Semantic Analysis of Crowded Scenes Based on Non-Parametric Tracklet Clustering

Allam S. Hassanein[1], **Mohamed E. Hussein**[1,2], **Walid Gomaa**[1,2]
[1]Cyber Physical Systems Lab.,
Egypt-Japan University of Science and Technology, Alexandria, Egypt
[2] Faculty of Engineering, Alexandria University, Alexandria, Egypt
{allam.shehata,mohamed.e.hussein,walid.gomaa}@ejust.edu.eg

Abstract

In this paper we address the problem of semantic analysis of structured/unstructured crowded video scenes. Our proposed approach relies on tracklets for motion representation. Each extracted tracklet is abstracted as a *directed line segment*, and a novel tracklet similarity measure is formulated based on line geometry. For analysis, we apply non-parametric clustering on the extracted tracklets. Particularly, we adapt the Distance Dependent Chinese Restaurant Process (DD-CRP) to leverage the computed similarities between pairs of tracklets, which ensures the spatial coherence among tracklets in the same cluster. By analyzing the clustering results, we can identify semantic regions in the scene, particularly, the common pathways and their sources/sinks, without any prior information about the scene layout. Qualitative and quantitative experimental evaluation on multiple crowded scenes datasets, principally, the challenging New York Grand Central Station video, demonstrate the state of the art performance of our method.

1 Introduction

Due to the increase of the population and diversity of human's activities and behaviors, crowded scenes have been more frequent in the real world than ever. Adding to this the escalating world-wide concerns about security, automatic crowded scene analysis has become one of the most attractive topics in computer vision and pattern recognition. The major goal of such research is extracting some kind of information from the scene about the moving objects' behaviors in order to serve multiple applications, such as, visual surveillance, crowd management, safety analysis of public places or sports arenas, etc.

Two main analysis levels for crowded scenes are introduced: macroscopic and microscopic [Li *et al.*, 2015]. At the macroscopic level , we deal with crowd motions as global motion pattern(s) of a mass of objects, without being concerned with the movements of the individual objects [Hu *et al.*, 2008]. On the other hand, the microscopic level is concerned with the movements of individual moving objects as well as the interactions among them [Zhou *et al.*, 2012].

To serve the aforementioned levels of analysis, two major approaches for the computational modeling of crowd behavior are introduced. The first is the continuum-based approach (holistic), which works better at the macroscopic level for medium and high density crowds [Ali and Shah, 2007]. Such kind of techniques usually try to obtain global information about the scene regardless of any local activities, such as the identification of global active regions which have high traffic as well as the main directions of flows. The second approach is agent-based, which is more suitable for low-density crowds at the microscopic level, where the movement of each individual moving object is taken into account [Zhou *et al.*, 2012; Zhao *et al.*, 2011].

Both continuum-based and agent-based approaches rely on some form of motion representation in order to conduct their analyses. In this regard, three main levels of motion representation have been introduced. The first is flow-based representation, which extracts motion features at the pixel level [Wang *et al.*, 2014]. The second is local spatio-temporal representation, which represents the scene in terms of local information extracted from 2D patches [Kratz and Nishino, 2012]. The third level is the trajectory/tracklet representation, which represents motion information at a higher level dealing with individual tracks as a basic unit [Zhou *et al.*, 2011; Topkaya *et al.*, 2015].

The trajectories/tracklets representation is more semantically-rich than the other representations because it incorporates information about a semantically meaningful moving entity (e.g. a feature point or an object) for a period of time. A *tracklet* is defined as a fragment of a trajectory obtained by the tracker within a short period of time. It may terminate when occlusions or scene clutters occur [Li *et al.*, 2015]. Thus, tracklets are more conservative and less likely to drift compared to complete trajectories.

In this paper, we introduce a new macroscopic-level approach for crowded scenes analysis that relies on tracklets as the basic motion representation. Particularly, we are interested in grouping motion patterns in a way that enables the discovery of the underlying scene structure, namely, the common pathways of moving objects and the sources/sinks of the scene, which we collectively call as *semantic regions*. On doing so, we do not assume prior information about the numbers or the spatial extents of such scene structural elements.

Our proposed approach first extracts tracklets of detected interest points in the foreground (motion) areas of the scene. Then, tracklets are clustered hierarchically over two levels such that the resulting clusters correspond to common pathways in the scene. To accomplish this goal, a novel and flexible tracklet similarity measure, which is based on line geometry, is introduced. The discovered pathways are then analyzed to find the common sources and sinks of the scene. In order to achieve these goals, we adopted a non-parametric clustering algorithm that is based on the Distance Dependent Chinese Restaurant Processes (DD-CRP) [Blei and Frazier, 2011].

The main contributions of this work can be summarized as follows: (i) a novel tracklet similarity measure based on line geometry, (ii) an adaptation of DD-CRP to the problem of grouping tracklets into common pathways using a two-level hierarchical clustering, (iii) a method for discovering the scene structure and its sources and sinks from the resulting clustering, and (iv) a novel evaluation framework for the resulting scene analysis that takes into account both the detected scene structural elements and their geometric extents.

The rest of the paper is organized as follows. Section 2 outlines the most related work. In Section 3, a detailed explanation of our proposed tracklet similarity measure is provided. The adaptation of DD-CRP model to serve our tracklet clustering problem, and the discovery of the scene's semantic regions are provided in Section 4. Experiments are included in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

Similar to our approach, many crowded scene analysis approaches in the literature are based on tracklets. In [Zhou *et al.*, 2011], a Random Field Topic (RFT) model is proposed to learn semantic region analysis in crowded scenes from tracklets. The RFT model is an advancement of the existing Latent Dirichlet Allocation (LDA) topic model [Blei *et al.*, 2003], where a Markov Random Field (MRF) is integrated as a prior to impose spatial and temporal coherence between tracklets during the inference process. To improve the inference of semantic regions and clustering of tracklets, sources and sinks are included as a high-level semantic prior. Our approach also identifies semantic regions; however, it does not require sources and sinks to be a priori known. In [Wang *et al.*, 2013], an approach is proposed to analyze motion patterns in dynamical crowded scenes based on hybrid generative-discriminative feature maps, which are in turn based on the collected tracklets. Automatic hierarchical clustering algorithm is used to analyze motion patterns. These motion patterns are analogous to the common pathways identified by our work. However, our approach is simpler and also produces sources and sinks.

Tracklets are frequently used as building blocks to enhance tracking in crowded scenes. For instance, in [Zhao and Medioni, 2011], an unsupervised manifold learning framework is proposed to infer motion patterns in videos. Tracklet points are embedded into a 3D space (x,y,θ) that represents the image space and motion direction. In this space, points automatically form intrinsic manifold structures, each of which corresponds to a motion pattern. The extracted motion patterns can be used as a prior to improve the performance of object tracking techniques. Also, in [Kuo *et al.*, 2010], an algorithm is proposed for Online Learning of Discriminative Appearance (OLDA) models for different targets in crowded scenes based on collected tracklets. Spatial-temporal relations between tracklets in a time window are examined to discriminate between targets. OLDA models are integrated into a hierarchical association framework to improve the tracking system's accuracy.

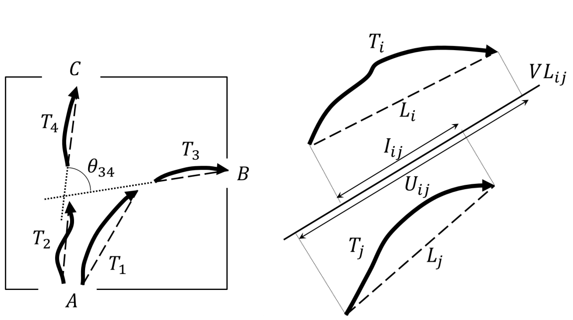
DD-CRPs are adopted in language modeling, computer vision problems, and mixture modeling for clustering applications. For example, DD-CRP is examined in [Ghosh *et al.*, 2011] in the spatial domain for image segmentation, where a novel hierarchical extension, better suited for efficient image segmentation, is proposed. A tracklets-clustering approach based on DD-CRP is proposed in [Topkaya *et al.*, 2015] for the purpose of tracking enhancement. In this work, two-level robust object tracking is employed to generate tracklets, which are then clustered based on their color, spatial, and temporal similarities. In our work, we adopted a similar model; however, the application is different. In their work, a cluster is supposed to contain a single whole trajectory of one object. In contrast, in our work, a cluster is supposed to contain a group of tracklets in a common pathway. Therefore, the similarity measure and cluster probability functions are totally different.

3 The Tracklet Similarity Measure

The purpose of clustering tracklets in our approach is to identify semantic regions in the scene, which are, namely, the common pathways, the sources, and the sinks. In this section, we focus on the tracklet similarity measure, variants of which are used in multiple levels of non-parametric clustering.

We would like tracklets to be clustered together when they belong to the same common pathway. For two tracklets to belong to a common pathway, they have to belong to a single object, or two objects that are originating from the same source and moving towards the same sink. In this case, the two tracklets are expected to bear similarity to one another in terms of their spatial layouts and their global orientations. However, encoding this similarity in a single measure is not trivial due to the many cases that can be encountered in practice.

Figure 1a shows a hypothetical scene having one source (A) and two sinks (B and C), with four overlaid tracklets. Consider the two tracklets T_1 and T_2. Although both of them originate from the same source and are spatially close to each other, perceptually, they do not seem to belong to a common pathway. This can be interpreted by inspecting the geometric relationship between the two tracklets: If they belonged to the same common pathway, they would have been *in the same stage* (the beginning here) of that pathway, which means they should have been almost parallel. However, because of their divergence in orientation, they are not perceived to be in 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. Nevertheless, perceptually, tracklet T_3 seems to be a continuation of T_1, i.e. the two tracklets can be in the same pathway but in *two different stages*.



(a) (b)

Figure 1: (a) A hypothetical scene with one source (A), two sinks (B and C), and four tracklets ($T_1..T_4$). The directed line segment associated with each tracklet is shown in dashed style. θ_{34} is the estimated angle between T_3 and T_4. (b) The 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 way we interpret the geometric relationship between two tracklets depends on the degree to which they are perceived to be in the same stage of a common pathway. In our approach, we estimate this by the degree of overlap between the two tracklets. The intuition is that the higher the overlap between two tracklets, the more likely they belong to the same stage of a common pathway, and vice versa. Our similarity measure incorporates both the *spatial* and *orientation* similarities between tracklets while taking into account the *overlap* between them. Now, we will explain each of these components.

Spatial similarity between two tracklets is estimated using two different distance functions: the Hausdorff distance and the minimum point-to-point distance. Let $T_i = (p_{i1}, p_{i2}, ..., p_{in})$, and $T_j = (p_{j1}, p_{j2}, ..., p_{jn})$ be two different 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. The Hausdorff distance $d_H(T_i, T_j)$ between the two tracklets can be computed as

$$d_H(T_i, T_j) = \max\left\{\Delta(T_i, T_j), \Delta(T_j, T_i)\right\}\ , \quad (1)$$

$$\Delta(T_i, T_j) = \max_{p_{ik} \in T_i} \min_{p_{jl} \in T_j} \mathrm{d}(p_{ik}, p_{jl})\,, \quad (2)$$

where $\mathrm{d}(p_{ik}, p_{jl})$ is the Euclidean distance between the k^{th} point of T_i and the l^{th} point of T_j. On the other hand, the minimum point-to-point distance can be expressed as

$$d_M(T_i, T_j) = \min_{p_{ik} \in T_i, p_{jl} \in T_j} \mathrm{d}(p_{ik}, p_{jl})\,, \quad (3)$$

In the following, we refer to the distance between two tracklets T_i and T_j by δ_{ij}, regardless of the type. In the following section, we will explain when we apply each type.

To estimate the orientation similarity between a pair of tracklets, we approximate each tracklet as a *directed line segment* that extends from its starting to its ending points, as shown in Figure 1a. Note that since tracklets are typically constructed over short time periods, approximating them by directed line segments should be acceptable for 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

$$Sim(T_i, T_j) = e^{-\left(\frac{\theta_{ij}}{\sigma_{\theta ij}}\right)^2} e^{-\left(\frac{\delta_{ij}}{\sigma_{\delta ij}}\right)^2}\ , \quad (4)$$

where the two variables $\sigma_{\theta ij}$ and $\sigma_{\delta ij}$ represent the tolerance values in the orientation and spatial dimensions. The higher the tolerance value, the less sensitive the similarity function to the changes in the associated variable. The similarity measure takes values in the range $[0, 1]$.

As the notation in Equation 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.

$$\sigma_{\theta ij} = \sigma_{\theta max} + O_{ij} \cdot (\sigma_{\theta min} - \sigma_{\theta max})\,, \quad (5)$$

$$\sigma_{\delta ij} = \sigma_{\delta min} + O_{ij} \cdot (\sigma_{\delta max} - \sigma_{\delta min})\,. \quad (6)$$

where O_{ij} indicates the degree of overlap between the two tracklets T_i and T_j, which is takes a value in the interval $[0, 1]$ (as explained below). Each tolerance value is chosen from an interval, i.e. $\sigma_{\theta ij} \in [\sigma_{\theta min}, \sigma_{\theta max}]$ and $\sigma_{\delta ij} \in [\sigma_{\delta min}, \sigma_{\delta max}]$. We linearly choose a value in the interval based on the overlap between the two tracklets such that the higher the overlap between them, the more tolerance we give to the spatial dissimilarity and less tolerance we give to orientation dissimilarity.

To estimate the degree of overlap between two tracklets, we resort again to 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 adopted the idea of the virtual line from [Etemadi *et al.*, 1991]. The computation is illustrated in Figure 1b.

4 Tracklet Clustering and Semantic Scene Analysis

In this section, we first provide a brief background about DD-CRP. Then, we introduce the adaptation of DD-CRP to our tracklet clustering problem. Finally, we explain how the semantic regions are discovered from the resulting clustering.

4.1 Distance Dependent CRP

The main issue in high dimensional data clustering problems is finding a flexible clustering algorithm. One of the recent valuable models is the *Dirichlet Process Mixture Models (DPMMs).* DPMMs provide an efficient way to model a set of data points O as a mixture of unknown number of distributions sampled from the same base distribution G_0 [Antoniak, 1974]. The clustering problem in a DPMM is represented as a distribution over an infinite number of mixture components (i.e., clusters). One of DPMMs representations is the *Chinese Restaurant Process (CRP).* In the CRP analogy, a sequence of customers are going to be seated at an infinite number of tables in a restaurant. The first comer will gain a probability one to sit at a given table. Any subsequent customer sits at a previously occupied table with probability proportional to the number of people already seated at the table, and sits at a new table with probability proportional to a scaling parameter α. Based on the Gibbs sampling method, CRP iteratively sample every table assignment z_i from the following probability:

$$P(z_i = j|z_{-i}, \alpha) \propto \begin{cases} N_j & j \leq K \\ \alpha & j = K+1 \end{cases} \tag{7}$$

where z_i is the table assignment of the *ith* customer, N_j is the number of customers sitting at table *j*, tables $1, \ldots, K$ are occupied, and z_{-i} is all table assignment except for the assignment of customer *i*.

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) which reflect dependencies among them, and violate the exchangeability property of the basic Dirichlet process. So, the DD-CRP model is developed in order to handle these dependencies [Blei and Frazier, 2011]. DD-CRP model represents the data partitioning through customer assignments rather than table assignments, and the customer's assignments depend only on the distance among customers. Furthermore, customers are assigned to tables by considering customers reachability to each other through their assignments. According to this analogy, customer assignments will be conditioned on the distances between customers and drawn independently according to the following scheme.

$$P(c_i = j|D, \alpha) \propto \begin{cases} f(d_{ij}) & i \neq j \\ \alpha & i = j \end{cases}, \tag{8}$$

where d_{ij} is the distance between customers i and j, D denotes the distance matrix between all customers, α is the scaling parameter, and f is the decay function (for decay function details see [Blei and Frazier, 2011]). Additionally, the posterior inference for DD-CRP, based on Gibbs sampling, is implemented by iteratively sampling the customer assignments from the conditional distribution for each new customer (latent one) given the previous already seated customers c_{-i} and all customers O. So, the posterior probability looks like the following:

$$\begin{aligned} P(c_i^{new}|c_{-i}, &O, D, f, \alpha, G_0) \propto \\ &P(c_i|D, \alpha) \times P(O|W(c_{-i} \cup c_i^{new}), G_0) \end{aligned} \tag{9}$$

Note that $P(c_i|D, \alpha)$ represents the DD-CRP prior term from Equation (8) and $P(O|W(c_{-i} \cup c_i^{new}), G_0)$ is the likelihood of the data points under the partitioning given by $W(c_{-i} \cup c_i^{new})$.

4.2 Tracklet Clustering Based on DD-CRP

We adapted DD-CRP by using our similarity measure between tracklets rather than the distance between them. Within the DD-CRP clustering framework, tracklets correspond to observations, whereas pathways are the output clusters. Let S denote the similarity matrix among all tracklets, so Equation (8) is modified as follows.

$$P(c_i = j|S, \alpha) \propto \begin{cases} s_{ij} \ \ if \ \ i \neq j \\ \\ \alpha \ \ if \ \ i = j \end{cases} \tag{10}$$

where s_{ij} is the pairwise similarity between tracklets i and j. We represent the likelihood term as the factorization of maximal pairwise similarities among a group of directly/indirectly connected tracklets (i.e. cluster).

$$P(t_{1:N}|G_0) = P(t_1|G_0) \prod_{n=2}^{N} \max_{j=1..n-1} P(t_n|t_j, G_0) \tag{11}$$

Note that $P(t_n|t_j, G_0)$ is chosen to be proportional to the pairwise similarity between tracklet t_n and tracklet t_j.

4.3 Identifying Semantic Regions

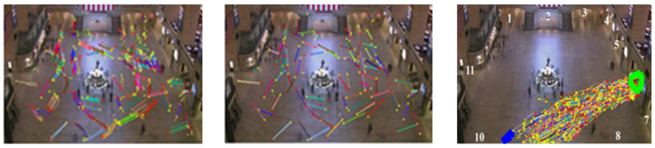
Semantic regions (i.e. pathways) correspond to spatial regions of the scene that have high degrees of local similarities. We define the pathway as a series of spatially coherent linked groups of tracklets. Each pathway has its preferred source and sink, and the motion flow is from the source to the sink. The collected tracklets are clustered hierarchically over two levels. In both levels, the adapted DD-CRP is deployed.

At the first level of clustering, the collected tracklets are clustered using DD-CRP based on a parallelism criteria, which tries to group only parallel tracklets together. This criteria is incorporated into our clustering framework by adjusting the limits on the tolerance values in the similarity measure $\sigma_{\theta min/max}$, and $\sigma_{\delta min/max}$, and using the Hausdorff distance function (Equation 1). As an output of this level, each resulting cluster is represented by a single directed representative line segment which is obtained from the associated cluster's tracklets. The representative line segment of a cluster of tracklets has the average orientation of tracklets in the clusters and passes through the center of mass of the union of all tracklets points. Its terminal points are identified by projecting all tracklets on it and taking the extreme projected points. Figure 2a shows the clusters of parallel tracklets and Figure 2b shows the associated representative line segments for the clusters correspondances.

At the second level of clustering, all of the resulting representative lines are clustered based on DD-CRP again using the same similarity function. However, in this case, the similarity function is adjusted to group line segments continuing after one another, by adjusting the limits of the tolerance again and by deploying the minimum point-to-point distance function (Equation 3). The output clusters from this level correspond to the common pathways in the scene as shown in Figure 2c.

We choose the Hausdorff distance for the spatial similarity in the first level of clustering because it captures the separation between parallel tracklets whether they are parallel or intersecting, while the minimum point-to-point distance becomes zero if the two tracklets intersect. On the other hand, for the second level clustering, the Hausdorff distance can become too large for tracklets continuing one another.

Once a pathway is identified, it is represented again as a single directed representative line segment, estimated from all corresponding pathway's tracklets (Figure 2c). The two terminals points of each such line segment represent the detected pathway's source and sink regions, respectively. To determine the spatial extents of a pathway's source and sink regions, we consider the convex hulls of the tracklet terminal points lying within a small distance[1] from the representative line segment's terminal points. The x-y coordinates of all the points within the convex hulls are then clustered using DD-CRP to identify the scene's sources and sinks (gates).



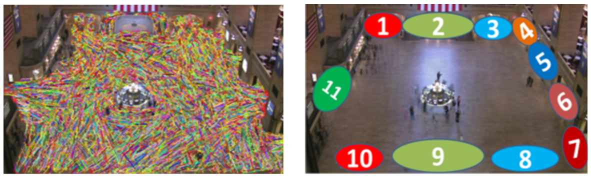
(a) (b) (c)

Figure 2: (a) Coherent parallel tracklets are clustered together. (b) Representative directed line segments for obtained clusters. (c) Detected pathway with associated source/sink regions overlaid in blue/green color, respectively.

5 Experimental Evaluation

In our implementation, to obtain tracklets in a given crowded scene, we first detect interest points using the minimum eigen features method [Shi and Tomasi, 1994] in foreground regions, which are identified using background subtraction via Gaussian Mixture Models [Stauffer and Grimson, 1999], learned from the first five frames. Then, the detected points are tracked using the standard Kanade-Locus-Tomasi (KLT) tracker [Tomasi and Kanade, 1991].

Experiments are conducted on multiple datasets. However, most of our analysis is performed on the challenging New York's Grand Central station video [Zhou *et al.*, 2011], which is a 33-minute video with 540×960 resolution and a frame rate of 25 FPS. More than 20,000 tracklets are extracted from this scene (Figure 3a). All tracklets are stopped, collected, and tracking is restarted every 25 frames, which makes all our tracklets having the same fixed length. In the following, we first present our experiments on the Grand Central Station's scene, then, on other datasets. More detailed results and resources associated with this work can be found online[2].



(a) (b)

Figure 3: (a) Collected tracklets in the Grand Central station's scene. (b) Scene gate layout according to the floor plan of the station.

5.1 The Grand Central Station Scene

We compare our approach against the Meta-Tracking (MT) approach [Jodoin *et al.*, 2013] in terms of pathway detection, pathway spatial layout coverage, and source/sink detection.

Method	TD-P	FD-P	TD-G	FD-G
MT [Jodoin *et al.*, 2013]	10	30	6	7
Proposed	**14**	**26**	**9**	**15**

Table 1: Pathway and Gate detection in our approach vs. the MT approach on the Grand Central scene. (TD-P/FD-P) are True Detections/False Detections for Pathways. (TD-G/FD-G) are the same for Gates.

Common Pathway Detection

To quantitatively evaluate the detection of common pathways, we used a recently released large scale annotation for the Grand Central video dataset [Yi *et al.*, 2015]. In this annotation, all pedestrians are manually tracked and the complete path for each pedestrian is labeled from the time of entering to the time of leaving the scene. We manually labeled eleven gates (pathway sources/sinks) in the scene, which are shown in Figure 3b. For each pair of gates, we extracted from the ground truth (GT) all pedestrian trajectories originating from the first and terminating in the second. If such trajectories exist, the pathway is considered existing in the ground truth. In this way, 108 GT pathways were found. The richness of these pathways (i.e. count of trajectories) ranges from 1 to 1338, with up to 57% of them having less than 40 trajectories.

To evaluate the detection of pathways, we sort the resulting pathways from the proposed approach and the MT approach based on their richness, which is measured by the number of tracklets in our approach and by the number of trajectories in the MT approach. Then, we take the richest 40 pathways of each approach and match them to the GT pathways[3]. The results of this experiment are presented in Table 1, which show that out of the richest 40 pathways, 14 are matched to true pathways in our proposed approach compared to only 10 in MT. It is worth noting that the MT approach can produce multiple pathways corresponding to the same GT pathway. We only count one match to a GT pathway as a true detection and the rest as false detections.

Common Pathway Spatial Layout Coverage

To our knowledge, all prior work evaluated only the count of identified pathways compared to GT. We introduce a new evaluation criteria based on measuring the similarity between the spatial layouts of a retrieved pathway and the corresponding GT pathway.

For each GT pathway, its trajectories are overlaid and accumulated on top of one another to construct a spatial probability map to represent the pathway's spatial extent in the scene and the level of activity at each point within it. Similarly, another probability map is constructed for each resulting pathway from the evaluated algorithm. From both probability maps, pixel-wise Precision and Recall are calculated, considering only the pixels with positive probability values, which we call *active pixels*.

[1] 60 pixels in our implementation.

[2] http://www.cps.ejust.edu.eg/index_files/ijcai_2016.htm

[3] The matching is done semi-automatically using bipartite graph matching followed by human inspection. Details are removed for space limitation.

GT Pathways (source gate-sink gate)			9-6	1-6	8-1	8-6	3-6	7-6	5-1
Method	MT [Jodoin *et al.*, 2013]	Precision	0.93	NA	NA	NA	0.82	NA	0.94
		Recall	0.48				0.67		0.61
	Proposed	Precision	0.98	0.95	0.99	NA	NA	NA	0.94
		Recall	0.48	0.33	0.35				0.46

Table 2: Pathway layout pixel-wise Precision/Recall scores for our approach and the MT approach [Jodoin *et al.*, 2013]. Results are shown for the 7 richest GT pathways (sorted descendingly by richness). NA indicates undetected pathways.

Table 2 shows the results of this experiment for our approach compared to MT. The scores are computed for the correctly detected pathways from the richest 7 GT pathways. Note that our approach detects more from these pathways than the MT approach. It also yields slightly better precision than the MT approach. However, for both of them the recall is lower than the precision. This can be explained by inspecting a sample of the scores as illustrated in Figure 4. As evident in the figure, GT pathways are sometimes very wide, either in the middle due to midway pedestrian diversion to avoid an obstacle, or at terminals due to the perspective effect which makes distances close to the camera appear much larger. This makes it hard for a detection algorithm to cover most of the spatial layout of the GT pathway. This is particularly true for our approach, which tends to produce coherent clusters. Sometimes, this results in distributing tracklets belonging to one GT pathway over multiple clusters, only one of them is matched with the corresponding GT pathway. We believe that the ignored clusters account for the recall loss.

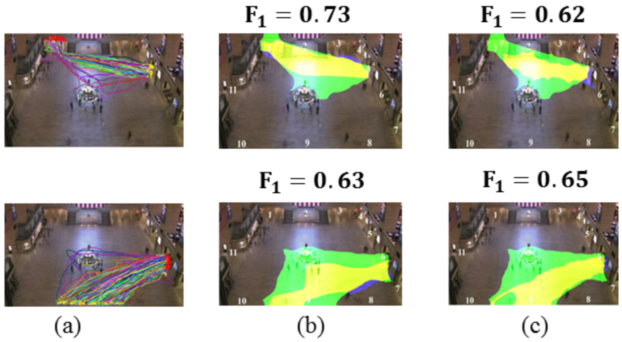


Figure 4: (a) GT pathway trajectories. (b), (c) Illustrate $\mathbf{F}_1$-score values for the pathway spatial layout coverage for the MT approach, and our approach, respectively. Yellow denotes True Positives, green False Negatives, and blue False Positives. Best viewed in color.

Source/Sink Region Detection

The detected source/sink regions (as explained in Section 4.3) are semi-automatically matched with the manually annotated Ground Truth source/sink regions (Figure 3b). Quantitative results are presented in Table 1. The results show that our approach detects 9 of the 11 GT gates, vs. only 6 detected by MT. On the other hand, the two approaches produce high count of false gates. This happens when a detected pathway starts from or terminates at an intermediate point that is not close to any GT gate.

5.2 Other Datasets

Qualitative results for other datasets are shown in Figure 5. These results are comparable to the results obtained by other approaches in the literature on the same scenes [Ali and Shah, 2007; Jodoin *et al.*, 2013].

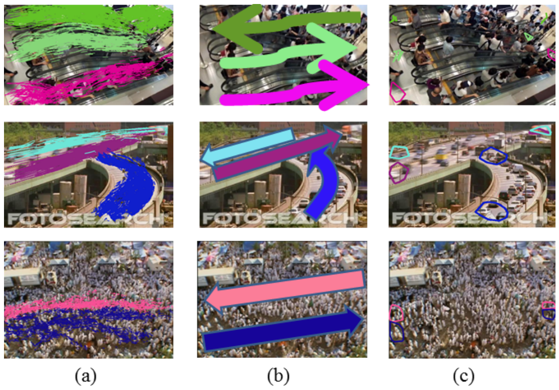


Figure 5: (a) Detected pathways using our approach for different datasets. (b) GT pathway motion directions. (c) Detected sources/sinks. (Best viewed in color).

6 Conclusion

In this paper, we propose a new approach for semantic region analysis of crowded scenes based on tracklet clustering. DD-CRP is adopted as a non-parametric clustering approach. Inspired by line geometry, a novel similarity measure is formulated, which effectively captures the spatial and directional similarity between tracklets during the clustering process. The proposed approach is evaluated against ground truth pathways from a recently released annotation for a challenging dataset. Pathways' spatial probability maps are constructed and active pixels of both identified pathway and ground truth are matched. Pixel-wise Precision/Recall measures are utilized to evaluate the spatial coverage of pathways. Our proposed work demonstrates state of the art performance both in pathways detection, their associated gates and spatial layout coverage. The proposed approach is also tested on different crowd scene datasets and demonstrates good qualitative performance.

Acknowledgment

This work is partially supported by the Information Technology Industry Development Agency (ITIDA) under the ITAC Program Grant no. PRP2015.R19.4-Automatic Crowd Scene Analysis and Anomaly Detection From Video Surveillance Cameras. It also has been supported by the Ministry of Higher Education (MoHE) of Egypt through a Ph.D. scholarship.

References

[Ali and Shah, 2007] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[Antoniak, 1974] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[Blei and Frazier, 2011] David M Blei and Peter I Frazier. Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488, 2011.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Etemadi *et al.*, 1991] A Etemadi, J-P Schmidt, George Matas, John Illingworth, and Josef Kittler. Low-level grouping of straight line segments. In *BMVC91*, pages 118–126. Springer, 1991.

[Ghosh *et al.*, 2011] Soumya Ghosh, Andrei B Ungureanu, Erik B Sudderth, and David M Blei. Spatial distance dependent chinese restaurant processes for image segmentation. In *Advances in Neural Information Processing Systems*, pages 1476–1484, 2011.

[Hu *et al.*, 2008] Min Hu, Saad Ali, and Mubarak Shah. Detecting global motion patterns in complex videos. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–5. IEEE, 2008.

[Jodoin *et al.*, 2013] Pierre-Marc Jodoin, Yannick Benezeth, and Yi Wang. Meta-tracking for video scene understanding. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 1–6. IEEE, 2013.

[Kratz and Nishino, 2012] Louis Kratz and Ko Nishino. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):987–1002, 2012.

[Kuo *et al.*, 2010] Cheng-Hao Kuo, Chang Huang, and Ramakant Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 685–692. IEEE, 2010.

[Li *et al.*, 2015] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 25(3):367–386, 2015.

[Shi and Tomasi, 1994] Jianbo Shi and Carlo Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994.

[Stauffer and Grimson, 1999] Chris Stauffer and W.E.L Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition (CVPR), 1999 IEEE Conference on*. IEEE, 1999.

[Tomasi and Kanade, 1991] Carlo Tomasi and Takeo Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[Topkaya *et al.*, 2015] Ibrahim Saygin Topkaya, Hakan Erdogan, and Fatih Porikli. Tracklet clustering for robust multiple object tracking using distance dependent chinese restaurant processes. *Signal, Image and Video Processing*, pages 1–8, 2015.

[Wang *et al.*, 2013] Chongjing Wang, Xu Zhao, Zhe Wu, and Yuncai Liu. Motion pattern analysis in crowded scenes based on hybrid generative-discriminative feature maps. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 2837–2841. IEEE, 2013.

[Wang *et al.*, 2014] Xiaofei Wang, Xiaomin Yang, Xiaohai He, Qizhi Teng, and Mingliang Gao. A high accuracy flow segmentation method in crowded scenes based on streakline. *Optik-International Journal for Light and Electron Optics*, 125(3):924–929, 2014.

[Yi *et al.*, 2015] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496, 2015.

[Zhao and Medioni, 2011] Xuemei Zhao and Gérard Medioni. Robust unsupervised motion pattern inference from video and applications. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 715–722. IEEE, 2011.

[Zhao *et al.*, 2011] Jing Zhao, Yi Xu, Xiaokang Yang, and Qing Yan. Crowd instability analysis using velocity-field based social force model. In *Visual Communications and Image Processing (VCIP), 2011 IEEE*, pages 1–4. IEEE, 2011.

[Zhou *et al.*, 2011] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3441–3448. IEEE, 2011.

[Zhou *et al.*, 2012] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2871–2878. IEEE, 2012.