Trajectories As Topics: Multi-Object Tracking by Topic Discovery

Wenhan Luo, Björn Stenger, Xiaowei Zhao and Tae-Kyun Kim

***Abstract*—This paper proposes a new approach to multi-object tracking by semantic topic discovery. We dynamically cluster frame-by-frame detections and treat objects as topics, allowing the application of the Dirichlet Process Mixture Model (DPMM). The tracking problem is cast as a topic-discovery task where the video sequence is treated analogously to a document. It addresses tracking issues such as object exclusivity constraints as well as tracking management without the need for heuristic thresholds. Variation of object appearance is modeled as the dynamics of word co-occurrence and handled by updating the cluster parameters across the sequence in the dynamical clustering procedure. We develop two kinds of visual representation based on super-pixel and deformable part model, and integrate them into the model of automatic topic discovery for tracking rigid and non-rigid objects respectively. In experiments on public data sets we demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—Multi-object tracking, Topic model, DPMM, Gibbs sampling.**

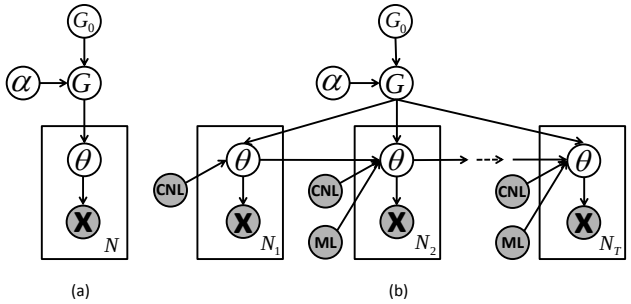

Fig. 1: Graphical model of the standard Dirichlet Process Mixture Model (DPMM) (a) and our topic model (b). In our model the document is temporally divided into epochs to model the temporal dynamics. ***CNL*** and ***ML*** denote cannot-link and must-link constraints.

I. Introduction

MULTIPLE object tracking is a mid-level computer vision task, which is used in applications such as action recognition, scene understanding and automatic video summarization. The task is to link a number of detection hypotheses into trajectories corresponding to different objects in a video. There has been significant progress in multi-object tracking [1], [2], [3], [4], [5], [6], [7]. However, issues like tracking management, appearance variation and occlusions remain challenging. Traditionally, the multi-object tracking task is cast as a data association problem in which detection hypotheses are associated into trajectories. Standard methods, such as the Hungarian algorithm, can be readily applied. However several practical considerations remain: Temporal gaps between observations may lead to disconnected trajectories of the same object [1], [4]. Determining the maximum allowable gap is difficult: low values will cause more fragmentation while higher values lead to more incorrect associations (ID switches). Handling track initialization and termination (also known as *tracking management*) is often based on heuristics. An existing trajectory may be terminated in the case of a single missing detection, resulting in fragmentation in some sequential approaches [8], [6]. Some approaches retain a "buffer" [9] for tracking initialization, so a trajectory is not counted until its length exceeds a pre-defined threshold. Appearance variation of objects may lead to fragmentation or ID switches as a result of inappropriate similarity measures. Physical constraints are rarely modeled explicitly, the work in [5] being one exception.

In this paper we propose an alternative approach to temporal data association by clustering detection instances, where each cluster corresponds to a unique object. An object is represented as a set of visual words, and we assume that this representation leads to similar patterns for the same object, while discriminating between different objects. An object thus corresponds to a semantic topic within a video sequence and the object is tracked as a new topic that evolves over time and eventually disappears.

We employ a Dirichlet Process Mixture Model (DPMM) to dynamically cluster detection responses into sets of objects (see Fig. 1 for the graphical model of our approach). The merit of applying a DPMM is that the number of semantic topics is learned automatically. Furthermore, it is naturally feasible to model dynamics in the clustering procedure for semantic topic discovery based on DPMM [10].

Specifically, we treat a detection hypothesis as a set of visual words. The uniqueness of word co-occurrence results in individual clusters, corresponding to individual objects by the application of DPMM. In a standard DPMM, when we consider the assignment of a given instance, the prior of which cluster the instance should belong to only depends on the number of existing instances in the cluster. However, in our problem, we also take the temporal distances between clusters and a given instance into consideration. Therefore, instead of treating the whole video as a single document, we

Wenhan Luo is with the Tencent AI Lab. Email: whluo.china@gmail.com
Björn Stenger is with the Rakuten Institute of Technology. Email: bjorn@cantab.net
Xiaowei Zhao is with the Alibaba Group. Email: zhiquan.zxw@alibaba-inc.com
Tae-Kyun Kim is with the Department of Electrical and Electronic Engineering, Imperial College London. Email: tk.kim@imperial.ac.uk

divide it into sequential *epochs* to model the dynamics of prior knowledge. On the other hand, adopting the temporal phenomena, the appearance variations of objects are dealt with by updating cluster parameters in different epochs in the clustering procedure.

In terms of the exclusivity constraints, which are (a) one detection could be assigned to no more than one trajectory and (b) no more than one detection could be assigned to an individual object/trajectory. By adopting clustering, the first exclusivity constraint is handled naturally by the assignment of each detection to only one cluster. To deal with the second constraint, we introduce the so called *cannot-link* constraints, which prohibits two detections in the same frame being assigned to one trajectory.

This paper extends our conference paper [11] in several ways. Firstly, by taking the scheme of dynamic clustering as a basic framework, we additionally tackle the problem of multiple pedestrian tracking. Based on the Deformable Part Model (DPM), we develop a representation model which not only considers the holistic but also the part-wise visual information. The combination of this model with DPMM handles the nonrigidness well. Secondly, we conduct experiments on public pedestrian datasets and compare with two state-of-the-art methods that also use the DPM in the visual representation. The comparison validates the effectiveness of the proposed approach.

To summarize, the contributions of automatic topic discovery for multi-object tracking are (1) multi-object tracking is cast as dynamic and sequential clustering by the application of DPMM without heuristics like "buffer" or maximum allowable temporal gap. Tracking management is handled automatically in the clustering procedure, (2) appearance variation of objects is modeled by the dynamics of cluster parameters, (3) exclusivity constraints are handled naturally due to the cluster assignments and the introduction of the *cannot-link* constraints to the model and (4) we provide a dynamic clustering algorithm as a tracking solution which could serve as a basic framework to integrate various kinds of appearance or motion models for improved tracking performance.

The remainder of the paper is organized as follows: Section II discusses related work. Section III briefly introduces the DPMM. Section IV proposes our dynamic clustering model for multiple object tracking, specifically for both rigid objects and non-rigid objects. Section V describes the inference of the proposed model. In Section VI we report experiment results corresponding to the two problems mentioned in Section IV. Section VII concludes the paper with some discussion.

II. Related work

Our work is related to both topic model and multiple object tracking. In the following, we discuss previous work about both aspects.

A. Topic Model

Topic model has been a long history in natural language processing, pattern recognition and computer vision. It can at least be dated back to Latent Semantic Indexing (LSI) [12] or Latent Semantic Analysis (LSA), which is applied in text analysis. This method is based on the principle that terms (or words) used in the same context are probably related to an identical topic. Based on LSA, probabilistic Latent Semantic Analysis (pLSA) [13] is proposed. pLSA models the co-occurrence of words and documents as a mixture of conditionally independent multinomial distributions. Further, Latent Dirichlet Allocation (LDA) [14] is developed for the task of text analysis. It is a generative model based on the bag-of-word assumption. Each document is assumed to be a combination of multiple topics, and each word is generated by one of the topics the document includes. Compared with pLSA, the distribution of topics in LDA has a Dirichlet prior. Dirichlet Process [15] is a distribution over distributions, *i.e.*, each draw from Dirichlet process is a distribution. It is particularly employed by Dirichlet Process Mixture Model, which is also called the infinite mixture model. Dirichlet process is the basis of the hierarchical Dirichlet process [16], within which the distribution of child is itself a Dirichlet process.

Topic models typically employ the concepts of words, topics and documents. Specifically, by treating a document as a bag of exchangeable words, documents are modeled as distributions over topics and topics are modeled as distributions over words. Thus topic models are naturally employed to deal with tasks of text analysis and natural language processing. On the other hand, they have been adopted to computer vision tasks in recent years due to the merits of these methods for discovering thematic structure. For example, a latent topic model is developed for object segmentation and classification [17]. This so-called Spatial LTM enforces the spatial coherency of the model and can simultaneously segment and classify objects. Similarly, spatial information is also integrated into a LDA model in [18] for image segmentation by Wang and Grimson. Topic models have been applied to numerous other tasks, such as region classification [19], trajectory analysis [20], image annotation [21], and image scene categorization [22].

B. Multiple Object Tracking

Different from classical visual tracking [23], [24], the task of multiple object tracking is to obtain the object-wise trajectories of objects in a given video (image sequence). In general, we discuss the work of multiple object tracking according to the following two kinds of problem settings. The first one is that the number of objects to be tracked is fixed and the initial states (location, size, *etc*) of these objects are manually given by human. This kind of setting is popular in the early work of multi-object tracking, such as [25], [26]. Recently, due to the advances of object detection, the second setting, *i.e.* the number of objects tracked is varying and initialization is provided by object detection, becomes more and more popular. Most of the work focuses on tracking pedestrians, since there are off-line trained human detectors [27], [28], [29], [30] which are ready to be employed.

Roughly, the first kind of multi-object tracking is usually handled in an online/sequential fashion. Approaches of this kind typically attain the observations up to the current frame.

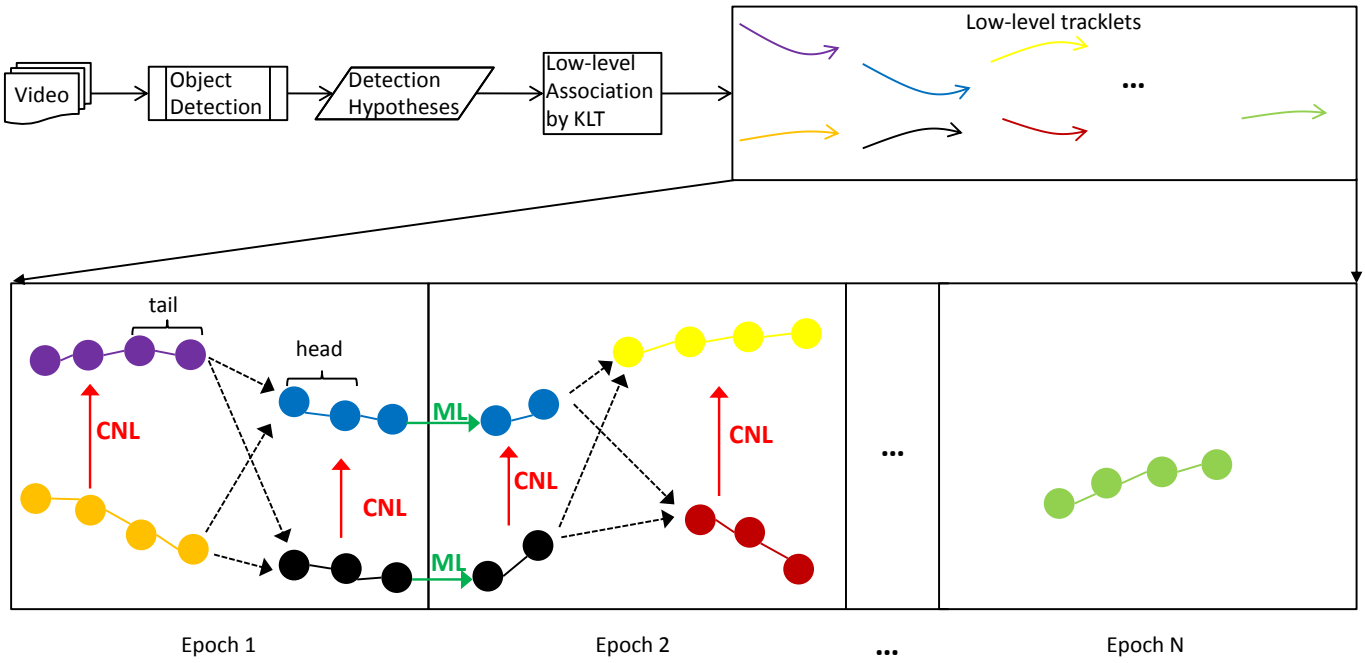

Fig. 2: **Schematic of the proposed method.** Tracklets are shown in different colors. Potential assignments are shown by dashed arrows. Temporally overlapping tracklets cannot be clustered together due to the ***cannot-link constraint*** (solid red arrows). The black tracklet and the blue tracklet temporally cross continuous epochs, and the segments of them are connected by the ***must-link constraint*** (solid green arrow). Note that, the purple and the orange tracklets in Epoch 1 could be directly connected to the yellow tracklet and the dark red tracklet in Epoch 2. In the last epoch, there is only one tracklet. Considering the temporal damping effect, the prior that this tracklet is linked to tracklets in previous epochs is limited if there is no intermediate tracklet bridging them. We dismiss some possible assignment arrows (for example the purple and the yellow tracklets could be possibly associated without linking the blue one) for the clarity of the figure (best viewed in color).

Based on the up-to-time observation, appearance [9], motion [31] and interaction models [2], [32], [33] are designed to discover appropriate candidates to extend the existing trajectories.

Solutions to the second kind of multi-object tracking generally deal with the observations in an offline/batch way. They attain the observations through the whole sequence, and usually tackle the tracking problem as a data association problem, associating observations into trajectories. The ideal association is obtained by minimizing a cost function, which is constructed based on pairwise observation similarities. Popular approaches include (but not limited to) Hungarian algorithms [34], [35], [36], [37], [3], [38], K-shortest path [39], [40], min-cost network flow [1], [41], [42], [43], [44], [45], [46], subgraph multicut [47], [48], maximum multi-clique [49], Conditional Random Field [50], [5] and Maximum Weight Independent Set [51]. Please refer to [52] for a more extensive review.

There are some works which cannot be categorized into the above two kinds. For example, some methods handle detection and tracking at the same time, such as [9]. There are also some approaches performing segmentation and clustering [53], [54] and conducting tracking based on the results.

Dirichlet Process Mixture Model has been applied in multi-object tracking [55], [56], [57], [58], while they are different from our method in various aspects. For example, aside from the techniques differences, [57], [56] are free of detectors. They are based on difference images computed across adjacent frames or GMM modeling of background, which limits their application in the scenery of fixed camera. The work of [55] adopts Dirichlet Processes for tracking maneuvering targets, while it does not need to consider appearance modeling, which is an important difference. We employ different appearance models based on superpixel and deformable part model, while the appearance model in [58] is simply based on color histogram.

In terms of clustering, Hu et al. [59] propose to employ Dirichlet Process Mixture Model for trajectory clustering. There are obvious differences between their work and ours. In our work, the instance to be clustered is trackerlet, which is not independent considering the temporal relation. The instance to be clustered in [59] is complete trajectory. Our task is to link/cluster tracklets belonging to one identical target as one complete trajectory. While their work focuses on mining the common patterns existing among multiple trajectories for applications such as retrieval.

Our work is also closely related to data association. PMHT (probabilistic multiple hypothesis tracking) [60], [61] and JPDA (Joint Probabilistic data association) [62] are classical methods for data association. PMHT is a method based on Bayesian framework. It proposes multiple hypothesizes over all possible data associations and employ the up-to-time observation to resolve ambiguities in the current time

step. JPDA seeks an optimal assignment between target and observation while maintaining contributions of all potential hypotheses from all tracks. On the contrary, our work of dynamic clustering provides an alternative to these classical methods and shows effectiveness.

III. DPMM

The Dirichlet Process Mixture Model (DPMM) [63] is a non-parametric model which assumes the data is governed by an infinite number of mixtures while only a fraction of these mixtures are activated by the data. Fig. 1(a) shows the graphical model of a DPMM. Assuming that the k-th mixture is parameterized by θ_k, each sample $\mathbf{x}_i$ is generated as follows:

$$\begin{aligned} G|\alpha, G_0 &\sim DP\left(\alpha, G_0\right) , \\ \theta_k|G &\sim G , \\ \mathbf{x}_i|\theta_{z_i} &\sim F\left(\theta_{z_i}\right) , \end{aligned} \tag{1}$$

where $DP\left(\bullet\right)$ is a Dirichlet process, G_0 is a base distribution, α is a concentration parameter, θ_k is drawn from G, which itself is a distribution drawn from the Dirichlet process. $F\left(\theta_{z_i}\right)$ denotes the distribution of observation $\mathbf{x}_i$ given θ_{z_i}, where z_i is the mixture indicator of $\mathbf{x}_i$. When this model is applied to clustering, z_i is the cluster index. Note that the number of mixtures in the model is determined by the data, *i.e.* the number of clusters is learned automatically, in contrast to parametric models such as K-means.

The Chinese Restaurant Process (CRP) illustrates the DPMM intuitively: Assuming an infinite number of tables (clusters), a new customer (observation) chooses an empty table with probability depending on α or joins an occupied table with a probability proportional to the number of people seated at that table. Formally,

$$\theta_i|\theta_{-i}, G_0, \alpha \sim \sum_k \frac{n_k}{i-1+\alpha}\delta\left(\phi_k - \theta_i\right) + \frac{\alpha}{i-1+\alpha}G_0 , \tag{2}$$

where ϕ_k is the parameter of cluster k, θ_{-i} is the set of associated parameters of $\mathbf{x}_{-i}$, *i.e.* observations except $\mathbf{x}_i$, n_k is the number of customers already at table k and $\delta\left(\cdot\right)$ is the Dirac delta function centered at 0. $\phi_{1:k}$ is the discrete set of values of $\{\theta_i\}$. It can also be written as $\theta_i = \phi_k$ with probability $\frac{n_k}{i-1+\alpha}$, and $\theta_i = \phi_{new}, \phi_{new} \sim G_0$ with probability $\frac{\alpha}{i-1+\alpha}$.

IV. Automatic topic discovery

In this section we develop a topic model to address the multi-object tracking problem. For different types of objects, *i.e.* rigid objects or non-rigid objects, we adopt different kinds of representation as visual words. We treat videos as documents and trajectories/objects as topics discovered in the video. We cluster coherent detection hypotheses (word co-occurrences) into trajectories (topics). As the number of objects/trajectories is not known in advance, it is learned from the data using a Dirichlet Process Mixture Model (DPMM). Fig. 2 illustrates the schematic of the proposed approach. Given a video, detection hypotheses are obtained by applying a ready-to-use object detector. Then these detection hypotheses are linked via KLT as low-level but reliable tracklets. These tracklets are the input of the dynamic clustering procedure, which groups tracklets belonging to an identical object into a cluster. In this stage, the exclusivity (cannot-link and must-link constraints) and the temporal damping effect are taken into account.

TABLE I: This table lists the correspondences between the topic model and the -object tracking problem.

Multi-object tracking	**Automatic topic discovery**
detections	word occurrence
trajectory	topic
video	document
video segment	epoch
exclusivity constraints	cluster membership exclusivity & cannot link
data association	dynamic clustering

To adopt the DPMM, we make the following analogies, which are described in Table I. The left column lists some concepts in the multi-object tracking problem, and the right column represents the corresponding entities in our approach of automatic topic discovery. Classical text-analysis applications of the DPMM assume that the document consists of a bag of exchangeable words, *i.e.* without specific order and without any dynamic modeling. In our problem, words are not assumed to be exchangeable as we consider a set of visual words in a detection hypothesis jointly as an observation and the representation of visual words in this paper additionally encodes the spatial information. As appearance of object varies (temporal dynamics of word-occurrence), the distribution of visual words in an object (topic) is dynamic across the video. In light of this, the video is divided temporally into epochs and each epoch is modeled by a DPMM with associated hyper-parameters. During the clustering procedure, the states of clusters are updated across epochs.

Further, as objects appear and disappear, corresponding to the birth and death of topics, the distributions of topics also vary across different epochs. We also observe that between two adjacent epochs, the distribution of words in a topic and the distributions of topics in a document are closely related to each other due to temporal continuity. Thus the relation between continuous DPMMs is modeled as a first-order Markov process. Fig. 1(b) shows the graphical model of the proposed approach. Each epoch (except the first one) is closely related to its previous epoch. Additionally, we take the spatio-temporal exclusivity, *i.e.* the cannot-link and the must-link constraints, into consideration.

A. DPMM-SP

In this section, we describe the superpixel-based visual representation and the likelihood computation in the tracking problem of multiple objects.

1) Visual Representation: Fig. 3 shows the visual representation. We adopt superpixels, pixel groups of similar color and location [64], for representing visual appearance. In our implementation, a detection bounding box is segmented into approximately 200 SLIC superpixels [64], each described as a 5-dimensional vector (r, g, b, x, y), where (r, g, b) and (x, y)

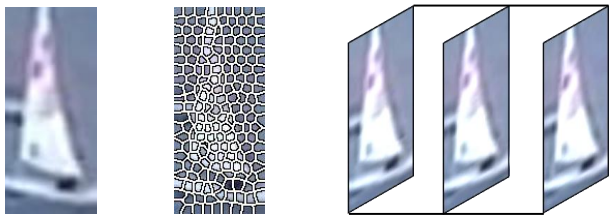

Fig. 3: **Visual representation based on superpixels.** A detection bounding box (a) which is segmented into a set of superpixels shown in (b). The rightmost side (c) shows an exemplar tracklet.

are the mean color and position, respectively. We cluster all superpixels from all frames in a video by K-means (K is the number of clusters) and define a dictionary according to the cluster prototypes. Each bounding box is quantized using this dictionary and represented as a histogram. Similar to part-based models [27] this object representation exhibits some robustness to partial occlusion since some superpixels representing the object will remain visible.

Usually the detection responses are linked into low-level reliable tracklets [65] in a pre-processing step. Here we employ KLT tracking to obtain N low-level tracklets, $\mathbf{x}_{1:N}$. Each tracklet is represented as a tuple $\mathbf{x}_i = \left\langle A_i^{head}, A_i^{tail}, T_i^{head}, T_i^{tail}, \widetilde{A}_i, \widetilde{V}_i \right\rangle$, where A_i^{head} and A_i^{tail} are the appearance representations (histograms) of the head and tail element within tracklet $\mathbf{x}_i$, $\widetilde{A}_i$ and $\widetilde{V}_i$ are the average appearance (center) and the covariance of histograms of the complete tracklet, T_i^{head} and T_i^{tail} are the time indexes of the head and tail element in $\mathbf{x}_i$.

2) Likelihood: Based on the object representation, let the parameters of cluster k at epoch t be $\phi_{t,k}$, including the center $\widetilde{A}_{t,k}$ and covariance matrix $\widetilde{V}_{t,k}$, are computed from the super-pixel representation of all the detection within the concerned cluster up to the current epoch. Given $\mathbf{x}_{t,i}$, the likelihood of an observation with the concerned cluster is estimated as

$$\begin{aligned} & f\left(\mathbf{x}_{t,i}|\phi_{t,k}, \mathbf{x}_{t,k,\cdot}\right) \\ & \propto s\big(A_{t,i}^{head}, A_{t,k,m}^{tail}\big)\, s\big(A_{t,i}^{tail}, A_{t,k,n}^{head}\big)\, p\Big(\widetilde{A}_{t,i}; \phi_{t,k}\Big) \ , \end{aligned} \tag{3}$$

where $\mathbf{x}_{t,k,\cdot}$ is the set of observations associated with $\phi_{t,k}$, $A_{t,k,m}^{tail}$ and $A_{t,k,n}^{head}$ are the feature vector of tail tracklet $\mathbf{x}_{t,k,m}$ and head tracklet $\mathbf{x}_{t,k,n}$ which are closest to $\mathbf{x}_{t,i}^{head}$ and $\mathbf{x}_{t,i}^{tail}$ respectively regarding temporal distance.

$s\left(\cdot,\cdot\right)$ is the similarity between two histograms. It has the following form:

$$s\left(A_1, A_2\right) = \exp\left(-Bhatt\left(A_1, A_2\right)\right) \ . \tag{4}$$

$p(\widetilde{A}_{t,i}; \phi_{t,k})$ is the likelihood of $\widetilde{A}_{t,i}$ given $\phi_{t,k}$. It is computed with regard to the distance between two Gaussian distributions, one corresponding to the cluster and the other one corresponding to the concerned tracklet. To be concrete, it is reversely proportional to the distance D between the cluster and tracklet, as

$$p\left(\widetilde{A}_{t,i}; \phi_{t,k}\right) \propto \exp\left(-D\left(\widetilde{A}_{t,i}, \widetilde{V}_{t,i}, \widetilde{A}_{t,k}, \widetilde{V}_{t,k}\right)\right) \ , \tag{5}$$

where $D\left(\widetilde{A}_{t,i}, \widetilde{V}_{t,i}, \widetilde{A}_{t,k}, \widetilde{V}_{t,k}\right)$ of with the following form

$$\begin{aligned} & D\left(\widetilde{A}_{t,i}, \widetilde{V}_{t,i}, \widetilde{A}_{t,k}, \widetilde{V}_{t,k}\right) \\ & = \left(\widetilde{A}_{t,i} - \widetilde{A}_{t,k}\right)^T \left(\frac{\widetilde{V}_{t,i} + \widetilde{V}_{t,k}}{2}\right)^{-1} \left(\widetilde{A}_{t,i} - \widetilde{A}_{t,k}\right) . \end{aligned} \tag{6}$$

Note that in Eq. (3) the first two terms compute the local affinity and the last term computes the global affinity in terms of temporal span.

B. (DPM)^2

In this section, we combine DPMM with the Deformable Part Model (DPM) [27] to deal with the problem of tracking multiple pedestrians. Compared with tracking multiple rigid objects in Section IV-A, the visual representation and the likelihood are different, which are illustrated as follows.

1) Visual Representation: Due to the non-rigid property of pedestrians, rather than the super-pixel representation, we adopt the Deformable Part Model (DPM) [27] to represent objects. The DPM has been successful in object detection, but has not often been applied to object tracking, [9], [66] being exceptions. The DPM represents an object with a root filter and a set of part filters. The score accounts for the part appearances along with the deformation of part locations with respect to the root. We see that the position, size, and appearance of parts of pedestrians exhibit unique patterns from person to person. On the other hand, the co-occurrence of these parts is similar if they belong to the same person. Therefore, we treat parts as words in the document, and the tracking problem as a topic discovery task where pedestrians are treated analogously to topics.

To be specific, we represent an object based on the holistic bounding box along with a set of parts (the number of parts is 8), which are the outputs of a DPM detector. We extract HOG and color features from the holistic bounding box and the associated parts as the appearance information for this detection hypothesis. Additionally, we exploit the configuration of the set of parts within the holistic bounding box. Generally, the head part is visible almost all the time, even in case of (partial) occlusion. By taking the head part as an anchor, the offset of the other parts could be calculated. We stack the spatial offset as feature, termed as deformable feature. The appearance feature and the deformable feature constitute the visual representation. Similar to the case of tracking rigid objects, we employ KLT to link detections into low-level tracklets, and represent the head and the tail of the tracklet separately. Each tracklet is represent as a tuple as $\mathbf{x}_i = \left\langle A_i^{head}, A_i^{tail}, \widetilde{D}_i, \widetilde{V}_i, T_i^{head}, T_i^{tail} \right\rangle$, where A_i^{head} and A_i^{tail} are the appearance features of the head and the tail, $\widetilde{D}_i$ and $\widetilde{V}_i$ are the mean and the covariance of the deformable configuration of the tracklet, T_i^{head} and T_i^{tail} are the frame indexes of the starting frame and the ending frame of the tracklet.

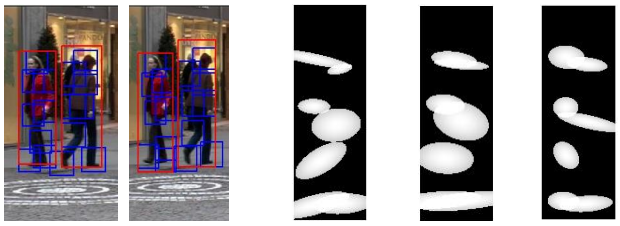

Fig. 4: **Visual representation based on DPM.** (a) Detection samples of continuous frames from the TUD-Stadtmitte data set. Note the part configuration of the same person and different persons. (b) and (c) show the visualization results of the part configuration of the same person at different times, while (d) shows the visualization result of a different person. Based on the likelihood represented in the following, similarity value between (b) and (c) is larger than that between (c) and (d).

2) Likelihood: Concerning the representation of pedestrian in our problem, parameters associated to cluster k are the center $\widetilde{D}_t^k$ and covariance $\widetilde{V}_t^k$ of the deformable feature, which are computed from all the deformable configurations within the cluster up to the current epoch. Given a tracklet $\mathbf{x}_{t,i}$, the likelihood of the tracklet belonging to the concerned cluster is

$$\begin{aligned} &f\left(\mathbf{x}_{t,i}|\phi_{t,k},\mathbf{x}_{t,k,\cdot}\right) \\ &\propto s\left(A_{t,i}^{head},A_{t,k,m}^{tail}\right)s\left(A_{t,i}^{tail},A_{t,k,n}^{head}\right)p\left(\widetilde{D}_{t,i},\widetilde{V}_{t,i};\phi_{t,k}\right), \end{aligned} \tag{7}$$

where $\mathbf{x}_{t,k,\cdot}$ is the set of observations associated with $\phi_{t,k}$, $A_{t,k,m}^{tail}$ and $A_{t,k,n}^{head}$ are visual representation of the tail tracklet $\mathbf{x}_{t,k,m}$ and head tracklet $\mathbf{x}_{t,k,n}$ which are closest to $\mathbf{x}_{t,i}^{head}$ and $\mathbf{x}_{t,i}^{tail}$ respectively regarding temporal distance.

$s\left(\bullet,\bullet\right)$ is the appearance similarity between the head of one tracklet and the tail of another tracklet by considering both the holistic bounding box and the parts. To be specific, it is computed as:

$$s\left(A_1,A_2\right)=l\left(\mathbf{h}_1^B,\mathbf{h}_2^B\right)+\frac{1}{8}\sum_{j=1}^{8}l\left(\mathbf{h}_{1,j}^P,\mathbf{h}_{2,j}^P\right), \tag{8}$$

where the first term corresponds to appearance of the holistic bounding box, and the second term accounts for the appearance information of the set of parts, $\mathbf{h}$ is the feature vector. B and P abbreviate "box" and "part" respectively. $l\left(\bullet,\bullet\right)$ is a similarity measurement between two feature histograms. We adopt the similar formula in Eq. (4) as the measurement.

$p\left(\widetilde{D}_{t,i},\widetilde{V}_{t,i};\phi_{t,k}\right)$ is the likelihood of $\mathbf{x}_{t,i}$ given cluster parameter $\phi_{t,k}$ considering the deformable configuration of parts. It is formulated as

$$p\left(\widetilde{D}_{t,i},\widetilde{V}_{t,i};\phi_{t,k}\right)\propto\exp\left(-d\left(\widetilde{D}_{t,i},\widetilde{V}_{t,i},\widetilde{D}_{t,k},\widetilde{V}_{t,k}\right)\right), \tag{9}$$

where $d\left(\widetilde{D}_{t,i},\widetilde{V}_{t,i},\widetilde{D}_{t,k},\widetilde{V}_{t,k}\right)$ is the distance between the cluster and the concerned tracklet considering the deformable configuration, as the following

$$\begin{aligned} &d\left(\widetilde{D}_{t,i},\widetilde{V}_{t,i},\widetilde{D}_{t,k},\widetilde{V}_{t,k}\right)= \\ &\frac{1}{7}\sum_{j=1}^{7}\left(\widetilde{D}_{t,i,j}-\widetilde{D}_{t,k,j}\right)^T\left(\frac{\widetilde{V}_{t,i,j}+\widetilde{V}_{t,k,j}}{2}\right)^{-1}\left(\widetilde{D}_{t,i,j}-\widetilde{D}_{t,k,j}\right) \end{aligned} \tag{10}$$

Here we compute the deformable configuration in a part-wise fashion. Note that the number of parts we are considering is 7, rather than 8 in Eq. (8). This is because we adopt the head part as the anchor, which is not taken into account. It is also worthy to note that in Eq. (7) the first two terms locally account for the appearance information and the third term globally considers the deformable information in terms of temporal span.

C. Cannot Links & Must Links

The first temporal exclusion constraint, that one detection can be assigned to no more than one trajectory, is modeled by the exclusive property of cluster membership of each object detection. The second one, *i.e.* one trajectory cannot occupy more than one detection within the same frame, is modeled by the cannot-link constraint. If two tracklets in the same epoch overlap temporally, they cannot have the same cluster label, *i.e.* they cannot be linked to be part of an identical object. We represent the set of cannot-link constraints in epoch t as

$$\mathbf{CNL}_t=\left\{\left(\mathbf{x}_{t,i},\mathbf{x}_{t,j}\right)|z_{t,i}\neq z_{t,j}\right\}, \tag{11}$$

where $z_{t,i}$ and $z_{t,j}$ are cluster membership indicators of tracklets $\mathbf{x}_{t,i}$ and $\mathbf{x}_{t,j}$ which overlap in epoch t. The partitioning of the video into epochs may split tracklets into segments. We use must-link constraints to connect these tracklets from adjacent epochs. This kind of constraints for epoch t is given by

$$\mathbf{ML}_t=\left\{\left(\mathbf{x}_{t,i},\mathbf{x}_{t-1,j}\right)|z_{t,i}=z_{t-1,j}\right\}. \tag{12}$$

Fig. 2 shows some examples of the cannot links and the must links. Note that there are no must-link constraints for the first epoch.

D. Temporal Damping

Temporal effects need to be included during the process of clustering observations. Let us illustrate this by the Chinese Restaurant Process (CRP) representation. In CRP, prior knowledge depends only on the existing number of customers belonging to the table. However, in our multi-object tracking problem, this is not sufficient. When we calculate the cluster to which a tracklet belongs, we additionally need to take the temporal gap between this tracklet and existing clusters into account. For example, considering a cluster which is temporally distant from the given tracklet, the probability that the tracklet is assigned to this cluster is low, even if there are already many tracklets assigned to this cluster. In other words, the assignment prior probability should decay with the

temporal gap between a cluster and the tracklet. Considering a tracklet at epoch t, suppose some clusters already exist, the number of members belonging to cluster k at epoch τ is damped by a weight, similar to [67], as:

$$n_{k,\tau} = \sum_j \delta\left(z_{\tau,j} - k\right) \exp\left(-\eta\left(t - \tau\right)\right) , \quad \tau < t , \tag{13}$$

where z is the cluster membership indicator and η is a damping factor.

V. Inference

Assuming there are N tracklets as $\mathbf{x}_{1:N}$ and T epochs, let us denote the observations in epoch t as $\mathbf{x}_{1:N_t}$, the corresponding estimations as $\theta_{1:N_t}$. We consider the first-order relation in our model, *i.e.* the first epoch is a normal DPMM and subsequent DPMMs are closely related to the previous DPMM. The posterior probability is written as

$$\begin{aligned}
&P\left(\theta_{1:N}|\mathbf{x}_{1:N}, \alpha, G_0, \mathbf{CNL}, \mathbf{ML}\right) \\
&= P\left(\theta_{1:N_1}|\mathbf{x}_{1:N_1}, \alpha, G_0, \mathbf{CNL}_1\right) \times \\
&\prod_{t=2}^{T} P\left(\theta_{1:N_t}|\mathbf{x}_{1:N_t}, \theta_{1:N_{t-1}}, \alpha, G_0, \mathbf{CNL}_t, \mathbf{ML}_t\right) \\
&\propto P\left(\theta_{1:N_1}|\mathbf{x}_{1:N_1}, \alpha, G_0, \mathbf{CNL}_1\right) \times \\
&\prod_{t=2}^{T} f\left(\mathbf{x}_{1:N_t}|\theta_{1:N_t}\right) P\left(\theta_{1:N_t}|\theta_{1:N_{t-1}}, \alpha, G_0, \mathbf{CNL}_t, \mathbf{ML}_t\right) ,
\end{aligned} \tag{14}$$

where $f\left(\cdot\right)$ is the likelihood function, $P\left(\theta_{1:N_t}|\theta_{1:N_{t-1}}, \alpha, G_0, \mathbf{CNL}_t, \mathbf{ML}_t\right)$ encodes the evolution over time.

Computing the posterior is intractable, thus we use Gibbs sampling for inference [10], introducing the latent cluster indicator variable of $\mathbf{x}_{t,i}$ as $z_{t,i}$. For each epoch, the input are the tracklets in this epoch and existing clusters up to this epoch; the output are the clusters after being assigned tracklets in the current epoch. The state of the sampler contains both the cluster indicators $z_{t,\cdot}$ of all observations and the states $\phi_{t,\cdot}$ of all clusters. We iterate between two steps: (1) given the current states of clusters, sample cluster indicators for all the observations, (2) given all cluster indicators of observations, update the states of clusters.

(1) Enforcing must-link and cannot-link constraints, cluster indicators are sampled as follows:

(a) if $\mathbf{x}_{t,i}$ is a member of the must-link set, *i.e.* $\mathbf{ML}_t$, the cluster indicator of $\mathbf{x}_{t,i}$ should be identical to that of its must-link counterpart $\mathbf{x}_{t-1,j}$;

(b) else the cluster indicator of $\mathbf{x}_{t,i}$ is sampled according to the conditional posterior as $P\left(z_{t,i}|z_{1:t-1}, z_{t,-i}, \mathbf{x}_{t,i}, \mathbf{x}_{t,k,\cdot}, \phi_{t,1:k}, \alpha, G_0\right)$. This is analogous to standard DPMM sampling, thus this probability can be written as:

$$\begin{aligned}
&P\left(z_{t,i} = k|\cdots\right) \\
&\propto \frac{n_{k,1:t-1} + n_{k,t,-i}}{N_{1:t-1} + N_t + \alpha - 1} f\left(\mathbf{x}_{t,i}|\phi_{k,t}, \mathbf{x}_{t,k,\cdot}\right) ,
\end{aligned} \tag{15}$$

where $n_{k,1:t-1} = \sum_{\tau=1}^{t-1} n_{k,\tau}$ is the number of past observations with cluster indicator k, $n_{k,t,-i} = \sum_{j \in -i} \delta\left(z_{t,j} - k\right)$, $N_{1:t-1} = \sum_{k \in \mathbf{K}} n_{k,1:t-1}$, $\mathbf{K}$ is the set of indicators of existing clusters.

We also allow the emergence of a new cluster with probability

$$\begin{aligned}
&P\left(z_{t,i} = new\ cluster|\cdots\right) \\
&\propto \frac{\alpha}{N_{1:t-1} + N_t + \alpha - 1} \int_\theta f\left(\mathbf{x}_{t,i}|\theta\right) dG_0(\theta) .
\end{aligned} \tag{16}$$

(c) due to the cannot-link set, if $\mathbf{x}_{t,i}$ belongs to $\mathbf{CNL}_t$, then $z_{t,i}$ must be different from all its cannot-link counterparts. Thus $z_{t,i}$ should be sampled from the indicators of all existing clusters excluding those of all $\mathbf{x}_{t,i}$'s cannot-link counterparts. According to this, when we compute the probability, we replace $\phi_{t,1:k}$ with $\phi_{t,1:k} \backslash \phi_{t,-i}$, where $\phi_{t,-i}$ is the set of clusters which $\mathbf{x}_{t,i}$'s cannot-link counterparts belongs to, and $\backslash$means the set difference operation.

(2) We update cluster parameters given cluster indicators by estimating $P\left(\phi_{t,k}|z_{t,\cdot}, \mathbf{x}_{t,\cdot}, \phi_{t-1,k}\right)$. Since a cluster is conditionally independent from other clusters given the cluster indicators, this probability can be written as $P\left(\phi_{t,k}|\mathbf{x}_{t,k,\cdot}, \phi_{t-1,k}\right) \propto G_0\left(\phi_{t,k}\right) f\left(\mathbf{x}_{t,k,\cdot}|\phi_{t,k}\right) P\left(\phi_{t,k}|\phi_{t-1,k}\right)$, where $\mathbf{x}_{t,k,\cdot}$ is the set of observations associated with $\phi_{t,k}$ and $f\left(\mathbf{x}_{t,k,\cdot}|\phi_{t,k}\right)$ is the likelihood. $P\left(\phi_{t,k}|\phi_{t-1,k}\right)$ encodes the cluster parameter dynamics, which is reversely proportional to the distance between the two Gaussian distributions corresponding to $\phi_{t,k}$ and $\phi_{t-1,k}$. To be more specific, it is with the following form:

$$\phi_{t,k}|\phi_{t-1,k} \sim \mathcal{N}\left(\phi_{t-1,k}, \gamma\mathbf{I}\right) . \tag{17}$$

Next we sample to update the parameters of the cluster.

These two steps are carried out iteratively in each epoch, resulting in observations with the same cluster indicator being linked into one trajectory, which corresponds to one object.

VI. Experiments

In this section, experiment settings, metrics, and results of DPMM applications based on the described two kinds of visual representation are reported.

A. Settings

We divide videos into epochs. There is an extreme case with regard to the division. One can treat the whole video sequence as one epoch, *i.e.*, without temporal dynamics in the clustering. To set the epoch size appropriately, we investigate the performance of different sizes. Empirically we found that the division of sequence into epochs does improve performance compared with the case without division, while it is not a key factor in the performance. Thus we empirically set the epoch size to be $50 - 200$ frames, depending on the length of the video. We set the dictionary dimension to 50, η to 0.2 and γ in Eq. (17) to 0.01 in all experiments. In the inference stage, for each epoch we run Gibbs sampling for 500 iterations and report results after the last iteration.

TABLE II: **MULTI-OBJECT TRACKING RESULTS.** The proposed method is compared with GMOT [8] and BLP [6], in terms of MT, ML, FM and IDS values. Results of the proposed method are in the shaded columns. The arrows next to the metrics indicate the direction of better performance, *e.g.* ↑ means larger values are better.

Sequence	*MT*↑				*ML*↓				*FM*↓				*IDS*↓			
	GMOT	GMOT-ATD	BLP	BLP-ATD	GMOT	GMOT-ATD	BLP	BLP-ATD	GMOT	GMOT-ATD	BLP	BLP-ATD	GMOT	GMOT-ATD	BLP	BLP-ATD
Zebra	0.44	0.43	0.58	0.61	0.29	0.30	0.25	0.25	36	27	30	26	6	3	7	1
Crab	0.10	0.15	0.21	0.25	0.71	0.68	0.69	0.69	243	134	205	163	114	77	63	15
Antelope	0.37	0.38	0.69	0.74	0.37	0.37	0.18	0.16	33	28	54	32	19	16	31	6
Goose	0.64	0.71	0.79	0.79	0.07	0.07	0.04	0.04	52	38	36	19	28	27	33	12
Sailing	0.25	0.50	0.83	0.83	0.08	0.08	0.08	0.08	99	85	45	40	33	11	12	8
Hockey	0.68	0.71	0.61	0.68	0.11	0.11	0.14	0.11	27	23	24	10	17	9	20	3
Overall	0.34	0.38	0.51	0.55	0.41	0.39	0.34	0.34	490	335	394	290	217	143	166	45

B. Metrics

To evaluate tracking performance we employ the metrics proposed in [68], [69]. These metrics include mostly tracked (MT) ground-truth trajectories, mostly lost (ML) ground-truth trajectories, fragmentation (FM), and ID switches (IDS). MT is the percentage of the ground-truth trajectories which are covered temporally for over 80% in time. ML is the percentage of the ground-truth trajectories which are recovered for less than 20% in length. The FM metric counts the number of interruptions of the ground-truth trajectories and IDS the number of times that the ground-truth trajectories change their matched ID. The Multiple Object Tracking Accuracy (MOTA) metric combines the false positive rate, false negative rate and mismatch rate into a single number, giving a fairly reasonable quantity for the overall tracking performance. Multiple Object Tracking Precision (MOTP) describes how precisely the objects are tracked measured by bounding box overlap. Multiple Object Detection Accuracy (MODA), which considers the relative number of false positives and miss detections. The Multiple Object Detection Precision (MODP) metric measures the quality of alignment between predicted detections and the ground truth.

C. MOT by DPMM-SP

1) Data sets: We apply our algorithm to two problems, (1) generic multi-object tracking [8], [6], [70], where multiple objects of any type are detected and tracked and (2) multi-pedestrian tracking, requiring the output of an off-line trained pedestrian detector as input. For the first problem, we employ public six data sets from [6] named *Zebra, Crab, Goose, Hockey, Sailing* and *Antelope*. For the second problem, we use the public *ETHMS* and *TUD-Stadtmitte* data sets.

2) Results: In this section, we represent the results of tracking multiple objects based on the super-pixel visual representation. The experiments are conducted in three parts. In the first part we compare our approach to existing sequential approaches [8], [6] in solving the generic multi-object tracking problem. The second part compares our algorithm with several state-of-the-art data association algorithms [4], [3] using the same detection results and visual representation. We additionally conduct the experiment of tracking multiple pedestrians based on the super-pixel visual representation, which is intended to check the effectiveness of the dynamic clustering solution. The results are reported in the third part, and are compared with those of other methods to multi-pedestrian tracking [2], [1], [5], [7].

Part 1 – Comparison with generic multi-object trackers. In this part, we compare our automatic topic discovery (ATD) algorithm with two state-of-the-art generic multi-object trackers, GMOT [8] and BLP [6], respectively. Note that, GMOT and BLP address both detection and tracking, while ATD focuses on tracking (data association) only. Thus in the comparison, we use the same detection results used in the compared method to directly compare the tracking part in the compared method with ATD. For example, GMOT and GMOT-ATD use the same detection results. BLP and BLP-ATD use the same detection results (while different from those used in GMOT and GMOT-ATD). We suspect this kind of comparison could directly show the improvements from ATD when fixing other parts. The results are shown in Table II. GMOT-ATD and BLP-ATD are the proposed algorithms based on the same detection results from the corresponding counterparts. The results of GMOT and BLP are quoted from [8] and [6], respectively. Compared with GMOT our algorithm reduces the quantity of FM and IDS by 32% and 34%. Compared with BLP, the FM and IDS values are reduced by 26% and 73%, respectively. This means that the proposed algorithm tracks objects in the test sequences more consistently. Note however, that the proposed algorithm is a batch algorithm while both GMOT and BLP process the data sequentially. The next set of experiments therefore directly compares with batch data association methods.

Part 2 – Comparison with data association algorithms. In this section we compare our method with a number of data association algorithms, including (1) DA-H: the Hungarian algorithm [3], (2) DA-DP: dynamic programming in network flow [4], (3) DA-SSP: successive shortest path in network flow [4] and (4) BL: a baseline method of our algorithm without temporal dynamics, *i.e.* the video sequence is treated as a single document without division into epochs. This can be viewed as the application of standard DPMM to our problem.

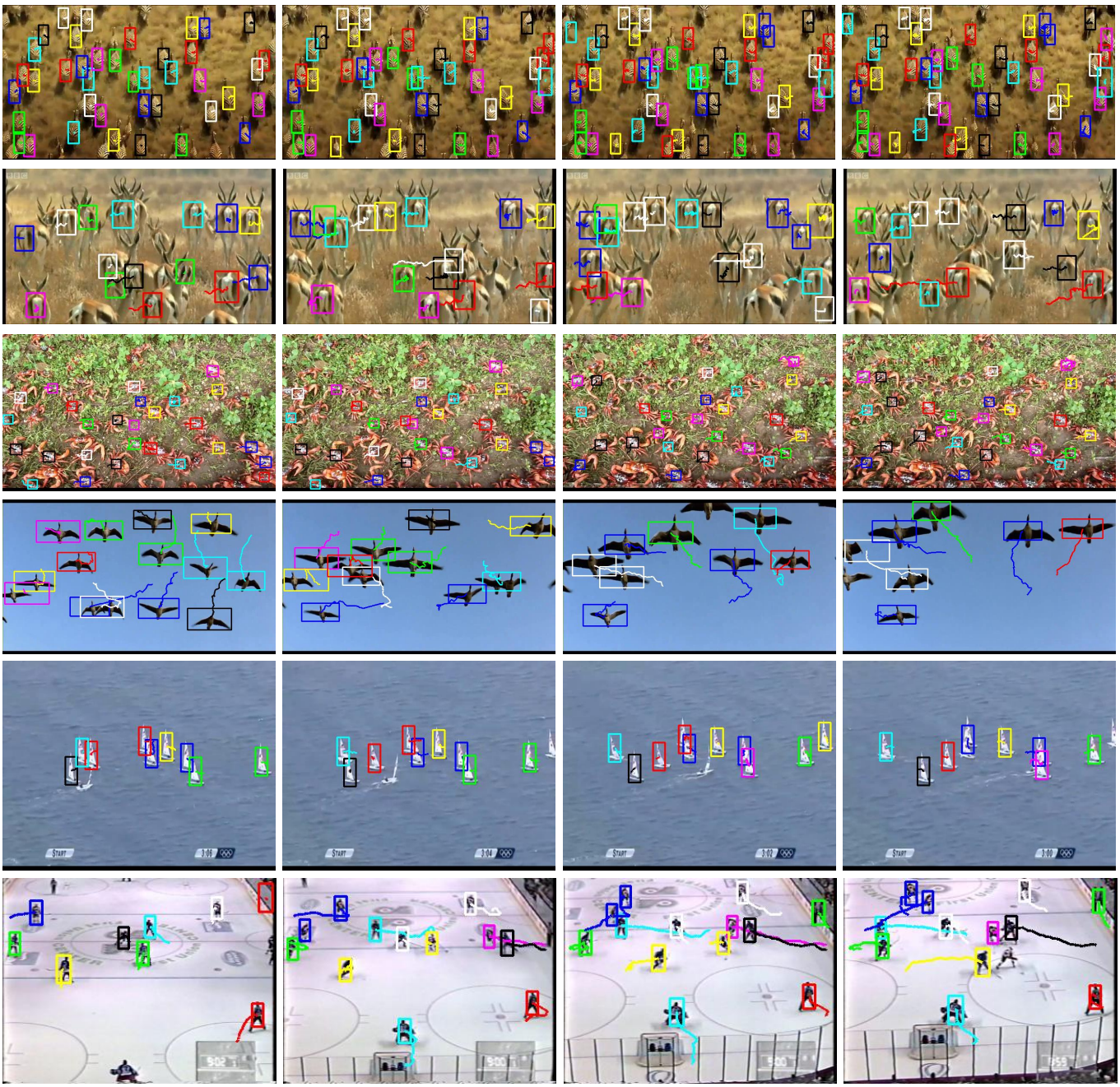

Fig. 5: Qualitative results on the (from top to bottom) *Zebra*, *Antelope*, *Crab*, *Goose*, *Sailing* and *Hockey* sequences.

Note that, the data association methods we compare with are classical and popular. They are still employed by more recent work, such as [45], [43]. For fairness, all algorithms are given the same detection results from [6]. The results of DA-DP and DA-SSP are obtained using the code from [4].

Results in Table III indicate that (1) generally DA-H tends to achieve good MT and ML values, meaning it is able to track objects more completely. On the other hand, the performance in terms of FM and IDS are worse than ours; (2) DA-DP and DA-SSP obtain good FM and IDS values, indicating that they can track objects more robustly and consistently. DA-SSP achieves slightly better FM and IDS than DA-DP. However, compared with DA-H, they tend to ignore parts of trajectories, thus MT and ML values are worse than those of DA-H; (3) compared with DA-H, BL has similar MT and ML values while achieving better FM and IDS values, showing the effectiveness of applying a DPMM; (4) the proposed method (DPMM-SP) achieves the best performance. Compared with BL, it further reduces the IDS and FM values.

Some example result images of the six data sets for GMOT problem are shown in Fig. 5.

Part 3 – Comparison with pedestrian trackers. In this part, we evaluate our method on the multiple pedestrian tracking problem where the raw detection results are those in [5]. We compare our results with those in [2], [1], [5], [7]. [2] develops a sophisticated dynamic model based on social

TABLE III: **DATA ASSOCIATION COMPARISON**, in terms of MT, ML, FM and IDS values. The best results are shown in bold.

Sequence	*MT*↑					*ML*↓					*FM*↓					*IDS*↓				
	DA-H	DA-DP	DA-SSP	BL	DPMM-SP	DA-H	DA-DP	DA-SSP	BL	DPMM-SP	DA-H	DA-DP	DA-SSP	BL	DPMM-SP	DA-H	DA-DP	DA-SSP	BL	DPMM-SP
Zebra	0.59	0.55	0.54	0.60	**0.61**	**0.25**	0.35	0.35	**0.25**	**0.25**	28	32	31	27	**26**	3	2	7	3	**1**
Crab	0.24	0.19	0.19	**0.25**	**0.25**	**0.69**	0.70	0.70	**0.69**	**0.69**	170	168	166	168	**163**	27	31	30	28	**15**
Antelope	**0.75**	0.63	0.63	0.72	0.74	**0.15**	0.27	0.27	**0.15**	0.16	36	33	**32**	37	**32**	14	10	10	25	**6**
Goose	**0.79**	0.64	0.68	**0.79**	**0.79**	**0.03**	0.25	0.32	0.04	0.04	34	31	29	25	**19**	25	20	18	14	**12**
Sailing	0.83	0.83	0.83	0.83	0.83	0.08	0.08	0.08	0.08	0.08	42	45	44	**40**	**40**	10	9	**8**	**8**	**8**
Hockey	0.64	0.54	0.54	**0.68**	**0.68**	0.14	0.18	0.18	**0.11**	**0.11**	12	11	**10**	12	**10**	11	7	6	6	**3**
Overall	0.54	0.47	0.46	0.54	**0.55**	0.34	0.41	0.42	**0.33**	0.34	322	320	312	309	**290**	90	79	73	84	**45**

TABLE IV: **MULTI-PEDESTRIAN TRACKING RESULTS** compared with other state-of-the-art methods in terms of MT, ML, FM and IDS values. The best results are shown in bold.

Sequence	*TUD-Stadtmitte*		*ETHMS*				
	[5]	DPMM-SP	[2]	[1]	[5]	[7]	DPMM-SP
MT↑	.400	**.900**	.516	.556	.664	**.720**	.589
ML↓	0	0	.056	.062	.082	**.047**	.073
FM↓	**13**	16	206	178	**69**	85	156
IDS↓	15	**13**	77	138	**57**	71	103

forces during association. [1] casts data association as finding the min-cost in network flow. [5] adopts a CRF model for data association.

Qualitative results (result images) are shown in Fig. 6. The *ETHMS* data set is composed of two sub sets, so the results are shown separately in two rows. Please note the green bounding box and the blue one in the top row of Fig. 6. These two pedestrians involve in occlusion. The proposed method maintains the identities correctly in the occlusion. However, the proposed approach also fails in some cases. For example, in the bottom row of Fig. 6, the red bounding box and white bounding box correspond to an identical person. The proposed method fails to link them due to miss detections. This case is expected to be solved if we could include some sophisticated motion models.

Quantitative results are shown in Table IV. On the *TUD-Stadtmitte* data set, our algorithm achieves better ML and IDS performance while obtaining worse FM performance. On the *ETHMS* data set, the results of the proposed method are comparable to those of [2] and [1] but worse than those of [5] and [7], which are all methods tailored to the task of pedestrian tracking. We suspect the reason is that although we take the same raw detection hypotheses as input, our approach does not include sophisticated appearance or motion models. In contrast, the motion model in [2] takes the effect of pedestrians in a group into account, which is helpful in reducing ID switches in the case of occlusion. The method in [1] includes a model named Explicit Occlusion Model (EOM) which especially handles occlusion by generating occlusion hypotheses and integrating them in the network. Besides considering exclusivity constraints, a motion model based on angular velocity is taken into consideration in [5]. [7] achieves the best MT and ML performance as a result of their contextual motion model, which is able to recover more trajectory components, even in the case of missed detections, by learning a dictionary of interaction features among objects. In our method, only the plain but general super-pixel representation is considered for appearance modeling. The super-pixel representation could perform well in representing rigid objects in the first application - generic multiple object tracking. It is specifically robust to illumination changes. As the clustering of super-pixels into prototype super-pixels would alleviate the noise resulted from illumination changes. While it inevitably suffers from clutters from backgrounds in representing non-rigid objects such as pedestrians. On the other hand, our approach can serve as a basic model to include more sophisticated appearance or motion models. In the next section, we would demonstrate the results of a $(\text{DPM})^2$ model, which includes visual representation developed by specifically considering the non rigidness of human, for multiple pedestrian tracking.

Our algorithm, as described in Section V, carries out inference iteratively, which is time consuming. As a batch method, the time to process a whole video depends on factors such as the length of the video, the number of pedestrians in the video (proportional to the number of tracklets). The run-time for each frame is approximately two seconds using Matlab code on a desktop PC (i5 CPU, 8G RAM).

D. MOT by $(DPM)^2$

Results of multiple pedestrian tracking indicate that, performance of pedestrian tracking partly relies on the representation model. This motivates us to develop a better visual representation model (Section IV-B) than the plain super-pixel representation. In this part, the results of tracking multiple pedestrians based on the proposed $(\text{DPM})^2$ model are reported.

1) Data sets: Three data sets are employed in the experiment. The first two are *TUD-Stadtmitte* and *ETHMS*, the same as those used in the last part of the previous section. The third

Fig. 6: Qualitative results of the proposed approach on the TUD-Stadtmitte, ETHMS (subset 1), ETHMS (subset 2) data sets (from top to bottom).

one is named *ParkingLot*. The reason of the usage of this data sets is that it is employed in [66], [9], which handle the problem of multiple pedestrian tracking by the employment of Deformable Part Model. Thus the adoption of this data set directly compares the proposed method with the other two [66], [9].

2) Results: In this section, the results are presented in two parts. In the first part, we show the comparison between $(DPM)^2$ and DPMM-SP for the task of multiple pedestrian tracking on the data sets of *TUD-Stadtmitte* and *ETHMS*. In the second part, the comparison between the performance of the proposed $(DPM)^2$ and some other state-of-the-art methods for multi-pedestrian tracking is presented.

Part 1 – Comparison between DPMM-SP and $(DPM)^2$. As shown in Table V, $(DPM)^2$ outperforms DPMM-SP on the data sets of *TUD-Stadtmitte* and *ETHMS*. More specifically, on *TUD-Stadtmitte* the values of MT and ML remain the same while the values of FM and IDS decrease by 43.8% and 15.4%. The improvement of performance is supposed to result from the DPM representation in $(DPM)^2$ as these two methods differ only from the representation of pedestrians. The similar observation can be obtained when we compare the results of $(DPM)^2$ and DPMM-SP on *ETHMS*. The model based on deformable part model is more suitable for the task of tracking pedestrians, as it can model the intrinsic configuration of the articulated human body.

Part 2 – Comparison between $(DPM)^2$ and other pedestrian trackers. Table VI shows the comparison between the proposed method $(DPM)^2$ and the other two state-of-the-art methods, which are termed as $(MP)^2$ [66] and PMT [9]. The results reveal that, 1) $(DPM)^2$ outperforms the PMT method [9] and 2) except the value of DP, DPM does not outperform the $(MP)^2$, but the values could still be comparable.

TABLE V: **MULTI-PEDESTRIAN TRACKING RESULTS** compared between DPMM-SP and $(DPM)^2$ in terms of MT, ML, FM and IDS values. The best results are shown in bold.

Sequence	*TUD-Stadtmitte*		*ETHMS*	
	DPMM-SP	$(DPM)^2$	DPMM-SP	$(DPM)^2$
MT↑	0.900	0.900	0.589	0.589
ML↓	0	0	0.073	**0.048**
FM↓	16	**9**	156	**140**
IDS↓	13	**11**	103	**102**

The comparison above suggests that, (1) treating multi-object tracking as automatic topic discovery presents an advancement over using specific SVM classifiers for individual objects, which is adopted in PMT [9] and (2) integrating motion or occlusion models such as in [66] lead to significant improvements over appearance-only models. The integration of such models is likely to lead to improvements of the proposed automatic topic discovery framework.

TABLE VI: **MULTI-PEDESTRIAN TRACKING RESULTS** on the *Parking Lot* data set, compared with other state-of-the-art methods in terms of MOTA, MOTP, MODA and MODP values. The best results are shown in bold.

Method	MOTA ↑	MOTP ↑	MODA ↑	MODP ↑
PMT [9]	0.793	0.741	0.798	0.742
$(MP)^2$ [66]	**0.889**	**0.775**	**0.965**	0.936
$(DPM)^2$	0.830	0.751	0.886	**0.959**

VII. Conclusions

This paper has introduced a topic model for the multi-object tracking problem. Thanks to the Dirichlet Process Mixture Model, tracking management is addressed by dynamical clustering. Along with the introduced cannot-link constraints, the exclusivity constraints are handled naturally. The dynamics of object appearance variation are modeled by segmenting the video into temporal epochs. Two different types of visual representations have been implemented and used in the dynamic clustering procedure to track rigid and non-rigid objects. Experiments on public data sets have shown the advantages of applying a topic discovery method over other data association methods and sequential solutions.

References

[1] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *CVPR*, 2008, pp. 1–8.

[2] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *ICCV*, 2009, pp. 261–268.

[3] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *CVPR*, 2009, pp. 1200–1207.

[4] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR*, 2011, pp. 1201–1208.

[5] A. Milan, K. Schindler, and S. Roth, "Detection-and trajectory-level exclusion in multiple object tracking," in *CVPR*, 2013, pp. 3682–3689.

[6] W. Luo, T.-K. Kim, B. Stenger, X. Zhao, and R. Cipolla, "Bi-label propagation for generic multiple object tracking," in *CVPR*, 2014.

[7] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *CVPR*, 2014.

[8] W. Luo and T.-K. Kim, "Generic object crowd tracking by multi-task learning," in *BMVC*, 2013.

[9] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *CVPR*, 2012, pp. 1815–1821.

[10] A. Ahmed and E. P. Xing, "Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering." in *SDM*, 2008, pp. 219–230.

[11] W. Luo, B. Stenger, X. Zhao, and T.-K. Kim, "Automatic topic discovery for multi-object tracking," in *Proc. of the Association for the Advancement of Artificial Intelligence*, 2015.

[12] S. Dumais, G. Furnas, T. Landauer, S. Deerwester, S. Deerwester *et al.*, "Latent semantic indexing," in *Proceedings of the Text Retrieval Conference*, 1995.

[13] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[15] Y. W. Teh, "Dirichlet process," in *Encyclopedia of machine learning*, 2010, pp. 280–287.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[17] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *ICCV*, 2007, pp. 1–8.

[18] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *NIPS*, 2008, pp. 1577–1584.

[19] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in *CVPR*, 2007, pp. 1–8.

[20] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models," *IJCV*, vol. 95, no. 3, pp. 287–312, 2011.

[21] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *CVPR*, 2009, pp. 1903–1910.

[22] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, vol. 2, 2005, pp. 524–531.

[23] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *CVPR*, 2013, pp. 2411–2418.

[24] ——, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[25] M. Yang, T. Yu, and Y. Wu, "Game-theoretic multiple target tracking," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[26] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang, "Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 12, pp. 2420–2440, 2012.

[27] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[29] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[31] L. Kratz and K. Nishino, "Tracking with local spatio-temporal motion patterns in extremely crowded scenes," in *CVPR*, 2010, pp. 693–700.

[32] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *CVPR*, 2011, pp. 1345–1352.

[33] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 120–127.

[34] B. Song, T.-Y. Jeng, E. Staudt, and A. K. Roy-Chowdhury, "A stochastic graph evolution framework for robust multi-target tracking," in *ECCV*, 2010, pp. 605–619.

[35] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *CVPR*, 2012.

[36] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," in *ECCV*, 2010.

[37] A. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," in *CVPR*, 2006.

[38] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *ECCV*, 2008.

[39] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *PAMI*, vol. 33, no. 9, pp. 1806–1819, 2011.

[40] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *ECCV*, 2012.

[41] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, "Coupling detection and data association for multiple object tracking," in *CVPR*, 2012, pp. 1948–1955.

[42] A. A. Butt and R. T. Collins, "Multi-target tracking by lagrangian relaxation to min-cost network flow," in *CVPR*, 2013, pp. 1846–1853.

[43] P. Lenz, A. Geiger, and R. Urtasun, "FollowMe: Efficient online min-cost flow tracking with bounded memory and computation," in *ICCV*, 2015.

[44] A. Dehghan, Y. Tian, P. H. S. Torr, and M. Shah, "Target identity-aware network flow for online multiple target tracking," in *CVPR*, 2015.

[45] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *CVPR*, 2015.

[46] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, 2016.

[47] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Subgraph decomposition for multi-target tracking," in *CVPR*, 2015.

[48] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *CVPR*, 2017.

[49] A. Dehghan, S. Modiri Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *CVPR*, 2015.

[50] B. Yang and R. Nevatia, "An online learned crf model for multi-target tracking," in *CVPR*, 2012, pp. 2034–2041.

[51] W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *CVPR*, 2011, pp. 1273–1280.

[52] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," *arXiv:1409.7618*, 2014.

[53] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *CVPR*, 2015, pp. 5397–5406.

[54] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in *CVPR*, vol. 1. IEEE, 2006, pp. 594–601.

[55] E. B. Fox, E. B. Sudderth, and A. S. Willsky, "Hierarchical dirichlet processes for tracking maneuvering targets," in *Information Fusion, 2007 10th International Conference on*. IEEE, 2007, pp. 1–8.

[56] I. S. Topkaya, H. Erdogan, and F. Porikli, "Detecting and tracking unknown number of objects with dirichlet process mixture models and markov random fields," in *International Symposium on Visual Computing*. Springer, 2013, pp. 178–188.

[57] W. Neiswanger, F. Wood, and E. Xing, "The dependent dirichlet process mixture of objects for detection-free tracking and object modeling," in *Artificial Intelligence and Statistics*, 2014, pp. 660–668.

[58] I. S. Topkaya, H. Erdogan, and F. Porikli, "Tracklet clustering for robust multiple object tracking using distance dependent chinese restaurant processes," *Signal, Image and Video Processing*, vol. 10, no. 5, pp. 795–802, 2016.

[59] W. Hu, X. Li, G. Tian, S. Maybank, and Z. Zhang, "An incremental dpmm-based method for trajectory clustering, modeling, and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1051–1065, 2013.

[60] N. Chenouard, I. Bloch, and J.-C. Olivo-Marin, "Multiple hypothesis tracking for cluttered biological image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2736–3750, 2013.

[61] R. L. Streit and T. E. Luginbuhl, "Probabilistic multi-hypothesis tracking," NAVAL UNDERWATER SYSTEMS CENTER NEWPORT RI, Tech. Rep., 1995.

[62] S. Hamid Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid, "Joint probabilistic data association revisited," in *ICCV*, 2015, pp. 3047–3055.

[63] D. M. Blei, M. I. Jordan *et al.*, "Variational inference for dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[64] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[65] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *CVPR*, 2010, pp. 685–692.

[66] H. Izadinia, I. Saleemi, W. Li, and M. Shah, "Mp2t: Multiple people multiple parts tracker," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 100–114.

[67] X. Zhu, Z. Ghahramani, and J. Lafferty, "Time-sensitive dirichlet process mixture models," DTIC Document, Tech. Rep., 2005.

[68] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *CVPR*, vol. 1, 2006, pp. 951–958.

[69] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009.

[70] X. Zhao, D. Gong, and G. Medioni, "Tracking using motion patterns for very crowded scenes," in *ECCV*, 2012, pp. 315–328.