Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Automatic Topic Discovery for Multi-object Tracking

Wenhan Luo
Imperial College London

Björn Stenger
Toshiba Research Europe

Xiaowei Zhao
Imperial College London

Tae-Kyun Kim
Imperial College London

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. This formulation addresses tracking issues such as object exclusivity constraints as well as *cannot-link constraints* which are integrated without the need for heuristic thresholds. The video is temporally segmented into epochs to model the dynamics of word (superpixel) co-occurrences and to model the temporal damping effect. In experiments on public data sets we demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Multi-object tracking is a mid-level computer vision task, which is employed in applications such as action recognition or automatic video summarization. The task is to link a number of given detection hypotheses into trajectories corresponding to different objects in a video. There has been significant progress in multi-object tracking (Zhang, Li, and Nevatia 2008; Pellegrini et al. 2009; Xing, Ai, and Lao 2009; Pirsiavash, Ramanan, and Fowlkes 2011; Milan, Schindler, and Roth 2013; Luo et al. 2014; Leal-Taixé et al. 2014), however, issues like tracking management, appearance variations and occlusions remain challenging. Traditionally, the multi-object tracking task is cast as a data association problem in which detection hypotheses are grouped 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 (Zhang, Li, and Nevatia 2008; Pirsiavash, Ramanan, and Fowlkes 2011). 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 (Luo and Kim 2013; Luo et al. 2014). 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 (Milan, Schindler, and Roth 2013) being one exception. Uniqueness constraints model the fact that (a) at most one object per frame can be associated with each trajectory, and (b) no more than one trajectory can be assigned to the same detected object.

In this paper we propose an alternative approach to temporal data association by clustering detection instances, where each cluster corresponds to a unique object. We introduce a text-document analogy, where an object corresponds to a semantic topic within a video sequence. A semantic topic can be defined as co-occurring words – we use a bag-of-superpixel representation for objects, encoding appearance and spatial information. An object is thus tracked as a new topic that evolves over time and fades away.

We employ a Dirichlet Process Mixture Model (DPMM) to dynamically cluster detection responses into sets of objects (Ahmed and Xing 2008). 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 the 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 the given instance into consideration. Therefore, instead of treating the whole video as a single document, we divide it into sequential *epochs* in order to model the dynamics of prior knowledge and appearance variation of objects during clustering. In terms of constraints, by adopting clustering, the exclusivity constraint of one trajectory per object is handled naturally by the assignment of each detection to only one cluster. To deal with the other constraint, we introduce the so called *cannot-link* constraint, which prohibits two detections in the same frame being assigned to one trajectory.

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 applica-

tion of DPMM without heuristics like 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, (4) in a more general sense, we provide a dynamic clustering algorithm as a tracking solution which could serve as a basic framework to integrate other appearance, or motion models for multi-object tracking.

2 Related Work

The most relevant work on topic models and multiple object tracking is reviewed in the following.

2.1 Topic Model

Popular topic models for text document processing include Latent Semantic Indexing (LSI) (Dumais et al. 1995), probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999), Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), and Dirichlet Processes (DP) (Teh 2010). These 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. Due to the merits of these methods for discovering thematic structure, they have been adopted in computer vision tasks in recent years. For example, a latent topic model is developed for object segmentation and classification (Cao and Fei-Fei 2007). Spatial information is integrated into a LDA model (Wang and Grimson 2008) for image segmentation by Wang and Grimson. Topic models have been applied to numerous other tasks, such as region classification (Verbeek and Triggs 2007), trajectory analysis (Wang et al. 2011), image annotation (Wang, Blei, and Li 2009), and image scene categorization (Fei-Fei and Perona 2005).

2.2 Multi-object Tracking

Approaches for multi-object tracking may be classified into two sets, sequential (or online) and batch (or offline) methods. Sequential methods typically attain the observations up to the current frame. Based on these observations, appearance models (Shu et al. 2012), motion models (Kratz and Nishino 2010), and interaction models (Pellegrini et al. 2009; Yamaguchi et al. 2011) are designed to discover appropriate candidates to extend existing trajectories. Batch methods attain the observations through the whole sequence, and usually treat tracking as a data association problem. Associations are obtained by minimizing a cost function, which is constructed from pairwise observation similarities. Popular approaches include the Hungarian algorithm (Xing, Ai, and Lao 2009), K-shortest paths (Berclaz et al. 2011), min-cost network flow (Zhang, Li, and Nevatia 2008; Butt and Collins 2013), Conditional Random Fields (Yang and Nevatia 2012; Milan, Schindler, and Roth 2013), and Maximum Weight Independent Sets (Brendel, Amer, and Todorovic 2011). Please refer to (Luo, Zhao, and Kim 2014) for a more extensive review.

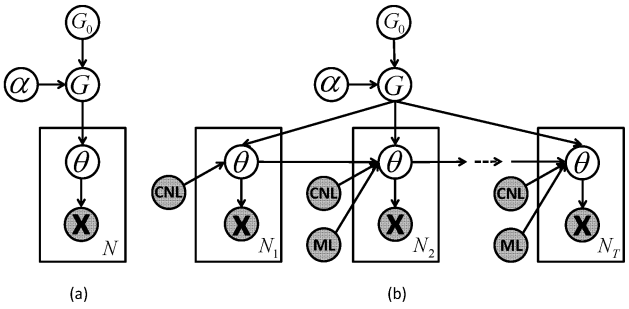

Figure 1: Graphical model of the DPMM (a) and our topic model (b). In our model the document is temporally divided into epochs to model the temporal dynamics. ***CNL*** and ***ML*** denote the introduced cannot-link and must-link constraints.

3 Dirichlet Process Mixture Model

The Dirichlet Process Mixture Model (DPMM) (Blei and Jordan 2006) is a non-parametric model which assumes the data is governed by an infinite number of mixtures where 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, and $F\left(\theta_{z_i}\right)$ 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 $\frac{\alpha}{i-1+\alpha}$.

4 Automatic Topic Discovery

In this section we develop a topic model to address the multi-object tracking problem. We treat superpixels as words,

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 DPMM.

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 superpixels (words) in a detection hypothesis jointly as an observation. We also take temporal information into account, *i.e.* the prior probability of which cluster an observation belongs to depends on their temporal distance. To model dynamics (Ahmed and Xing 2008), we divide the video into temporal epochs, modelling each epoch by a DPMM with associated hyper-parameters. As object appearance varies (temporal dynamics of word-occurrence), the distributions of superpixels (words) in an object (topic) is dynamic across the video (sequential documents). 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.

4.1 Visual Representation

We adopt superpixels, pixel groups of similar color and location (Achanta et al. 2012), for representing visual appearance. In our implementation, a detection bounding box is segmented into approximately 200 SLIC superpixels (Achanta et al. 2012), each described as a 5-dimensional vector (r, g, b, x, y), where (r, g, b) and (x, y) are the mean color and position, respectively. We cluster all superpixels from all frames in the video by K-means and define a dictionary from the cluster prototypes. Each bounding box is quantized using this dictionary and represented as a histogram. Similar to part-based models (Felzenszwalb et al. 2010) 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 (Kuo, Huang, and Nevatia 2010) in a preprocessing step. Here we employ KLT tracking to obtain N low-level tracklets, $\mathbf{x}_{1:N}$. Each tracklet is represented as a tuple $\mathbf{x}_i = \left\langle A_i^h, A_i^t, \widetilde{A}_i, \widetilde{V}_i, T_i^h, T_i^t \right\rangle$, where A_i^h and A_i^t 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 and the covariance of the appearance histograms of the complete tracklet $\mathbf{x}_i$, T_i^h and T_i^t are the time indexes of the head and tail element in $\mathbf{x}_i$.

4.2 Temporal Constraints

The first temporal exclusion constraint, that at most one object can be assigned to each trajectory, is modeled by the exclusive property of cluster membership of each object detection. The second one, *i.e.* one trajectory cannot be assigned more than one detection within the same frame, is modeled by a 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

$$\mathbf{CNL}_t = \left\{ (\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) \,|z_{t,i} \neq z_{t,j} \right\}, \tag{3}$$

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 between tracklets from adjacent epochs to connect these. The constraint for epoch t is given by

$$\mathbf{ML}_t = \left\{ (\mathbf{x}_{t,i}, \mathbf{x}_{t-1,j}) \,|z_{t,i} = z_{t-1,j} \right\}. \tag{4}$$

Note that there are no must-link constraints for the first epoch.

4.3 Temporal Damping

Temporal effects need to be included during the process of clustering the observations. Let us illustrate this by the Chinese Restaurant Process (CRP) representation. In the CRP, prior knowledge only depends 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 which the tracklet belongs to, we additionally need to take the temporal gap between this new 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 (Zhu, Ghahramani, and Lafferty 2005), as:

$$n_{k,\tau} = \sum_j \delta\left(z_{\tau,j} - k\right) \exp\left(-\eta\left(t - \tau\right)\right), \quad \tau < t, \tag{5}$$

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

4.4 Likelihood

Based on the object representation, the cluster parameter is a Gaussian distribution with center $\widetilde{A}$ and covariance matrix $\widetilde{V}$, which are computed from the appearance histograms within the cluster. The likelihood of an observation given a cluster is estimated as

$$\begin{aligned} &f\left(\mathbf{x}_{t,i} | \phi_{t,k}, \mathbf{x}_{t,k,\cdot}\right) \\ &\propto s\left(\mathbf{x}_{t,i}^{head}, \mathbf{x}_{t,m}^{tail}\right) s\left(\mathbf{x}_{t,i}^{tail}, \mathbf{x}_{t,n}^{head}\right) p(\widetilde{A}_{t,i}, \widetilde{V}_{t,i}; \phi_{t,k}), \end{aligned} \tag{6}$$

where $\mathbf{x}_{t,k,\cdot}$ is the set of observations associated with $\phi_{t,k}$, $\mathbf{x}_{t,m}^{tail}$ and $\mathbf{x}_{t,n}^{head}$ are the tail detection and head detection

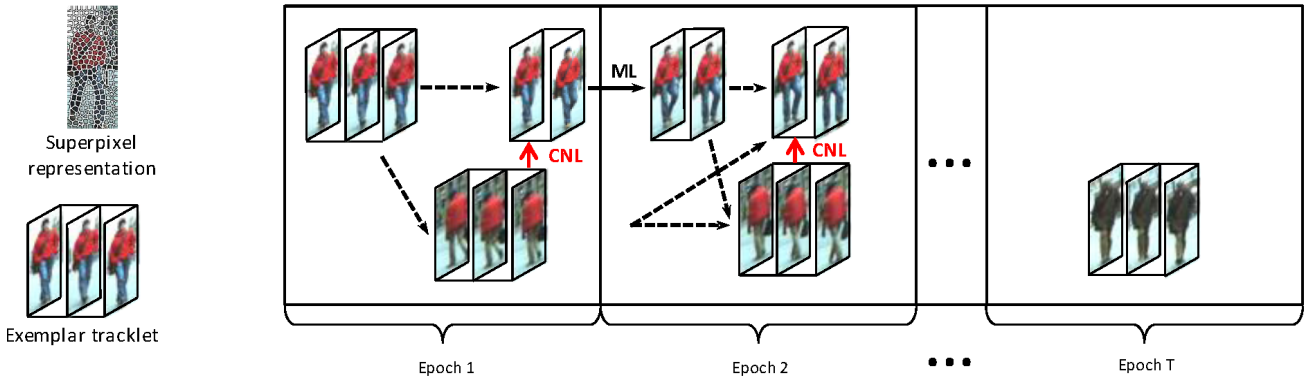

Figure 2: **Schematic of the proposed method.** The left side shows object representation and an exemplar tracklet. The right side shows the dynamic clustering procedure. Potential assignments are shown by dashed arrows. Temporally overlapping tracklets cannot be clustered together due to the ***cannot-link constraint*** (solid red arrows). Note the second tracklet and the third tracklet in the first row. They are from a single tracklet which crosses the first and the second epoch before we generate tracklets, so they are connected by the ***must-link constraint*** (solid black arrow). 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 (figure best viewed in color).

which are closest to $\mathbf{x}_{t,i}^{head}$ and $\mathbf{x}_{t,i}^{tail}$ respectively regarding temporal difference, $s(\cdot,\cdot)$ is the similarity between two histograms based on the superpixel representation, $p(\widetilde{A}_{t,i}, \widetilde{V}_{t,i}; \phi_{t,k})$ is the likelihood of $\widetilde{A}_{t,i}$ and $\widetilde{V}_{t,i}$ given $\phi_{t,k}$. Note that the first two terms compute the local affinity and the last term computes the global affinity in terms of temporal span.

5 Inference

Assuming there are N tracklets as $\mathbf{x}_{1:N}$ and T number of 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

$$
\begin{aligned}
&P\left(\theta_{1:N}|\mathbf{x}_{1:N},\alpha,G_0,\mathbf{CNL},\mathbf{ML}\right)\\
&= P\left(\theta_{1:N_1}|\mathbf{x}_{1:N_1},\alpha,G_0,\mathbf{CNL}_1\right)\times\\
&\prod_{t=2}^{T} P\left(\theta_{1:N_t}|\mathbf{x}_{1:N_t},\theta_{1:N_{t-1}},\alpha,G_0,\mathbf{CNL}_t,\mathbf{ML}_t\right)\\
&\propto P\left(\theta_{1:N_1}|\mathbf{x}_{1:N_1},\alpha,G_0,\mathbf{CNL}_1\right)\times\\
&\prod_{t=2}^{T} f\left(\mathbf{x}_{1:N_t}|\theta_{1:N_t}\right) P\left(\theta_{1:N_t}|\theta_{1:N_{t-1}},\alpha,G_0,\mathbf{CNL}_t,\mathbf{ML}_t\right)\ ,
\end{aligned}
\tag{7}
$$

where $f(\cdot)$ 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, and we use Gibbs sampling for inference (Ahmed and Xing 2008), introducing the latent cluster indicator variable of $\mathbf{x}_{t,i}$ as $z_{t,i}$. For each epoch, the inputs are the tracklets in this epoch and existing clusters up to this epoch; the outputs 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 with the difference of temporal damping, thus this probability can be written as:

$$
\begin{aligned}
&P\left(z_{t,i}=k|\cdots\right)\\
&\propto \frac{n_{k,1:t-1}+n_{k,t,-i}}{N_{1:t-1}+N_t+\alpha-1} f\left(\mathbf{x}_{t,i}|\phi_{k,t},\mathbf{x}_{t,k,\cdot}\right)\ ,
\end{aligned}
\tag{8}
$$

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

We also allow the emergence of a new cluster with probability

$$
\begin{aligned}
&P\left(z_{t,i}=new\ cluster|\cdots\right)\\
&\propto \frac{\alpha}{N_{1:t-1}+N_t+\alpha-1}\int_{\theta} f\left(\mathbf{x}_{t,i}|\theta\right) dG_0(\theta)\ .
\end{aligned}
\tag{9}
$$

Table 1: **Multi-object tracking results.** The proposed automatic topic discovery (ATD) is applied to GMOT (Luo and Kim 2013) and BLP (Luo et al. 2014) and is evaluated 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	.44	.43	.58	.61	.29	.30	.25	.25	36	27	30	26	6	3	7	1
Crab	.10	.15	.21	.25	.71	.68	.69	.69	243	134	205	163	114	77	63	15
Antelope	.37	.38	.69	.74	.37	.37	.18	.16	33	28	54	32	19	16	31	6
Goose	.64	.71	.79	.79	.07	.07	.04	.04	52	38	36	19	28	27	33	12
Sailing	.25	.50	.83	.83	.08	.08	.08	.08	99	85	45	40	33	11	12	8
Hockey	.68	.71	.61	.68	.11	.11	.14	.11	27	23	24	10	17	9	20	3
Overall	.34	.38	.51	.55	.41	.39	.34	.34	490	335	394	290	217	143	166	45

(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 $P\left(\phi_{t,k}|\mathbf{x}_{t,k,\cdot}, \phi_{t-1,k}\right) \propto G_0\left(\phi_{t,k}\right) f\left(\mathbf{x}_{t,k,\cdot}|\phi_{t,k}\right) P\left(\phi_{t,k}|\phi_{t-1,k}\right)$, 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 inversely proportional to the distance between the two Gaussian distributions corresponding to $\phi_{t,k}$ and $\phi_{t-1,k}$. Next we sample to update the state 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. After each epoch we update the cluster parameters by assigning the new instances.

6 Experiments

6.1 Data Sets & Settings

We apply our algorithm to two problems, (1) multi-pedestrian tracking, requiring the output of an off-line trained pedestrian detector as input and (2) generic multi-object tracking (Zhao, Gong, and Medioni 2012; Luo and Kim 2013; Luo et al. 2014), where multiple objects of any type are detected and tracked. For the first problem, we use the public *ETHMS* and *TUD Stadtmitte* data sets. For the second problem, we employ public six data sets from (Luo et al. 2014) named *Zebra, Crab, Goose, Hockey, Sailing*, and *Antelope*. We divide videos into epochs which are composed of approximate $50-200$ frames, depending on the length of the video. We set the dictionary dimension to 50 and η to 0.2 in all experiments. In the inference stage, for each epoch we run Gibbs sampling for 500 iterations and report results after the last iteration.

6.2 Metrics

To evaluate tracking performance we employ the metrics proposed in (Wu and Nevatia 2006). 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.

6.3 Results

The experiments are conducted in three parts. In the first part we compare our approach to existing sequential approaches (Luo and Kim 2013; Luo et al. 2014) in solving the generic multi-object tracking problem. The second part compares our algorithm with several state-of-the-art data association algorithms (Pirsiavash, Ramanan, and Fowlkes 2011; Xing, Ai, and Lao 2009) using the same detection results and visual representation. The third part compares the method to multi-pedestrian tracking approaches (Pellegrini et al. 2009; Zhang, Li, and Nevatia 2008; Milan, Schindler, and Roth 2013; Leal-Taixé et al. 2014).

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 (Luo and Kim 2013) and BLP (Luo et al. 2014). For fairness we use the same detection results as used in the methods that we compare with, allowing direct comparison of the association performance. The results are shown in Table 1. 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 (Luo and Kim 2013) and (Luo et al.

Table 2: **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	Ours	DA-H	DA-DP	DA-SSP	BL	Ours	DA-H	DA-DP	DA-SSP	BL	Ours	DA-H	DA-DP	DA-SSP	BL	Ours
Zebra	.59	.55	.54	.60	**.61**	**.25**	.35	.35	**.25**	**.25**	28	32	31	27	**26**	3	2	7	3	**1**
Crab	.24	.19	.19	**.25**	**.25**	**.69**	.70	.70	**.69**	**.69**	170	168	166	168	**163**	27	31	30	28	**15**
Antelope	**.75**	.63	.63	.72	.74	**.15**	.27	.27	**.15**	.16	36	33	**32**	37	**32**	14	10	10	25	**6**
Goose	**.79**	.64	.68	**.79**	**.79**	**.03**	.25	.32	.04	.04	34	31	29	25	**19**	25	20	18	14	**12**
Sailing	.83	.83	.83	.83	.83	.08	.08	.08	.08	.08	42	45	44	**40**	**40**	10	9	**8**	**8**	**8**
Hockey	.64	.54	.54	**.68**	**.68**	.14	.18	.18	**.11**	**.11**	12	11	**10**	12	**10**	11	7	6	6	**3**
Overall	.54	.47	.46	.54	**.55**	.34	.41	.42	**.33**	.34	322	320	312	309	**290**	90	79	73	84	**45**

2014), 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 (Xing, Ai, and Lao 2009), (2) DA-DP: dynamic programming in network flow (Pirsiavash, Ramanan, and Fowlkes 2011), (3) DA-SSP: successive shortest path in network flow (Pirsiavash, Ramanan, and Fowlkes 2011), (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. For fairness, all algorithms are given the same detection results from (Luo et al. 2014). The results of DA-DP and DA-SSP are obtained using the code from (Pirsiavash, Ramanan, and Fowlkes 2011).

Results in Table 2 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 achieves the best performance. Compared with BL, it further reduces the IDS and FM values.

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 (Milan, Schindler, and Roth 2013). We compare our results with those of the methods called SB (Pellegrini et al. 2009), GDANF (Zhang, Li, and Nevatia 2008), DTE (Milan, Schindler, and Roth 2013), and MC (Leal-Taixé et al. 2014). SB develops a sophisticated dynamic model based on social forces during association. GDANF casts data association as finding the min-cost in network flow. DTE adopts a CRF model for data association. The results are shown in Table 3. On the *TUD-Stadtmitte* dataset, 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 SB and GDANF, but worse than those of DTE and MC, which are all methods tailored to this task. 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 SB 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 GDANF 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 DTE. MC achieves the best MT and ML performance as a result of the contextual motion model, which is able to recover trajectory components, even in the case of missed detections, by learning a dictionary of interaction features among objects. In our method we only consider the simple but general super-pixel representation for appearance modeling. The super-pixel representation inevitably suffers ef-

Table 3: **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*				
	DTE	Ours	SB	GDANF	DTE	MC	Ours
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

fects from background clutter when 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. These models may be integrated into Eq. 6 to provide better likelihood functions.

7 Conclusion

This paper has introduced a topic model for the multi-object tracking problem. Thanks to the DPMM, 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 and the temporal damping are modeled by segmenting the video into temporal epochs. Experiments on public data sets show the advantages of our method over sequential solutions and other data association methods. Future work includes the incorporation of more sophisticated models of appearance, motion and context in order to improve the performance of specific applications such as pedestrian tracking.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Susstrunk, S. 2012. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI* 34(11):2274–2282.

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

Berclaz, J.; Fleuret, F.; Turetken, E.; and Fua, P. 2011. Multiple object tracking using k-shortest paths optimization. *PAMI* 33(9):1806–1819.

Blei, D. M., and Jordan, M. I. 2006. Variational inference for dirichlet process mixtures. *Bayesian analysis* 1(1):121–143.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Brendel, W.; Amer, M.; and Todorovic, S. 2011. Multiobject tracking as maximum weight independent set. In *CVPR*, 1273–1280.

Butt, A. A., and Collins, R. T. 2013. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *CVPR*, 1846–1853.

Cao, L., and Fei-Fei, L. 2007. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 1–8.

Dumais, S.; Furnas, G.; Landauer, T.; Deerwester, S.; Deerwester, S.; et al. 1995. Latent semantic indexing. In *Proceedings of the Text Retrieval Conference*.

Fei-Fei, L., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 524–531.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *PAMI* 32(9):1627–1645.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57.

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

Kuo, C.-H.; Huang, C.; and Nevatia, R. 2010. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, 685–692.

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

Luo, W., and Kim, T.-K. 2013. Generic object crowd tracking by multi-task learning. In *BMVC*.

Luo, W.; Kim, T.-K.; Stenger, B.; Zhao, X.; and Cipolla, R. 2014. Bi-label propagation for generic multiple object tracking. In *CVPR*.

Luo, W.; Zhao, X.; and Kim, T.-K. 2014. Multiple object tracking: A review. *arXiv:1409.7618*.

Milan, A.; Schindler, K.; and Roth, S. 2013. Detection-and trajectory-level exclusion in multiple object tracking. In *CVPR*, 3682–3689.

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

Pirsiavash, H.; Ramanan, D.; and Fowlkes, C. C. 2011. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 1201–1208.

Shu, G.; Dehghan, A.; Oreifej, O.; Hand, E.; and Shah, M. 2012. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 1815–1821.

Teh, Y. W. 2010. Dirichlet process. In *Encyclopedia of machine learning*. 280–287.

Verbeek, J., and Triggs, B. 2007. Region classification with markov field aspect models. In *CVPR*, 1–8.

Wang, X., and Grimson, E. 2008. Spatial latent dirichlet allocation. In *NIPS*, 1577–1584.

Wang, X.; Ma, K. T.; Ng, G.-W.; and Grimson, W. E. L. 2011. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *IJCV* 95(3):287–312.

Wang, C.; Blei, D.; and Li, F.-F. 2009. Simultaneous image classification and annotation. In *CVPR*, 1903–1910.

Wu, B., and Nevatia, R. 2006. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR*, 951–958.

Xing, J.; Ai, H.; and Lao, S. 2009. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*, 1200–1207.

Yamaguchi, K.; Berg, A. C.; Ortiz, L. E.; and Berg, T. L. 2011. Who are you with and where are you going? In *CVPR*, 1345–1352.

Yang, B., and Nevatia, R. 2012. An online learned crf model for multi-target tracking. In *CVPR*, 2034–2041.

Zhang, L.; Li, Y.; and Nevatia, R. 2008. Global data association for multi-object tracking using network flows. In *CVPR*, 1–8.

Zhao, X.; Gong, D.; and Medioni, G. 2012. Tracking using motion patterns for very crowded scenes. In *ECCV*. 315–328.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2005. Time-sensitive dirichlet process mixture models. Technical Report CMU-CALD-05-104.