or motion speed within the pathway. Therefore, we deploy a time-less similarity measure. In this regard, the longest common subsequence (LCSS) between tracklets is an attractive choice. However, LCSS does not capture the continuation relationship between tracklets, which is a critical component in our similarity measure (Section 3). Different approaches have been used to overcome this limitation of LCSS. For example, Cheriyadat and Radke (2008) resorted to using full noisy trajectories instead of tracklets. The obtained motion pathways were then smoothed by fitting polynomials to cluster centers. Conversely, Chongjing et al. (2013) resorted to using tracklets of densely sampled feature points. However, computing LCSS for all pairs of densely sampled tracklets adds a significant time complexity to the analysis framework. By contrast, our proposed similarity measure deploys efficient line geometry-based calculations in estimating the similarities between tracklets. Moreover, it captures the continuation relationship between tracklets, even if they are sparsely sampled.

3. Tracklet similarity measure

The purpose of clustering tracklets in our approach is to identify motion pathways in a crowded scene. In this section, we focus on the tracklet similarity measure, variants of which are used at multiple levels of non-parametric clustering.

We would like tracklets to be clustered together when they belong to the same motion pathway. For two tracklets to belong to a motion pathway, they have to belong to a single object or two objects that originate from the same source and move toward the same sink. In this case, the two tracklets are expected to be similar in terms of their spatial layouts and global orientations. However, encoding this similarity in a single measure is not trivial because of the many cases that can be encountered in practice.

Fig. 1(a) shows a hypothetical scene that has one source (A) and two sinks (B and C), with four overlaid tracklets. Consider the two tracklets t_1 and t_2. Although they both originate from the same source and are spatially close, perceptually, they do not seem to belong to the same motion pathway. This phenomenon can be understood by inspecting the geometric relationship between the two tracklets: If they belonged to the same motion pathway, then they would have been at *the same stage* (which is the beginning, in this case) of that pathway, which means that they should have been almost parallel. However, because of their divergence in orientation, they are not perceived to be on the same pathway. Now, consider the two tracklets t_1 and t_3. The difference in orientation between them is higher than that between t_1 and t_2. Despite this, perceptually, tracklet t_3 seems to be a continuation of t_1, that is, the two tracklets could be on the same pathway but at *two different stages*.

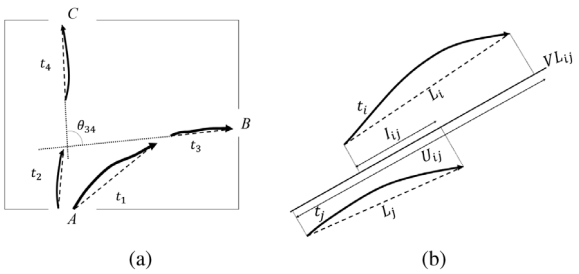


Fig. 1. (a) Hypothetical scene with one source (A), two sinks (B and C), and four tracklets ($t_1, \ldots, t_4$). The directed line segment associated with each tracklet is indicated by a dashed line. θ_{34} is the estimated angle between t_3 and t_4. (b) Computation of the overlap ratio between two tracklets, t_i and t_j, as $O_{ij} = \frac{I_{ij}}{U_{ij}}$.

From the discussion above, the approach that we use to interpret the geometric relationship between two tracklets depends on the degree to which they are perceived to be at the same stage of a motion pathway. In our approach, we estimate this by the degree of overlap between the two tracklets. Our intuition is that the greater the overlap between two tracklets, the more likely that they belong to the same stage of a pathway, and vice versa. Our proposed similarity measure incorporates both the *spatial* and *orientation* similarities between tracklets while taking into account the *overlap* between them (see Fig. 1(b)). Next, we explain each of these components.

The spatial similarity between two tracklets is estimated using two distance functions: the *Hausdorff* distance and minimum point-to-point distance. Each distance function will be considered later in the proposed clustering model (see Section 4.3).

Let $t_i = (p_{i1}, p_{i2}, \ldots, p_{in})$ and $t_j = (p_{j1}, p_{j2}, \ldots, p_{jn})$ be two tracklets such that each tracklet is identified by n points, and each point is identified by its (x, y) coordinates in the image's frame. *Hausdorff* distance $d_H(t_i, t_j)$ between the two tracklets is computed as

$$d_H(t_i, t_j) = \max\left\{\Delta(t_i, t_j), \Delta(t_j, t_i)\right\}, \tag{1}$$

$$\Delta(t_i, t_j) = \max_{p_{ik} \in t_i} \min_{p_{jl} \in t_j} \mathrm{d}(p_{ik}, p_{jl}), \tag{2}$$

where $\mathrm{d}(p_{ik}, p_{jl})$ is the Euclidean distance between the kth point of t_i and the lth point of t_j.

By contrast, the minimum point-to-point distance is expressed as

$$d_M(t_i, t_j) = \min_{p_{ik} \in t_i, p_{jl} \in t_j} \mathrm{d}(p_{ik}, p_{jl}). \tag{3}$$

In the following, we refer to the distance between two tracklets t_i and t_j as δ_{ij}, regardless of the type of distance function. In the following section, we explain the cases in which we apply each type.

The orientation similarity between a pair of tracklets is estimated by approximating each tracklet as a *directed line segment*. This directed line extends from the starting point of a tracklet to its ending point, as shown in Fig. 1(a). Note that because tracklets are typically constructed over short time periods, approximating them using directed line segments should be acceptable in most cases. For two tracklets t_i and t_j, the angle between them, θ_{ij}, is estimated as the angle between their two associated directed line segments.

The overall similarity measure between a pair of tracklets t_i and t_j is defined as

$$Sim(t_i, t_j) = e^{-\left(\frac{\theta_{ij}}{\sigma_{\theta ij}}\right)^2} e^{-\left(\frac{\delta_{ij}}{\sigma_{\delta ij}}\right)^2}, \tag{4}$$

where the two variables $\sigma_{\theta ij}$ and $\sigma_{\delta ij}$ represent the tolerance values in the orientation and spatial dimensions, respectively. The higher the tolerance value, the less sensitive the similarity function to changes in the associated variable. The similarity measure takes values in the interval $[0, 1]$, where 0 indicates no similarity and 1 indicates identical tracklets.

As the notation in Eq. (4) indicates, the tolerance values are associated with the two particular tracklets for which the similarity is computed. These tolerance values are computed as follows:

$$\begin{aligned} \sigma_{\theta ij} &= \sigma_{\theta max} + O_{ij} \cdot (\sigma_{\theta min} - \sigma_{\theta max}), \\ \sigma_{\delta ij} &= \sigma_{\delta min} + O_{ij} \cdot (\sigma_{\delta max} - \sigma_{\delta min}), \end{aligned} \tag{5}$$

where O_{ij} indicates the degree of overlap between the two tracklets t_i and t_j, which takes a value in the interval $[0, 1]$ (as explained below). Hence, depending on the overlap ratio between the two tracklets, each tolerance value is linearly experimentally chosen from an interval bounded by the minimum and maximum values, that is, $\sigma_{\theta ij} \in [\sigma_{\theta min}, \sigma_{\theta max}]$ and $\sigma_{\delta ij} \in [\sigma_{\delta min}, \sigma_{\delta max}]$. Thus, the greater the overlap between the two tracklets, the more tolerance we give to spatial dissimilarity and less tolerance we give to orientation dissimilarity. Note that the min–max intervals for the tolerance values for our experiments are mentioned in Section 5.

To estimate the degree of overlap between two tracklets, we again use the directed line segment approximation. Particularly, we estimate the overlap between tracklets t_i and t_j as the overlap ratio between the two associated directed line segments, L_i and L_j, when projected on an intermediate line called the virtual line, VL_{ij}. We adopt the idea of the virtual line from Etemadi et al. (1991). The computation is illustrated in Fig. 1(b). Virtual line VL_{ij} has the average orientation of the two line segments and passes through the center of mass of all line segments' points.

4. Non-parametric tracklet clustering

In clustering analysis, determining the appropriate number of clusters within the data is a challenging problem. The number of clusters is required for many clustering algorithms as a pre-determined hyperparameter on which the quality of the resulting clustering strongly depends. However, in real-world applications, the number of clusters is mostly unknown (i.e., latent). Moreover, in dynamic scenarios, the number of clusters is expected to change over time as more observations are collected and incorporated. This problem can be generally overcome using non-parametric approaches; however, the choice of a flexible clustering algorithm is pivotal. One of the recent reliable and flexible models is the *Dirichlet process mixture model (DPMM)* (Antoniak, 1974). In this type of model, the clustering problem is represented as a distribution over an infinite unknown number of mixture components (i.e., clusters). DPMMs have already been applied in natural language understanding (Liang et al., 2007) and document clustering (Teh, 2006). Furthermore, they are currently being adopted in visual scene analysis (Sudderth et al., 2008) and image segmentation (Ghosh et al., 2011; Baldassano et al., 2015).

One DPMM representation is the *Chinese restaurant process (CRP)*. In the restaurant analogy, a sequence of customers are going to be seated at an infinite number of tables in a restaurant. The first customer gains probability one to sit at a given table. Any subsequent customer sits at a previously occupied table with a probability proportional to the number of people already seated at the table and sits at a new table with a probability proportional to a specific concentration parameter (Blei and Frazier, 2011).

4.1. DD-CRP clustering model

In infinite clustering models, the data points to be clustered may be ordered in time, such as time-stamped articles, or in space, such as pixels in an image. This reflects the dependencies among them and violates the exchangeability property of the basic Dirichlet process (Blei and Frazier, 2011). Therefore, the recent DD-CRP model has been developed as a variation of DPMMs to manage these types of dependencies. The DD-CRP model follows the same restaurant analogy of the traditional CRP process (Aldous, 1985). However, it represents

data partitioning through customer assignments instead of table assignments. Customer assignments depend on their pairwise distances. Moreover, customers are implicitly assigned to tables by considering their reachability to each other through their pairwise assignments. According to this analogy, customer assignments are conditioned on the distances between customers and drawn independently according to the following scheme:

$$P(c_i = j|D,\alpha) \propto \begin{cases} f(d_{ij}) & i \neq j \\ \alpha & i = j, \end{cases} \tag{6}$$

where the assignment of the ith customer, c_i, is the index of another customer with whom the ith customer is seated (i.e., both customers are seated at the same table); d_{ij} is the distance between customers i and j; D denotes the distance matrix between all customers; α is the concentration parameter; and f is a decay function (for decay function details, see Blei and Frazier, 2011). Note that j ranges over the entire set of customers; hence, any customer may sit with any other customer. The posterior inference for DD-CRP, based on Gibbs sampling, is implemented by sampling the customer assignments iteratively from the conditional distribution for each new customer assignment as follows (Blei and Frazier, 2011):

$$\begin{aligned} P(c_i^{new}|\mathbf{c}_{-i}, O, D, f, \alpha, G_0) \propto \\ P(c_i|D,\alpha) \times P(O|W(\mathbf{c}_{-i} \cup c_i^{new}), G_0), \end{aligned} \tag{7}$$

where c_i^{new} is a new assignment for the ith customer; $\mathbf{c}_{-i}$ is the vector of all customer assignments, excluding the ith customer; G_0 is the base distribution[2]; and O is the set of all customers. The first term on the right-hand side is the DD-CRP prior for the new customer assignment (Eq. (6)). The second term is the likelihood of the observations given customer partitioning over tables, which is denoted by $W(\mathbf{c}_{-i} \cup c_i^{new})$. This can be considered as removing the current link from the ith customer and then considering how each alternative new link affects the likelihood of the observations (Blei and Frazier, 2011).

4.2. Tracklet clustering based on DD-CRP

The recent DD-CRP model was developed to manage dependencies between data points during clustering. We adapt this model to manage the pairwise similarities between tracklets. In our proposed DD-CRP clustering framework, *tracklets correspond to customers, and pathways are the output clusters*. Given similarity matrix S among tracklets, we modify the prior probability (Eq. (6)), which represents the first term in the posterior (Eq. (7)), as follows:

$$P(c_i = j|S,\alpha) \propto \begin{cases} s_{ij} \quad if \quad i \neq j \\ \alpha \quad if \quad i = j \end{cases}, \tag{8}$$

where s_{ij} is the pairwise similarity between tracklets i and j. We also represent the observation likelihood, which is the second term of the posterior (Eq. (7)), as the product of likelihood estimates for individual clusters, which we refer to as the *TCL*. The TCL of cluster C is in turn defined as the geometric mean of the maximal pairwise similarities between the N_C tracklets in the cluster, $\{t_1, \ldots, t_{N_C}\}$. Hence, in our model,

$$P\left(O|W\left(c_{-i} \cup c_i^{new}\right), G_0\right) = \prod_{C \in \mathcal{R}} TCL_C \ , \tag{9}$$

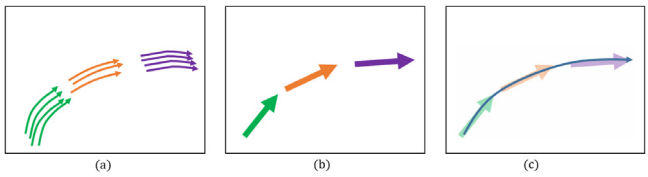


Fig. 2. Illustrative figure: (a) parallel tracklets appear in separate clusters that resulted from the first level of clustering, (b) corresponding collinear representatives for each cluster, and (c) global collinear representative obtained from the second level of clustering based on curve fitting.

$$TCL_C = \left(\prod_{n=1}^{N_C} \max_{j=1..N_C, j \neq n} P\left(t_n|t_j, G_0\right) \right)^{\frac{1}{N_C}}, \tag{10}$$

where $\mathcal{R}$ is the collection of tracklet clusters. $P(t_n|t_j, G_0)$ is chosen to be proportional to the pairwise similarity between tracklets t_n and t_j.

4.3. Two-level hierarchical tracklet clustering

The clustering of tracklets in our approach is hierarchically performed over two levels. We collect the tracklets of a crowd video at every time stage (each stage is 30 frames in our experiments). At the first level of clustering, the collected tracklets for each stage are clustered using the DD-CRP model based on a parallelism criterion for the similarity measure, which attempts to group only parallel tracklets, as shown in Fig. 2(a). This criterion is incorporated into our clustering framework by adjusting the limits on the tolerance values in the similarity measure $\sigma_{\theta min/max}$, and $\sigma_{\delta min/max}$ in Eq. (5), and using the *Hausdorff* distance from Eq. (1).

As an output of this level, we represent each resulting cluster by a single representative that is a directed line segment obtained from the associated cluster's tracklets, as shown in Fig. 2(b). Following (Etemadi et al., 1991), each cluster's representative directed line segment has the following properties:

- Its orientation is the average orientation of the cluster's tracklets.
- It passes through the center of mass of all tracklets' points in the cluster.

It extends between the limits of all tracklets' points in the cluster after projecting these tracklets' points on the line that contains the cluster's representative directed line segment.

At the second level of clustering, the resulting representatives are grouped again based on the DD-CRP model. However, in this case, the similarity function is adjusted to group collinear representatives[3] by slightly changing the limits of the tolerance and including the *minimum point-to-point* distance from Eq. (3). The actual values of the tolerance parameters that we use in the Grand Central Station experiments are mentioned in Section 5. For each cluster of representatives that results from the second level of clustering, a global representative is estimated by adopting a piecewise cubic interpolation curve that fits over the cluster representatives' points, as shown in Fig. 2(c). These fitted global representatives are considered as tracklets in subsequent clustering iterations, which means that they are approximated to directed line segments for similarity computation.[4] The two-level hierarchical clustering process is repeated incrementally until the end of the video.

Note that the simple fitting of tracklets in a cluster, as a directed line segment at the first clustering level or a curve at the second clustering

[2] In the original DD-CRP document clustering model setting, base distribution G_0 is typically a Dirichlet distribution over distributions of words (Blei and Frazier, 2011). In our adapted DD-CRP tracklet clustering model, we build our cluster likelihood function, which is mainly proportional to the similarities between tracklets as an empirical assumption. Thus, we do not take the base distribution into account in our calculations. We use it in our formulation to follow the same notation as the original model.

[3] For simplicity, we informally use the term 'collinear' to refer to pairs of cluster representatives whose geometric relationship indicates that one of them is a potential continuation of the other.

[4] The directed line segment representation obtained from a fitted curve, as introduced in this paper, has much more accurate end points compared with that obtained from directly fitting a line, as we did in our previous work (Hassanein et al., 2016).

level, makes the linking of clusters much more computationally efficient than using, for example, the Hungarian algorithm (Perera et al., 2006) or globally optimal greedy approach of Pirsiavash et al. (2011). This is performed without compromising our notion of parallelism or continuity between tracklets at the same stage or consecutive stages of a pathway, respectively.

It is worth noting that we choose the *Hausdorff* distance for the spatial similarity at the first level of clustering because it captures the separation between tracklets whether they are parallel or intersecting, whereas the *minimum point-to-point* distance becomes zero if the two tracklets intersect. However, at the second level of clustering, the *Hausdorff* distance can become too large for collinear tracklets, whereas the *minimum point-to-point* distance captures the notion of spatial proximity for tracklet continuity. At the end of the video frames, the final output clusters from the second level correspond to the motion pathways.

4.4. Semantic prior-based pathway detection

In our work, we introduce the notion of a *motion pathway* as a series of spatially coherent linked groups of tracklets that share the same motion route from a unique source to a unique sink. To identify such coherent pathways, we extend our DD-CRP clustering model by including the possible sources and sinks of tracklets through the crowded scene as a high-level semantic prior.

Source/sink priors have been explored in many existing works (Hassanein et al., 2016; Dehghan and Kalayeh, 2015; Jodoin et al., 2013) as essential scene structures. However, to the best of our knowledge, they have not been well explored in terms of improving the clustering of tracklets. Our work demonstrates that incorporating such priors in our non-parametric clustering model improves the results of tracklet clustering. Source/sink priors were introduced in Zhou et al. (2011) as hardwired gate labels (h = source and m = sink) to improve semantic region segmentation and trajectory clustering. In our work, we incorporate such priors in a more flexible probabilistic form by adopting the entering and exiting gates' probability distributions for each tracklet.

4.4.1. Prior-based tracklet cluster likelihood

We argue that introducing prior information about the scene layout, such as the spatial extents of crowded scene gates, can significantly improve the clustering of tracklets and produce more coherent pathways, as shown in Fig. 4(b). Therefore, based on this argument, we assume that the boundaries of the source and sink regions of a crowded scene are roughly known (e.g., manually annotated). To capture the gate annotations, we use image processing methods for the preprocessing step. First, we manually mark the polygons (i.e., gate boundaries) for each gate region and then save the (x, y) coordinates for all gate polygons. Next, we extract the binary mask for each gate region that is bounded by its polygon and then identify the centroid point coordinate for each extracted mask. As a result, we can estimate the entering and exiting gate probability distributions for the generated tracklets. We further reformulate the TCL term of the adapted DD-CRP model in Eq. (10) to incorporate these estimated distributions. We consider this new TCL term as the product of two terms: (1) the geometric mean of the product of the maximal pairwise similarities for the tracklets in a specific cluster C and (2) the source/sink consistency, $\mathcal{Q}^C_{sk}$, of that cluster (Eq. (15)). This consistency term reflects the interrelationships between tracklets that rely on their shared source/sink prior distributions. The derivation of this consistency term is introduced in detail in Section 4.4.2. The new TCL term is formulated as

$$TCL_C = \left(\prod_{n=1}^{N_C} \max_{j=1..N_C, j\neq n} P\left(t_n | t_j, G_0\right) \right)^{\frac{1}{N_C}} \mathcal{Q}^C_{sk}. \tag{11}$$

We then adapt the posterior inference of the DD-CRP model in Eq. (7). The new TCL term (Eq. (11)) is then used in place of the old one (Eq. (10)) in the observation likelihood model (Eq. (9)).

In the following, we explain the estimation of the source/sink distributions (i.e., the entering and exiting gate probability distributions) for a cluster of tracklets. As a result, we justify how these distributions contribute in the adapted DD-CRP clustering model.

4.4.2. Estimating cluster source/sink consistency

For the effective estimation of such source/sink distributions, we consider each tracklet as having frontal and rearward fields of view (FOVs). We adopt both FOVs for the tracklet terminal points. In our representation, the frontal FOV defines the expected sinks for that tracklet, whereas the rearward FOV defines its expected sources. For each tracklet, we extend a set of directed lines from its start terminal point to all gate centroids, and another set of directed lines from the end terminal point of the tracklet to the same centroids. For instance, the directed lines that extend from the start terminal of a tracklet t_i to the gate centroids are shown in Fig. 3(a).

As a special case, each tracklet terminal point (starting point or ending point) that lies within a gate polygon (source or sink) receives a probability of one for that gate and zeros for the remaining gates, and it is not necessary to extend the directed lines. For the other possible cases, we compute the angles between the associated directed line that represents the tracklet (simply, this directed line extends from the starting point of the tracklet to its ending point, as shown in Fig. 1(a)), and the directed lines that extend from the start and end terminal points of that tracklet to the gate centroids. The two obtained sets of angles (one for sources and one for sinks) are then normalized (in absolute values) to create a valid probability distribution. In Fig. 5, we show the normalized source and sink prior distributions for an example tracklet, t_i. Note how the most likely source and sink gates for that tracklet (i.e., gates 11 and 5, respectively, in this illustrative case) are those gates that are perceptually closest to the direction of motion represented by the tracklet.

Let t^C_i be a tracklet in a given cluster of tracklets C and g_j be a gate in the scene. Then, let $P_s(g_j|t^C_i)$ and $P_k(g_j|t^C_i)$ be the probabilities that gate g_j is the source gate and sink gate for tracklet t^C_i, respectively. Note that, to define the valid source and sink gate probability distributions for tracklet t^C_i, the two probabilities must satisfy the following two constraints:

i. $0 \leq P_h(g_j|t^C_i) \leq 1 \ , j \in 1, .., M, \quad h \in \{s, k\}$
ii. $\sum_{j=1}^{M} P(g_j|t^C_i) = 1,$

where M is the number of gates in the scene.

The source and sink gate probability distributions defined above can be expressed as two probability vectors for each tracklet t_i in cluster C: $V_s^{t^C_i}$ and $V_k^{t^C_i}$, respectively, as in Eq. (12). Each of these vectors has length M and sums to one (see Fig. 5). In our experiments, we used $M = 14$ for the Grand Central Station scene.

$$V_h^{t^C_i} = \begin{bmatrix} P_h(g_1|t^C_i) \\ . \\ . \\ . \\ P_h(g_M|t^C_i), \end{bmatrix} \tag{12}$$

where subscript h denotes the gate type, $h \in \{s, k\}$. We compute the overall gates distributions $\mathcal{D}_h$ for cluster C that has N_C tracklets as follows:

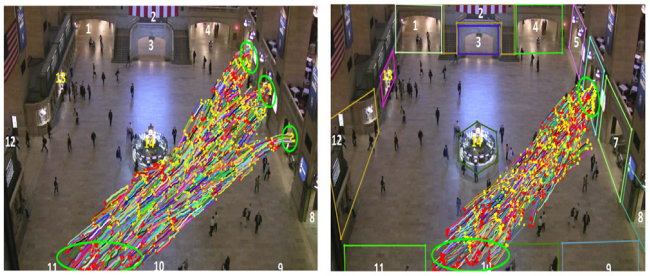
$$\mathcal{D}_h \propto \prod_{i=1}^{N_C} V_h^{t^C_i}. \tag{13}$$

We use Eq. (13) to compute both the source and sink gate probability distributions separately. The $\prod$ operator denotes element-wise vector multiplication for the N_C probability vectors. To create a valid probability distribution, the resulting distribution, $\mathcal{D}_h$, is then normalized so that its elements sum to one.

It is desirable for the estimated distributions for a cluster of tracklets to be concentrated at a particular source gate and particular sink gate

(a) (b) (c)

Fig. 3. (a) Hypothetical scene that demonstrates the directed lines that extend from the start terminal of a tracklet to the gate centroids, (b) our 14-gate annotation, and (C) the floor plan of Grand Central Station (image credit: https://www.cultofmac.com/82433/confirmed-apple-to-open-biggest-store-yet-in-grand-central-terminal).



(a) (b)

Fig. 4. (a) Motion pathway obtained from our clustering model without providing the semantic prior (green ellipses indicate the most probable gates). (b) Corresponding pathway after introducing the source/sink priors.

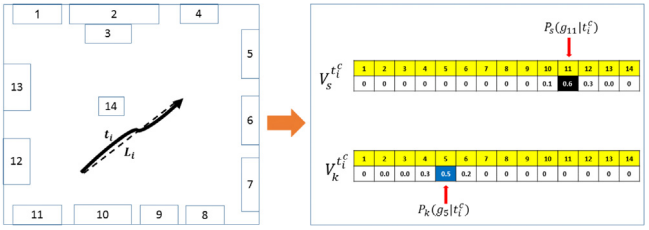


Fig. 5. Computation of both source and sink gate probability distributions for hypothetical tracklet t_i in cluster C based on its frontal and rearward FOVs. Note that the most likely source and sink gates are highlighted in black and blue, respectively (11 and 5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

because concentrated distributions reflect agreement among the cluster's tracklets on specific source and sink gates. This agreement may be considered as an *extra criterion* of similarity among tracklets in the same cluster. However, the real importance of this agreement is to eliminate cases of severe disagreement, such as two tracklets that are close in space and orientation but exiting or entering from two different, but nearby, gates. Such tracklets are expected to have source/sink gate probability distributions concentrated at different gates. Hence, putting them in the same cluster never results in concentrated distributions, and hence reduces the cluster likelihood.

To examine the concentration of the estimated distributions, we adopt the fundamental concept of *Shannon entropy* (Shannon, 1948) from information theory. The *entropy* measure provides its highest value for a uniform distribution and its lowest value for a distribution that is concentrated completely at a single point (i.e., a Kronecker delta function) (MacKay, 2003). Given gate (source or sink) distribution $\mathcal{D}_h$ of a cluster of tracklets, C, we measure the entropy of C as

$$H_h^C = -\sum_{l=1}^{M} D_h^l \log_2 D_h^l \tag{14}$$

where l is an index for the D_h elements. For a particular cluster of tracklets, C, we compute the entropies for both its source and sink distributions as H_s^C and H_k^C, respectively. Hence, we define the overall cluster source/sink consistency term $\mathcal{Q}_{sk}^C$ in Eq. (11) as follows:

$$\mathcal{Q}_{sk}^C = exp^{-\left(H_s^C + H_k^C\right)}. \tag{15}$$

This resulting term takes its maximum value if the two source/sink distributions of the cluster of tracklets are perfectly concentrated. By contrast, the flatter either or both of the two distributions, the closer the term to its minimum value. The involvement of this consistency term in the cluster likelihood function of the DD-CRP clustering model, Eq. (11), encourages having clusters with tracklets that originate from and reach the same gates. This is because such tracklets have concentrated distributions and force the entire cluster source/sink distribution to be concentrated as well. Conversely, it penalizes clusters that have tracklets with inconsistent source/sink gate distributions, for example, coherent tracklets entering or exiting nearby gates.

For instance, the output pathway in Fig. 4(b) demonstrates the impact of involving such estimated distributions on the clustering results. The pathways identified using our prior-based approach still respect our notion of a motion pathway in which the tracklets of pathways are sufficiently compact and spatially extended from a unique source to a unique sink. By contrast, the pathways obtained without incorporating the estimated distributions demonstrate pathways associated with multiple terminal gates, as shown in Fig. 4(a). In the latter case, it is understandable that the existence of such cases is due to the high pairwise spatial/orientation similarities between tracklets at those terminals' gates. Such results violate our notion of a motion pathway and negatively affect the evaluation results.

5. Identifying motion pathways

Our proposed clustering framework follows the two hierarchical levels of the DD-CRP clustering model that was explained in Section 4.2. We first collect the tracklets from a crowd video at a time stage of every 30 frames, and the two DD-CRP clustering levels mentioned above are applied incrementally for every group of tracklets that are captured during each time stage.

At the first level of clustering, the DD-CRP clustering model attempts to group parallel tracklets into separate clusters. As a result, we estimate a single representative for each cluster. Before providing those representatives to the second level of clustering, the source/sink gate probability distributions are estimated and assigned to each representative. Note that the TCL term at this level depends only on the geometric mean of the mean of maximal pairwise similarities between tracklets (Eq. (10)).

At the second level of clustering, the estimated representatives of the first level are clustered again, but based on collinear criteria. At this level, the new source/sink distribution-based TCL in Eq. (11) is

adopted. As a result, for each cluster of collinear representatives, we create a simple piecewise cubic interpolation curve that fits their xy point coordinates. The output fitted curve is adopted as a global cluster representative, as shown in Fig. 2(c), and its source/sink distributions are estimated from the source/sink distributions of the low-level representatives using Eq. (13). In fact, to have accurate source/sink distributions, only the low-level representatives close to an end point of the fitted curve are used to estimate the distribution for the gate associated with that end point. Then, we check the estimated source/sink distributions for each global representative before concatenating the results with the cluster representatives of the next time stage. If the global representative source and sink distributions are concentrated completely at a single source gate and single sink gate (i.e., a Kronecker delta function), then this global representative and its associated tracklets are separated, and a pathway is recognized between the two identified source and sink gates. As a result, each pathway is formed incrementally based on the evolution of dynamics through the different time stages of the crowd video.

It worth noting that, for our experiments on the Grand Central Station dataset, the tolerances of the proposed similarity measure were experimentally selected. For the parallel clustering stage, we selected the orientation tolerances $[\sigma_{\theta min}, \sigma_{\theta max}] = [3, 6]$ and distance tolerances $[\sigma_{\delta min}, \sigma_{\delta max}] = [25, 35]$. For the collinear clustering stage, we selected the orientation tolerances $[\sigma_{\theta min}, \sigma_{\theta max}] = [1, 3]$ and the distance tolerances $[\sigma_{\delta min}, \sigma_{\delta max}] = [7, 14]$.

6. Pathway matching score (PMS)

Introducing appropriate quantitative measures that capture our perception of the resulting pathways is an important objective toward achieving a robust evaluation of the proposed method against the GT. Additionally, this provides common ground for a comparison with other state-of-the-art methods. Therefore, we introduce a reliable measure that can serve two purposes: (1) to match a predicted pathway with a GT pathway and (2) provide a quantitative value for how good this match is. In the remainder of this section, we simply refer to this measure as a matching score. Our proposed matching score incorporates both the spatial and motion orientation information of the pathway of interest.

Consider an identified pathway $\mathcal{P}$ that we need to match with a GT pathway $\mathcal{G}$. We start by constructing probabilistic representations for the spatial and motion orientation information for the two pathways. For the spatial information, the pathway's trajectories/tracklets are overlaid and accumulated on top of one another to construct a spatial probability (heat) map. This map represents the pathway's spatial extent in the scene and the level of activity at each point within it, as shown in Fig. 6(b).

The ***spatial matching score*** between detected pathway $\mathcal{P}$ and corresponding GT pathway $\mathcal{G}$ is then captured by calculating the quadratic chi-square distance between their corresponding spatial probability maps. Note that both heat maps are normalized to represent valid probability distributions.

By contrast, the motion orientation through a given pathway is represented by computing the motion direction between each pair of adjacent (x, y) points of the pathway trajectories/tracklets. These motion orientations are quantized and aggregated into a motion orientation histogram that describes the distribution of motion orientations through the pathway, as shown in Fig. 6(c).

The ***motion orientation matching score*** is estimated based on the quadratic chi-square distance between both the motion orientation histograms of the discovered pathway and GT pathway.

Note that the quadratic chi-square distance is defined as

$$\chi^2(X, Y) = \frac{1}{2} \sum_i \frac{(X_i - Y_i)^2}{(X_i + Y_i)} \quad , \tag{16}$$

Table 1
Evaluation of our approach versus the GT. GT-P/GT-G are the GT counts for pathways and gates, respectively. TD-P/FD-P are true detections/false detections for pathways, respectively. D-P is the total number of discovered pathways. *Pr* and *Rc* denote the pathway detection's precision and recall, respectively.

Method	GT-P	GT-G	D-P	TD-P	FD-P	Pr (%)	Rc (%)
Proposed	176	14	125	**124**	**1**	**99.20**	**70.45**

where X and Y are two discrete probability distributions. Typically, the χ^2 distance is used to test the fit between a distribution and observed empirical frequencies (Pele and Werman, 2010).

The matching scores are then obtained by converting the spatial and orientation chi-square distances (χ^2_s and χ^2_o, respectively) into similarity scores ($S_{\chi^2_s}$ and $S_{\chi^2_o}$, respectively) simply by subtracting the distance from one (because Eq. (16) produces a normalized distance in the interval $[0, 1]$).

We introduce the overall ***pathway matching score*** (PMS) as the average of both the spatial and motion orientation matching scores as follows:

$$PMS(\mathcal{P}, \mathcal{G}) = 0.5 \left(S_{\chi^2_s} + S_{\chi^2_o} \right). \tag{17}$$

This proposed metric takes values in the interval $[0, 1]$, where 0 indicates very poor matching and 1 indicates perfect matching. Because we depend on this metric for demonstrating our qualitative perception of the resulting pathways, we qualify the corresponding results as 'qualitative results'.

7. Experimental evaluation

We conducted our experiments on multiple datasets. However, our analysis was mainly performed on the new challenging New York's Grand Central Station scene (Yi et al., 2015). We started by sampling 50,000 frames from the video scene and used the KLT tracker (Tomasi and Kanade, 1991) for collecting the tracklets. All tracklets had a fixed length of 30 frames.

In the following subsections, we divide the evaluation experiments into two main parts. The first part focuses on the evaluation of the detected motion pathways using our approach against the GT and state-of-the-art methods. The second part describes the evaluation of anomaly detection using normalcy models that are built based on the detected motion pathways.

7.1. Evaluation of pathway detection

For a quantitative evaluation, we adopted the recall (R_c) and precision (P_r) measures. The recall defines the number of truly identified pathways (TD-P) out of the total number of GT pathways (GT-P), whereas the precision defines the number of truly detected pathways (TD-P) out of the total number of predicted pathways (D-P).

7.1.1. Evaluation results: ground truth pathways

To the best of our knowledge, the Grand Central Station scene still lacks extracted GT pathways. To solve this problem we used the recent large-scale pedestrian walking route dataset (Yi et al., 2015) to extract the first GT pathways for such a scene.[5] The walking route dataset provides manually annotated walking paths of 12,684 pedestrians through Grand Central Station during crowded periods. The complete path (i.e., trajectory) for each pedestrian is labeled from the time of entering to the time of leaving the scene.

For our evaluation experiments, we introduced a new comprehensive GT pathways dataset that best describes the overall motion

[5] The GT pathways, gate annotations, and more detailed results are available online http://cps.ecce.ejust.edu.eg/cviu17/.

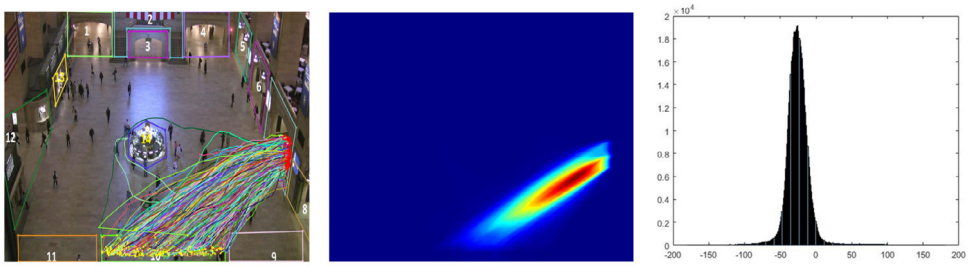


Fig. 6. (a) GT pathway, (b) heat map of the pathway's spatial extent, and (c) pathway motion orientation histogram.

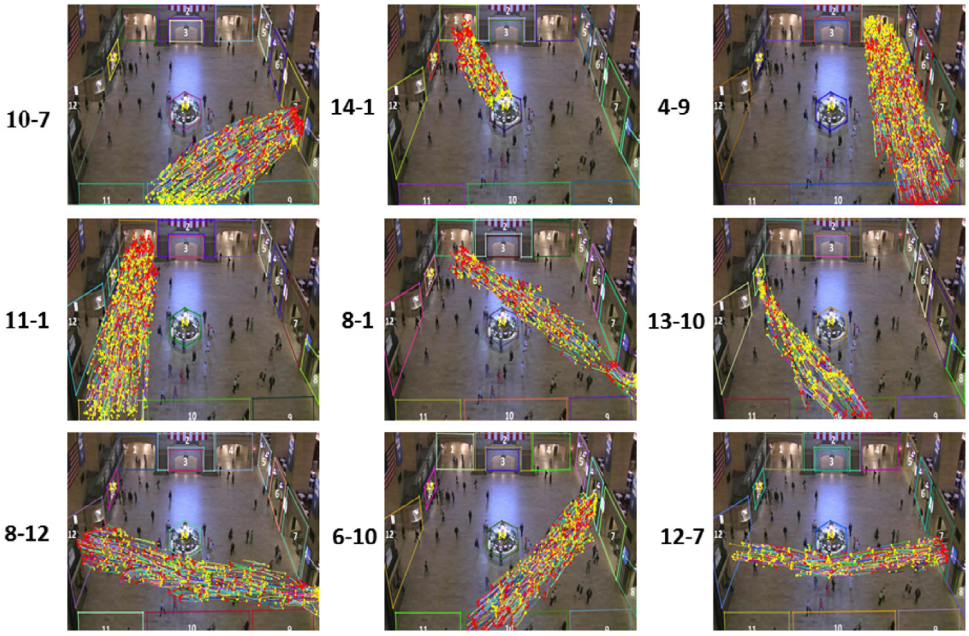


Fig. 7. Samples from our identified pathways (prior-based approach) using 14-gate annotation. Each pathway has its estimated source and sink on the left-hand side (**source–sink**).

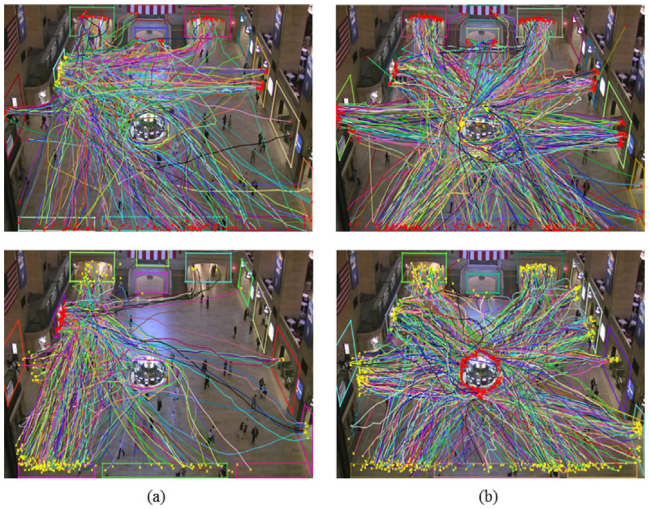


Fig. 8. (a) GT trajectories that start from/stop at the ticket window (gate 13) and (b) the same for the information booth (gate 14). Start terminal points of GT trajectories are yellow and end terminal points are red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dynamics in the Grand Central Station scene based on 14-gate annotation. First, we manually annotated the spatial extents for 14 gates of the scene. This gate annotation almost matched the actual spatial extents of the gates on the floor plan of the main concourse hall of Grand Central Station (see Figs. 3(b) and 3(c)). We assumed that all the gates were bi-directional. For each pair of annotated gates (one as a source and the other as a sink), we extracted, from the walking route dataset (Yi et al., 2015), the pedestrian trajectories that originated from the annotated source and terminated at the annotated sink. If such trajectories existed, then the pathway was considered to be in the new GT pathway dataset.

Furthermore, we considered the ticket window and information booth as gates, which were gate 13 and gate 14, respectively, in our annotation. Our motivation for considering them as gates (terminal points) is as follows (to the best of our knowledge, they have never been considered by previous approaches). Suppose, for instance, that a person enters at gate 1 with the intention to go to the ticket window (gate 13) and then proceed to gate 7 to leave. The same person has two goals, each of them is active within a given time period. Based on this understanding, we adopted them as gates because they were still active regions and represented desirable goals for the same individual to access them (in/out). The provided walking route trajectory dataset (Yi et al., 2015) is concerned only with a person's trajectory from the time that he/she appears to the time that he/she disappears, regardless of his/her intention or behavior. However, we are concerned with both the person and his/her goal for a specific period. Each person's goal for us represents a trajectory segment that belongs to a given pathway.

Based on the proposal above, we performed a preprocessing step on the original walking route trajectories. Specifically, the trajectory

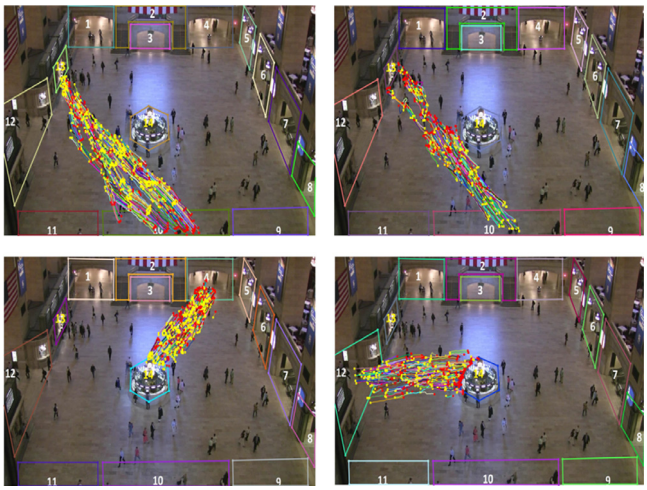


Fig. 9. Samples from the identified pathways from/to the ticket window, gate 13 (first row), and the information booth, gate 14 (second row).

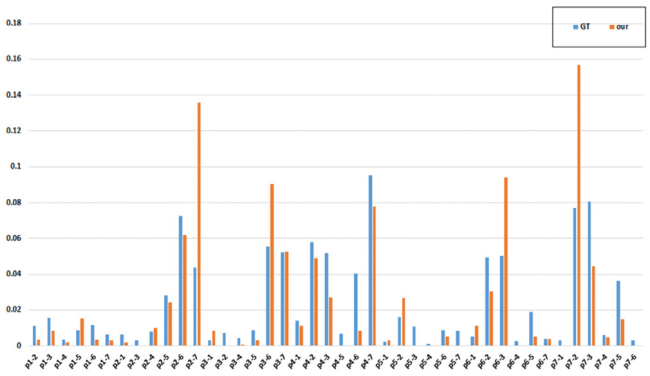


Fig. 10. Trajectory/tracklet densities through GT pathways and our discovered pathways based on seven-gate annotation. The density of the identified pathways relatively corresponds to the density of the GT pathways. The trajectory/tracklet counts are normalized. The horizontal axis represents the pathways label (source–sink) and the vertical axis represents the normalized count.

that entered either the ticket window or the information booth, stayed there for a short time, and then made a motion transition out of these regions, had to be broken at the point in time that this transition was made. Hence, a single original trajectory was segmented into multiple trajectories, and the GT pathways from/to these two gates and the remaining gates were considered. This preprocessing step increased the total number of preprocessed GT trajectories to 13,811. Fig. 8 shows a sample from the GT trajectories after this preprocessing step. Fig. 9 demonstrates samples of discovered motion pathways involving these two gates. Quantitatively, we successfully identified 30 motion pathways relevant to these two gates (from/to/in-between). After adopting all possible pathways, we obtained 176 GT pathways. We argue that this extended GT pathways dataset and our gate annotation are valuable for the robust evaluation of our pathway detection method in the Grand Central Station scene. The evaluation results of our prior-based approach against GT pathways based on 14-gate annotation are presented in Table 1. After applying our approach, we successfully identified 124 pathways out of 176 GT pathways and achieved recall rate $Rc = 70.45\%$. Additionally, observe that our identified pathways appear to be close to the GT pathways in their desirable compactness, spatial extents, and trajectory/tracklet densities. Fig. 7 shows samples of the pathways obtained in this experiment. Note how the spatial extent of a pathway goes from a unique source to a unique sink, which matches our intuition about a motion pathway. Fig. 11 demonstrates our identified pathways versus the GT for different gate annotations. These pathways also spatially extend from a unique source to a unique sink. These results strongly demonstrate the impact of involving source/sink prior information in our clustering model in addition to the pairwise similarities between tracklets.

Another approach to evaluate the predicted pathways versus the GT pathways is to estimate the similarity between the density of a predicted pathway and that of a real pathway. The probability distribution in Fig. 10 shows the detected pathways' densities versus the densities of the corresponding GT pathways. We can observe that the dense identified pathways roughly correspond to the dense GT pathways and the sparse pathways correspond to the sparse GT pathways.

7.1.2. Evaluation results: prior-based approaches

Many existing approaches have used the Grand Central Station dataset in their analysis. Recent approaches that have focused on identifying semantic regions and motion pathways have relied on the semantic prior in their analysis. These baseline approaches are the random field topic model (RFT) (Zhou et al., 2011) and understanding crowd collectivity (UCC) approach (Dehghan and Kalayeh, 2015). Both of these methods were evaluated on the Grand Central Station scene based on seven-gate and eight-gate annotations, respectively.

For a fair comparison, we applied the first baseline (RFT) (Zhou et al., 2011) using its default parameters with seven-gate annotation.[6] We focused on both quantitative and qualitative results for the identified motion pathways and their corresponding semantic regions. We extracted 42 GT pathways from the original GT trajectories of the walking route dataset (Yi et al., 2015) based on seven-gate annotation.

In our experiments, we used the recently released version of the Grand Central Station dataset (Yi et al., 2015), which is a 1-hour video. We sampled 50,000 frames throughout the entire video. Then, we collected more than 190,000 variable-length tracklets (several tracklets were hundreds of frames) and used them to estimate semantic regions by applying the RFT model (Zhou et al., 2011). For comparison, the total number of estimated semantic regions are presented in the first row in Table 2 as quantitative results. In the second row, we report our corresponding quantitative results that resulted from our prior-based clustering model. Note that our results were obtained based on our collected fixed-length (30 frames) tracklets. A sample of these motion pathways and their corresponding semantic paths, which were obtained from the RFT model, are shown, respectively, in the second and third columns of Fig. 12(b, c) compared with the GT pathways in column (a).

Note that the definition of semantic regions that was introduced in the RFT baseline method (Zhou et al., 2011) is quite similar to our motion pathway definition. Therefore, we considered their identified semantic regions as motion pathways and compared them with our motion pathways. Based on seven-gate annotation, our proposed approach successfully identified 40 out of 42 GT motion pathways that satisfy $Rc = 95.23\%$ in contrast to the RFT baseline method in Zhou et al. (2011), which identified only 30 semantic regions, as shown in the first and second rows in Table 2). We adopted all the semantic regions of RFT as true detection pathways, although multiple semantic regions were duplicated and spatially extended in a short range. However, our identified motion pathways were unique between each pair of source/sink gates. Furthermore, our identified pathways achieved good semantic interpretation by effectively capturing the global structures of the scene, and still respected our introduced notion of a motion pathway. Note that our intuition regarding the semantic interpretation of a pathway is that it is created at those regions in the crowded scene that are typically walked though by coherently moving

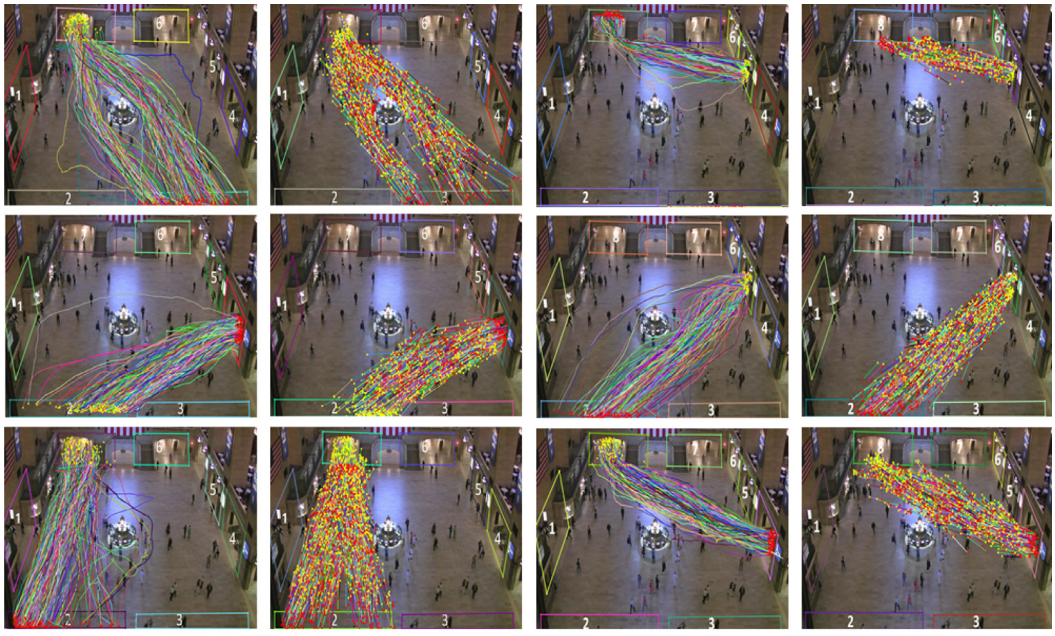
[6] We used the default parameters of the authors' source code, which set the total number of topics to 30, variable-length tracklets, and 800 iterations for the Gibbs sampling process (Zhou et al., 2011). https://github.com/metalbubble/RF_topic/tree/master/vc_RTM.

Table 2
Pathway/gate evaluation: the first and second rows present the quantitative results for our proposed method versus the RFT model (Zhou et al., 2011) using seven-gate annotation. Similarly, the third and fourth rows present the quantitative results for our proposed method versus the UCC method (Dehghan and Kalayeh, 2015) based on eight-gate annotation. We used the same metrics as those in Table 1.

Method	GT-P	GT-G	D-P	TD-P	FD-P	Pr (%)	Rc (%)
RFT (Zhou et al., 2011)	42	7	30	30	0	100	71.42
Our			40	**40**	0	100	**95.23**
UCC (Dehghan and Kalayeh, 2015)	64	8	54	54	0	100	84.37
Our	55		52	**52**	0	100	**94.54**

Table 3
Qualitative results for our proposed work and the RFT baseline (Zhou et al., 2011) versus the GT based on seven-gate annotation (columns labels:**source–sink**). The chosen pathways were the perceptually best five semantic regions that resulted from the RFT. The best matches are in red. The average metric in Eq. (17) was used as a matching measure. μ and σ are, respectively, the overall mean and standard deviation of matching scores for all identified pathways (40) and semantic regions (30) against GT.

Method	Score	6-3	7-3	6-2	7-4	2-4	μ	σ
RFT (Zhou et al., 2011)	spatial	0.5463	0.6773	0.3749	0.6263	0.8367		
	orientation	0.8920	0.9532	0.9558	0.8664	0.8364		
	avg.	**0.7192**	**0.8153**	**0.6654**	**0.7464**	**0.8366**	**0.6731**	**0.0937**
Our	spatial	0.8297	0.6571	0.6710	0.7071	0.8134		
	orientation	0.9610	0.9718	0.9545	0.9232	0.9479		
	avg.	**0.8954**	**0.8145**	**0.8128**	**0.8151**	**0.8807**	**0.7233**	**0.1124**



(a) (b) (c) (d)

Fig. 11. (a, c) GT pathways based on seven and eight-gate annotations, respectively, and (b, d) our identified pathways' correspondences.

pedestrians, where all pedestrians share the same motion orientation and speed. Additionally, these regions should spatially extend well from a particular source to a particular sink.

For qualitative results, we depended on our proposed pathway matching metric in Eq. (17) for evaluation. We computed both the spatial and motion orientation matching scores for all identified pathways and the corresponding GT pathways. We observe that our method still outperformed the RFT method and demonstrated state-of-the-art performance. In Table 3, the average PMSs (**avg**.) are reported for our obtained pathways and the semantic regions estimated from the RFT baseline (Zhou et al., 2011), both against the GT pathways. Moreover, for all our identified pathways and RFT semantic regions, we achieved overall mean and standard deviation of pathway matching scores of $\mu = 0.7233$, $\sigma = 0.1124$, respectively, versus $\mu = 0.6731$, $\sigma = 0.0937$ for the estimated semantic regions obtained from the RFT model.

Furthermore, our identified pathways also reflected the actual scene dynamics in terms of pathway densities. As shown in Fig. 10, the tracklet densities of the identified pathways relatively correspond to the densities of the corresponding GT pathways to a large extent. Fig. 12 shows samples of identified pathways obtained using our proposed method (column b) and their corresponding semantic regions that resulted from the RFT model (Zhou et al., 2011) (column c). We observe that our results demonstrated state-of-the-art performance in contrast to the baseline method, for which noisy tracklets, missing parts from discovered pathways, and pathways associated with multiple source/sink gates (column c) violated our adopted notion of motion pathways, and hence diminished the semantic interpretation of motion pathways.

Our method also outperformed the UCC baseline method (Dehghan and Kalayeh, 2015) in quantitative results based on eight-gate annotation. This is presented in the third and fourth rows of Table 2. Our method successfully identified 52 pathways out of 55 GT pathways in contrast to 54 pathways out of 64 GT for the UCC method. Note that the GT pathways count, GT-P, is different because the UCC model was applied on an older version of the Grand Central Station dataset (Zhou et al., 2011). This was before the release of the recent Grand Central

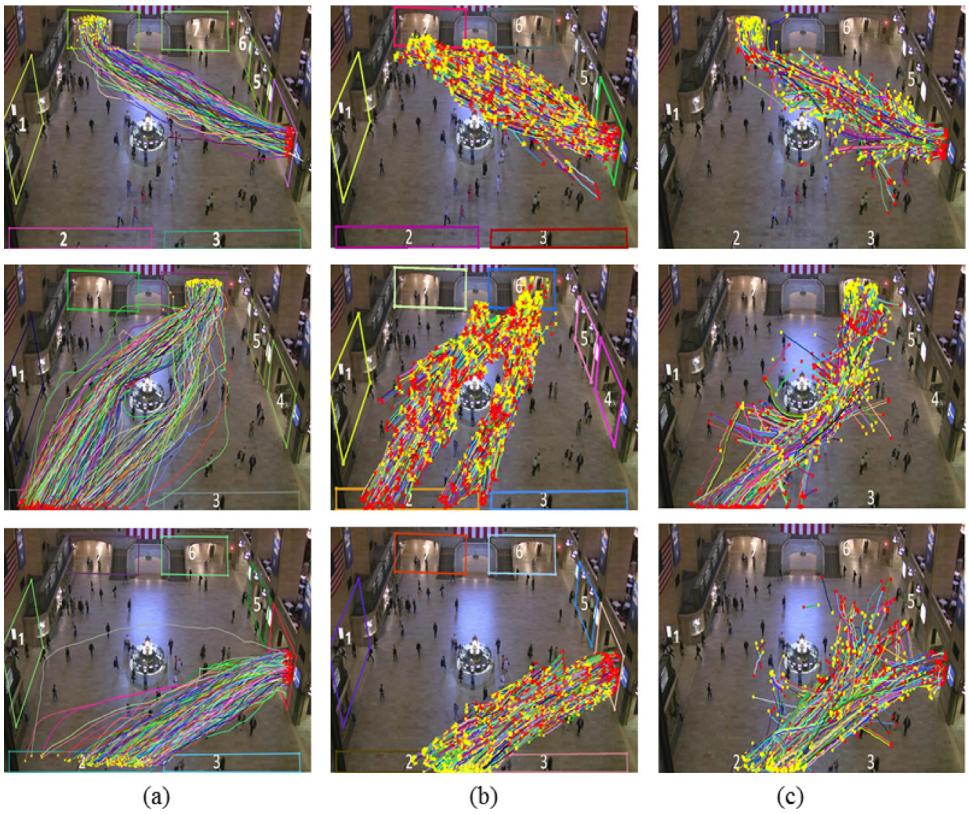


Fig. 12. (a) GT pathways, (b) pathways identified by our proposed method, and (c) corresponding semantic paths from Zhou et al. (2011). Both approaches incorporate source/sink priors in their clustering model.

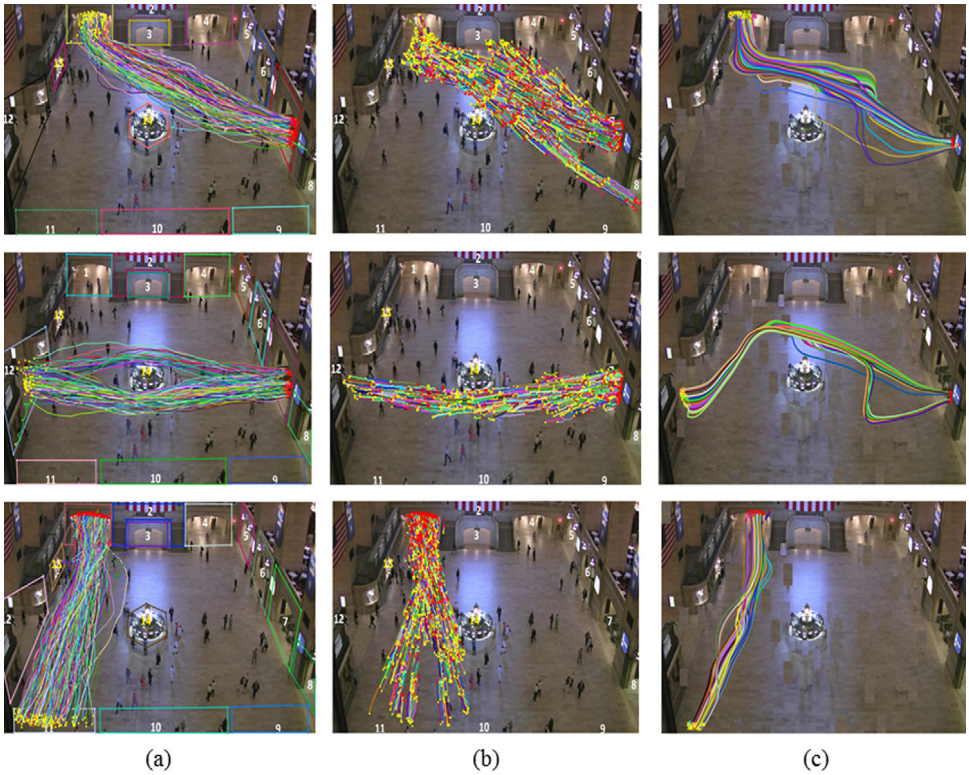


Fig. 13. (a) GT motion pathways, (b) pathways obtained using our no-prior approach, and (c) pathways that resulted from the MT method (Jodoin et al., 2013).

Station dataset and its walking route GT trajectories (Yi et al., 2015). Dehghan and Kalayeh (2015) argued that the total number of GT pathways throughout the Grand Central Station scene was 64 paths without any clear evidence regarding this GT number. However, we

Table 4
Pathway detection using our no-prior approach versus the MT approach for the Grand Central Station scene. TD-P/FD-P denote true detections/false detections for pathways, respectively.

Method	D-P	GT_G	TD-P	FD-P	Rc (%)
MT (Jodoin et al., 2013)	50	14	29	21	0.58
Ours (no-prior)			**48**	**2**	**0.96**

released the first GT pathways for the Grand Central Station scene based on several configurations of gate annotations. We approximately report the UCC method results for relative quantitative comparison. Our method outperformed the UCC method with $Rc = 94.54\%$ for our method versus $Rc = 84.37\%$ for the UCC method.

7.1.3. Evaluation results: no-prior approach

We also conducted further experiments to estimate the dominant motion paths from highly crowded scenes without considering any semantic prior information about the scene layout. We applied our approach (without a prior) on the Grand Central Station scene dataset and selected the baseline meta-tracking method (MT) (Jodoin et al., 2013) for evaluation.

To evaluate the detection of pathways, we sorted the resulting pathways from both our approach and the MT approach based on their richness. Richness was measured by the number of tracklets in our approach and number of trajectories (i.e., meta-tracks) in the MT approach. Then, we considered the richest 50 pathways of each approach and matched them with the GT pathways based on the proposed matching metric in Eq. (17). For each resulting pathway, the matching score was computed against the GT pathways and then thresholded such that if the score exceeded a given threshold (in our experiments $Thr = 0.5$), it was accepted as a detection, and the pathway that had the largest matching score was selected as the best match.

The results of this experiment are presented in Table 4, which shows that out of the richest 50 pathways, 48 were matched to true pathways using our proposed approach compared with only 29 using MT approach (Jodoin et al., 2013). Note that both the MT approach and our proposed model (without a prior) produced multiple pathways that corresponded to the same GT pathway. We report all the best matches in both approaches as true detections (TD-P) and the remainder as false detections (FD-P). However, by adopting only one match to a GT pathway as a true detection, we achieved TD-P = 30 versus 21 for the MT approach.

Some samples of our identified motion pathways and MT approach compared with GT pathways are shown in Fig. 13. Other sample results obtained from the crowd dataset released in Zhou et al. (2014) are shown in Fig. 14 to demonstrate the robustness of our approach for different crowd scenarios. Note that we visualized the identified pathways obtained using our no-prior approach for each scene (each row) and the first column shows the GT pathways.

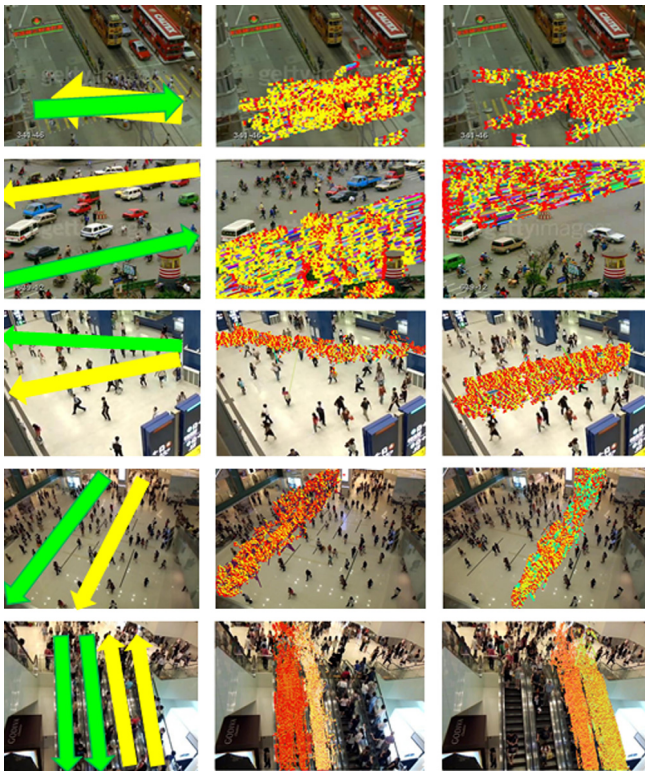


Fig. 14. Sample clustering results from different crowded scenes: the first column shows the GT pathways. The remaining columns show the identified pathways obtained using our approach without providing a semantic prior.

7.2. Anomaly detection evaluation

Detecting unusual actions/activities throughout the crowded scene has attracted a great deal of research interest recently. The problem is not only to detect if there is an abnormal action but also attempt to localize where and when the event occurred, or even identify how long it took. Visual scenes almost always contain normal behavior over the time evolution, and abnormal actions are rare cases. Thus, most of the proposed anomaly detection approaches depend on learning behaviors from labeled data or a corpus of unlabeled data in which most parts are normal. In anomaly detection approaches, several representations are considered for motion flow, such as optical flow, spatio-temporal patches, and tracklets (Li et al., 2015).

We introduce a simple tracklet representation that incorporates spatial, orientation, and speed information. We represent each tracklet

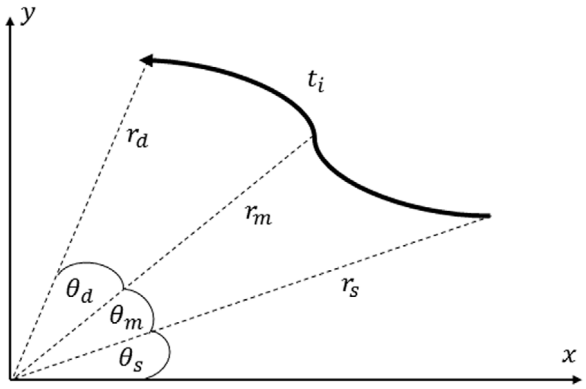


Fig. 15. Tracklet representation for anomaly detection.

by a **six-dimensional (6D)** feature vector. The pair of parameters (r, θ) represents each point on the tracklet in polar coordinates. These parameters express, respectively, the distance of the point from the frame origin and the amount of rotation required from the positive x-axis, as shown in Fig. 15. Let s, m, and d denote the start, middle, and end points of the tracklet, respectively. The proposed feature vector is a 6D vector $[r_s, r_m, r_d, \theta_s, \theta_m, \theta_d]$. Note that the θ angles are the rotation differences. This representation has proved to be reliable for the anomaly detection task for the discrimination between different types of tracklet abnormalities. In the following, we describe the evaluation results for both tracklet and frame levels of abnormalities.

7.2.1. Tracklet-level anomaly detection

Most crowd anomaly detection approaches depend on partitioning the crowd video into normal and abnormal clips. For the normal clips, we extracted the normal motion tracklets and then introduced

Table 5
Evaluation results for different anomalous scenes. AUC is the area under the curve, ACC is the maximum accuracy, and EER is the equal error rate. GC refers to the Grand Central Station scene. The three panic scenes of the UMN dataset (UMN, 2006) are shown in Fig. 16 from left to right, respectively.

	Data	AUC	ACC	EER		AUC	ACC	EER
Tracklet level	GC	0.97	0.94	0.07	Frame level		NA	
	UMN Scene 1	0.97	0.92	0.09		0.98	0.97	0.00
	UMN Scene 2	0.91	0.90	0.16		0.93	0.89	0.12
	UMN Scene 3	0.83	0.81	0.23		0.99	0.98	0.02

their corresponding feature vectors to train a Gaussian mixture model (GMM). The motion tracklets that were extracted from the abnormal clip were examined against the trained normalcy models. Every tracklet that deviated from the trained model with some certainty threshold was classified as abnormal. In our approach, we primarily adopted the existence of motion pathways of crowded scenes as a stipulation for analysis. Generally, any test tracklet that did not belong to any of the identified pathways was classified as abnormal (out of the active motion area).

To the best of our knowledge, the Grand Central Station dataset (Zhou et al., 2011; Yi et al., 2015) lacks anomalous data. Thus, defining actual anomalies in such a scene is a challenging task. Therefore, we sought to evaluate the robustness of our proposed representation for such a scene based on synthetic abnormal tracklets. We started by defining several types of anomalies that were expected to occur in such a scene. We synthetically generated 100 abnormal tracklets that represented the following five tracklet-based anomalous classes in the Grand Central Station scene: 1-*high speed*, 2-*low speed*, 3-*opposite direction*, 4-*gate blocking*, and 5-*out of the motion area*. For instance, a tracklet that had a higher speed than others in the same pathway was classified as abnormal, and a tracklet that may have passed through the same spatial extent of the pathway but in the opposite direction was also classified as abnormal.

From our output pathway results, we selected 20 pathways for anomaly detection evaluation. We started by training a normalcy GMM model for each pathway. All pathways' tracklets were adopted as normal tracklets and involved in the GMM training. For a new test tracklet, we computed all the GMM likelihoods for that tracklet. Finally, true/false detection rates were presented. As an experiment, we randomly selected 100 normal tracklets from the chosen detected pathways and added them to the 100 abnormal synthetic tracklets for testing. We used those 200 tracklets to test the trained normalcy GMM models based on the likelihood criterion. The evaluation results for tracklet-level abnormality are reported in Table 5 for the Grand Central Station scene (GC). Additionally, we report the evaluation results based on our tracklet representation for the three panic scenes (each clip contained individuals moving normally, followed by a panic motion) of the UMN dataset (UMN, 2006). The area under the curve (AUC) was adopted as an evaluation metric to compare against the GT (Table 5).

7.2.2. Frame-level anomaly detection

To evaluate our anomaly-based tracklet representation at the frame level, we used the available anomaly UMN dataset (UMN, 2006). It contains three panic scenes, as shown in Fig. 16. We used the same GT annotation of abnormal frames as those of Mehran et al. (2009). Based on the trained GMM models from the tracklet level, if a tracklet stopped at a particular frame and was classified as abnormal, then the frame was also classified as abnormal. The number of abnormal frames was examined against the GT frames, and the results are reported in Table 5. The frame-level anomaly evaluation of the UMN dataset using our approach accomplished comparable results (our average AUC = 0.97) compared with the simple optical flow and SFM (Mehran et al., 2009) (average AUC = 0.84 and 0.96, respectively). Note that we did not consider the frame-level anomaly evaluation for the Grand Central Station scene in our experiments because there is no frame-level anomaly GT for it yet.

7.3. Discussion: Automatic tolerance setting

In our experiments, we empirically selected the distance tolerances, $[\sigma_{\delta min}, \sigma_{\delta max}]$, and angle tolerances, $[\sigma_{\theta min}, \sigma_{\theta max}]$, in Eq. (5). It is, of course, desirable to automatically estimate the appropriate values for these tolerance limits. However, these limits depend on how much the moving objects in the same motion pathway tend to deviate from one another in terms of space and orientation when they are moving in the same stage of a pathway and when they are moving in two consecutive stages of a pathway. This requires prior knowledge about the nature of the scene and its motion dynamics. We think that it is possible to automatically estimate these limits from the GT trajectories or pathways if either of them is available, as well as from the scene layout. This topic requires deeper investigation and will be considered in our future work.

8. Conclusion

In this paper, we proposed a robust approach for identifying motion pathways in crowded scenes. We estimated pairwise similarities between motion tracklets based on a novel similarity measure inspired by line geometry. This measure effectively captured the spatial and orientation similarity between tracklets. We adapted the DD-CRP nonparametric clustering model based on the estimated similarities. This model was then extended by introducing the spatial extents of the scene gates as a high-level semantic prior, where the source and sink probability distributions were estimated for each tracklet and incorporated in the DD-CRP cluster likelihood computation. Moreover, we proposed a robust pathway matching metric for qualitative evaluation. A new comprehensive GT pathway and gate annotation for the Grand Central Station scene was provided for robust evaluation. The experimental results on the challenging Grand Central Station dataset demonstrated the superiority of our clustering approach for identifying more compact and the best spatially extended motion pathways between their associated gates. Furthermore, we examined the identified pathways against the anomalous activities on both the tracklet and frame levels, with competitive quantitative performance. In future work, we plan to study the identification of crowd-based events, such as congestion at gates and the influence of such events on the behaviors of pedestrians in highly crowded scenes. Moreover, we seek to study the relationship between crowd analysis and crowd simulation, and aim for a unified framework that helps to bridge the gap between these two interrelated tasks.

Acknowledgments

This work was supported by the Information Technology Industry Development Agency under the ITAC Program [grant number PRP2015.R19.4-Automatic Crowded Scene Analysis and Anomaly Detection from Video Surveillance Cameras]. It was also supported by the Ministry of Higher Education of Egypt through a Ph.D. scholarship. Part of this work was accomplished during a research visit to Yagi Laboratory, Department of Intelligent Media, ISIR, Osaka University, Japan. We thank Shuai Yi, CHUK, for providing us with the 1-hour video of the Grand Central Station scene. We thank Maxine Garcia, Ph.D. from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

Fig. 16. Three panic scenarios of the UMN dataset (UMN, 2006): the first row shows the normal behaviors and the second row shows the abnormal behaviors.

References

Alahi, A., Ramanathan, V., Fei-Fei, L., 2014. Socially-aware large-scale crowd forecasting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2203–2210.

Aldous, D., 1985. Exchangeability and related topics. École d'Été de Probabilités de Saint-Flour XIII—1983, 1–198.

Ali, S., Shah, M., 2008. Floor fields for tracking in high density crowd scenes. In: Computer Vision–ECCV 2008. Springer, pp. 1–14.

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist. 1152–1174.

Atev, S., Masoud, O., Papanikolopoulos, N., 2006. Learning traffic patterns at intersections by spectral clustering of motion trajectories. In: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on. IEEE, pp. 4851–4856.

Baldassano, C., Beck, D.M., Fei-Fei, L., 2015. Parcellating connectivity in spatial maps. PeerJ 3, e784.

Blei, D.M., Frazier, P.I., 2011. Distance dependent chinese restaurant processes. J. Mach. Learn. Res. 12, 2461–2488.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Chen, K., Kämäräinen, J.-K., 2016. Pedestrian density analysis in public scenes with spatiotemporal tensor features. IEEE Trans. Intell. Transp. Syst. 17 (7), 1968–1977.

Chen, X., Qin, Z., An, L., Bhanu, B., 2014. An online learned elementary grouping model for multi-target tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1242–1249.

Chen, M., Wang, Q., Li, X., 2017. Patch-based topic model for group detection. Sci. China Inf. Sci. 60 (11), 113101.

Cheriyadat, A.M., Radke, R.J., 2008. Detecting dominant motions in dense crowds. IEEE J. Sel. Top. Sign. Proces. 2 (4), 568–581.

Chongjing, W., Xu, Z., Yi, Z., Yuncai, L., 2013. Analyzing motion patterns in crowded scenes via automatic tracklets clustering. China Commun. 10 (4), 144–154.

Dehghan, A., Kalayeh, M.M., 2015. Understanding crowd collectivity: A meta-tracking approach. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1–9.

Etemadi, A., Schmidt, J.-P., Matas, G., Illingworth, J., Kittler, J., 1991. Low-level grouping of straight line segments. In: BMVC91. Springer, pp. 118–126.

Ghosh, S., Ungureanu, A.B., Sudderth, E.B., Blei, D.M., 2011. Spatial distance dependent chinese restaurant processes for image segmentation. Adv. Neural Inf. Process. Syst. 1476–1484.

Hassanein, A.S., Hussein, M.E., Gomaa, W., 2016. Semantic analysis of crowded scenes based on non-parametric tracklet clustering. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. AAAI Press, pp. 3389–3395.

Helbing, D., Johansson, A., Al-Abideen, H.Z., 2007. Dynamics of crowd disasters: An empirical study. Phys. Rev. E 75 (4), 046109.

Helbing, D., Molnar, P., 1995. Social force model for pedestrian dynamics. Phys. Rev. E 51 (5), 4282.

Hospedales, T., Gong, S., Xiang, T., 2009. A markov clustering topic model for mining behaviour in video. In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE, pp. 1165–1172.

Jodoin, P.-M., Benezeth, Y., Wang, Y., 2013. Meta-tracking for video scene understanding. In: Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on. IEEE, pp. 1–6.

Kratz, L., Nishino, K., 2012. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell. 34 (5), 987–1002.

Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S., 2015. Crowded scene analysis: A survey. IEEE Trans. Circuits Syst. Video Technol. 25 (3), 367–386.

Li, W., Mahadevan, V., Vasconcelos, N., 2014. Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell. 36 (1), 18–32.

Liang, P., Petrov, S., Jordan, M.I., Klein, D., 2007. The infinite PCFG using hierarchical Dirichlet processes. In: EMNLP-CoNLL. pp. 688–697.

MacKay, D.J., 2003. Information Theory, Inference and Learning Algorithms. Cambridge university press.

Mehran, R., Moore, B.E., Shah, M., 2010. A streakline representation of flow in crowded scenes. In: European Conference on Computer Vision. Springer, pp. 439–452.

Mehran, R., Oyama, A., Shah, M., 2009. Abnormal crowd behavior detection using social force model. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, pp. 935–942.

Pele, O., Werman, M., 2010. The quadratic-chi histogram distance family. In: European Conference on Computer Vision. Springer, pp. 749–762.

Perera, A.A., Srinivas, C., Hoogs, A., Brooksby, G., Hu, W., 2006. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1. IEEE, pp. 666–673.

Pirsiavash, H., Ramanan, D., Fowlkes, C.C., 2011. Globally-optimal greedy algorithms for tracking a variable number of objects. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, pp. 1201–1208.

Raghavendra, R., Del Bue, A., Cristani, M., Murino, V., 2011. Abnormal crowd behavior detection by social force optimization. In: International Workshop on Human Behavior Understanding. Springer, pp. 134–145.

Rodriguez, M., Ali, S., Kanade, T., 2009. Tracking in unstructured crowded scenes. In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE, pp. 1389–1396.

Saleemi, I., Hartung, L., Shah, M., 2010. Scene understanding by statistical modeling of motion patterns. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp. 2069–2076.

Shannon, C., 1948. A mathematical theory of communication, bell system technical journal 27: 379-423 and 623–656. Math. Rev. (MathSciNet): MR10, 133e.

Shao, J., Loy, C.C., Wang, X., 2014. Scene-independent group profiling in crowd. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, pp. 2227–2234.

Solmaz, B., Moore, B.E., Shah, M., 2012. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE Trans. Pattern Anal. Mach. Intell. 34 (10), 2064–2070.

Su, H., Yang, H., Zheng, S., Fan, Y., Wei, S., 2013. The large-scale crowd behavior perception based on spatio-temporal viscous fluid field. IEEE Trans. Inf. Forensics Secur. 8 (10), 1575–1589.

Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S., 2008. Describing visual scenes using transformed objects and parts. Int. J. Comput. Vis. 77 (1–3), 291–330.

Teh, Y.W., 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 985–992.

Tomasi, C., Kanade, T., 1991. Detection and tracking of point features. School of Computer Science, Carnegie Mellon Univ. Pittsburgh.

Topkaya, I.S., Erdogan, H., Porikli, F., 2015. Tracklet clustering for robust multiple object tracking using distance dependent chinese restaurant processes. Signal, Image Video Process. 1–8.

Treuille, A., Cooper, S., Popović, Z., 2006. Continuum crowds. ACM Trans. Graph. 25 (3), 1160–1168.

UMN, 2006. Unusual crowd activity dataset of University of Minnesota. http://mha.cs.umn.edu/movies/crowdactivity-all.avi. Accessed 30 September 2010.

Wang, X., Ma, K.T., Ng, G.-W., Grimson, W.E.L., 2008. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.

Wang, X., Yang, X., He, X., Teng, Q., Gao, M., 2014. A high accuracy flow segmentation method in crowded scenes based on streakline. Optik-Int. J. Light Electron Optics 125 (3), 924–929.

Yi, S., Li, H., Wang, X., 2015. Understanding pedestrian behaviors from stationary crowd groups. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3488–3496.

Zhang, T., Lu, H., Li, S.Z., 2009. Learning semantic scene models by object classification and trajectory clustering. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, pp. 1940–1947.

Zhao, J., Xu, Y., Yang, X., Yan, Q., 2011. Crowd instability analysis using velocity-field based social force model. In: Visual Communications and Image Processing (VCIP), 2011 IEEE. IEEE, pp. 1–4.

Zhou, B., Tang, X., Wang, X., 2012. Coherent filtering: Detecting coherent motions from crowd clutters. In: Computer Vision–ECCV 2012. Springer, pp. 857–871.

Zhou, B., Tang, X., Wang, X., 2015. Learning collective crowd behaviors with dynamic pedestrian-agents. Int. J. Comput. Vis. 111 (1), 50–68.

Zhou, B., Tang, X., Zhang, H., Wang, X., 2014. Measuring crowd collectiveness. IEEE Trans. Pattern Anal. Mach. Intell. 36 (8), 1586–1599.

Zhou, B., Wang, X., Tang, X., 2011. Random field topic model for semantic region analysis in crowded scenes from tracklets. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, pp. 3441–3448.

Zou, Y., Zhao, X., Liu, Y., 2015. Detect coherent motions in crowd scenes based on tracklets association. In: Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, pp. 4456–4460.