A CLUSTER SPECIFIC LATENT DIRICHLET ALLOCATION MODEL FOR TRAJECTORY CLUSTERING IN CROWDED VIDEOS

*Jialing Zou, Yanting Cui, Fang Wan, Qixiang Ye, Jianbin Jiao**

University of Chinese Academy of Sciences, Beijing 101408, China.
*jiaojb@ucas.ac.cn

ABSTRACT

Trajectory analysis in crowded video scenes is challenging as trajectories obtained by existing tracking algorithms are often fragmented. In this paper, we propose a new approach to do trajectory inference and clustering on fragmented trajectories, by exploring a cluster specif c Latent Dirichlet Allocation(CLDA) model. LDA models are widely used to learn middle level trajectory features and perform trajectory inference. However, they often require scene priors in the learning or inference process. Our cluster specif c LDA model addresses this issue by using manifold based clustering as initialization and iterative statistical inference as optimization. The output middle level features of CLDA are input to a clustering algorithm to obtain trajectory clusters. Experiments on a public dataset show the effectiveness of our approach.

Index Terms— Latent Dirichlet Allocation, Manifold, Trajectory clustering

1. INTRODUCTION

Trajectory clustering is a video analysis task whose goal is to assign individual trajectories with common cluster labels, with applications in activity surveillance, traff c f ow estimation and emergency response [1], [2].

In straightforward trajectory clustering approaches, a set of features, i.e. coordinates, velocities and/or geometrical shapes and scene specif c information, are extracted to represent trajectories, and then unsupervised learning methods are used to classify these features [1], [2], [3]. However, in videos of crowded scenes, it is often diff cult to obtain complete trajectories with off-the-shelf tracking algorithms [4]. In most cases, fragmented trajectories of different length are obtained, which are diff cult to be aligned and represented with low level features.

Recently, middle level feature based trajectory clustering approaches have attracted attentions. Middle level features are usually observed as dominant paths of moving objects, which can bridge the fragmented trajectories in low level feature space with their clusters [5]. Trajectory middle level features can be learned with non-Bayesian approaches, for example, similarity clustering [6] or dimension reduction [7]. In [6], Wang et al. propose two Euclidean similarity measures, and perform trajectories with the def ned measures. In [7], Hu et al. introduce a dimensional spectral clustering method. Trajectories are projected to a lower space through eigenvalue factorization, and are clustered in the lower sub-space with a k-means algorithm.

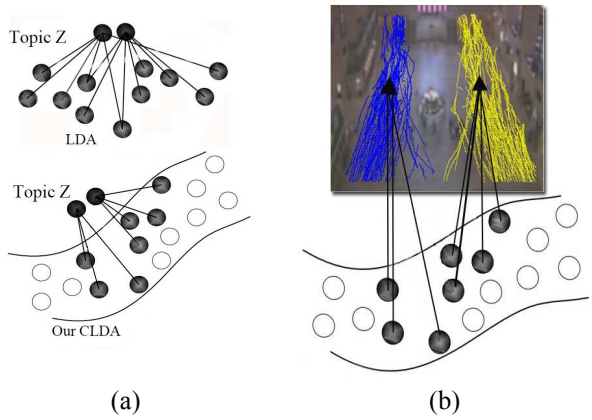


Fig. 1. (a) Illustration of the LDA based topic extraction in original feature space (up) and a manifold embedding space (down). (b) Illustration of correspondence between trajectory clusters and the manifold embedding.

Middle level features can also be learned with hierarchical latent variable Bayesian models, such as latent Dirichilet allocation (LDA) [8] models, which have been widely explored in classif cation and clustering tasks [9], [10], [11], [12], [13], [14]. These models are adopted from the natural language processing and are well known as ”topic models”. With LDA related models, trajectories are treated as documents and observations of trajectories are treated as visual words. Learned topics correspond to the middle level features of trajectories.

In [12], Li et al. proposed a theme model that poses class supervision to LDA, and enables different classes have different topics. In [13], Daniel et.al. proposed a labeled LDA that assigns each class a Dirichlet prior. In [11], Wang et al. proposed a topic-supervision LDA (ts-LDA) which enables different classes share different topics. In [14] Wang et al. proposed a semi-latent LDA, which enables both the latent labels and words are visible in the training process. In [10],

Wang et al. proposed a mixture of latent Dirichlet allocation (LDA) models for trajectory clustering. In [9], Zhou et al. proposed a Random Field Topic (RFT) model for trajectory clustering by integrating scene priors.

Despite of the effectiveness of above LDA models, however, most of them [10, 11, 12, 13] ignore the distribution of data, so they often require a complex parameter estimation and variable inference procedure. Some of them [9] use scene priors to improve performance, but can only be used in situations where priors are available. Some of them [10, 12] use cluster initialization to replace scene priors, however, to do a good initialization in a high dimensional feature space is often diff cult.

We propose a cluster-specif c LDA (CLDA) model with cluster initialization in a manifold embedding space and iterative optimization with Bayesian inference. As shown in Fig.1(a), the cluster initialization enables that our approach create topics (middle level features) that can ref ect data distribution and cluster information, effectively. After the iterative optimization, the CLDA model creates discriminative topics for clustering without using any scene prior. Here, "discriminative" implies that different sets of trajectories have different clusters in the manifold space, as shown in Fig.1(b).

The remainder of the paper is organized as follows: The CLDA model is described in Section 2. The trajectory clustering approach is presented in section 3. Experiments are presented in Section 4 and we conclude the paper in Section 5.

2. CLUSTER-SPECIFIC LATENT DIRICHLET ALLOCATION

This section presents the CLDA and the middle level feature (topic) extraction with the CLDA parameter estimation.

2.1. Latent Dirichlet Allocation (LDA)

To make the paper self-contained, we f rst review the LDA. Fig. 2(a) shows the graphical representation of LDA [8]. Four important notations about LDA are corpus, document, topic and word, which in our case correspond to path, trajectory, topic and visual word, respectively. M is number of documents. N_j is the number of words of each trajectory j. θ is assumed to follow a Dirichlet distribution with parameter α. z_{ji} is a latent variable being assumed to follow a parameterized multinomial distribution $Mult(\theta_j)$. x denotes words. β is hyper-parameter, corresponding to the middle level features.

2.2. Cluster-specif c Dirichlet Allocation (CLDA)

Fig.2(b) shows the graphical representation of CLDA. It can be seen that observed visual words (low level features) and labels are inputs of the CLDA. The visual words that are coordinates and velocities of tracked objects are described in Section 3. The joint probability of the model is given by Eq.(1), the θ probability is changed to $p(\theta|c,\alpha) = \prod_{j=1}^{C} Dir(\theta|\alpha_j)^{\delta(c,j)}$, where $Dir(\bullet)$ denotes the Dirichlet distribution based on parameter α_j.

$$p(x,z,\theta,c|\alpha,\beta,\eta) = p(c|\eta)p(\theta|\alpha,c)\cdot \prod_{j=1}^{M} p(z_j|\theta_j)p(x_j|z_j,\beta). \quad (1)$$

Other terms are consistent with that in [8]. Similar to [12], we set the uniform distribution $p(c) = 1/C$. And then leave out the estimation of η. The label C is updated by $\arg\max_c p(x|c,\alpha,\beta)$, where x denotes a set of words of a trajectory. The posterior probability of x is in Eq. (2).

$$p(x|\alpha,\beta,c) = \int p(\theta|\alpha,c) \left(\prod_{j}^{M} \sum_{z_j} p(z_j|\theta_j)p(x_j|z_j,\beta)\right) d\theta. \quad (2)$$

We use the variational breaking algorithm in [8] to do variable inference and parameter estimation. Fig.2(c) is the graphical representation of the approximate distribution of the CLDA. Therefore, we have

$$\log p(x|\alpha,\beta,c) = L(\gamma_c,\phi_c;\alpha_c,\beta) + KL(q(\theta,z|\gamma_c,\phi_c)||p(\theta,z|x,\alpha_c,\beta)). \quad (3)$$

We iteratively maximize the the term $L(\bullet)$ instead of $\log p(x|\alpha,\beta,c)$, which results in the minimum of difference between the distribution in Fig. 2(b) and Fig. 2(c). The details of computation can be seen in [8]. In Eq. (4) and Eq. (5), we present two terms related to middle level features.

$$\phi_{ki}^{c} \propto \beta \exp\left[\Psi(\gamma_k^c) - \Psi(\sum_{k=1}^{K}\gamma_k^c)\right] \quad (4)$$

$$\beta_k \propto \sum_i \phi_{k,i}^{c} n_i \quad (5)$$

where ϕ_{ki}^{c} denotes the probability that the ith word belongs to the kth topic's. $\Psi(\bullet)$ is a digamma function and n_i is the count of the ith word in the codebook. β_k is the kth topic's feature representation with respect to the codebook.

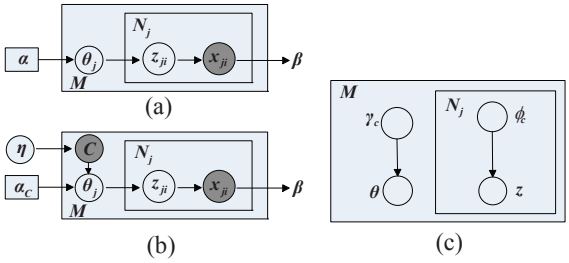


Fig. 2. LDA and CLDA models. (a) Graphical representation of LDA, (b) Graphical representation of CLDA, (c) Graphical representation of approximate distribution of CLDA.

3. TRAJECTORY CLUSTERING

In trajectory clustering, we f rst connect trajectory fragments into trajectory trees and extract low level features, which are

embedded in a dimensional manifold space for initial trajectory clustering. With initial cluster labels, the middle level features are extracted by the proposed CLDA for f ne trajectory clustering. The f owchart of the proposed approach is shown in Fig.3.

3.1. Low-level Feature Extraction

Given a crowded video, we f rst use a KLT tracker [15] to calculate trajectory segments and motion vectors of objects. A spanning tree algorithm [9] is used to uncover spatiotemporal relations among trajectory fragments and connect them into trajectory trees, on which visual words are extracted as low level features.

To extract low level features (visual words), we construct a codebook for each video scene. The scene image is divided into cells of 10*10 pixels, and the velocity of each trajectory point is quantized into 5 bins, as $v \in \{0, 1, 2, 3, 4\}$. Given scene video resolution of $W * H$, the size of the codebook is set to $(W/10) * (H/10) * 5$. With the codebook, we compute a word for each trajectory point with $word = v * (H/10 * W/10) + (x/10) * (H/10) + (y/10)$, where (x,y) is the coordinate, and v is the velocity bin. With the words for all trajectory points, each trajectory tree is represented with a $bag\text{-}of\text{-}words$ [8].

3.2. Initial Clustering

We use the Laplacian Eigenmap method (LE) [16] for trajectory manifold embedding, and empirically set the dimensionality of the manifold space to 4. The goal in LE is to minimize the term $\sum_{i,j} (y_i - y_j)^2 w_{ij}$, where y denotes the coordinates of trajectories in the low dimensional feature space, and w_{ij} denotes the neighbor weights. w_{ij} is 1 if trajectory i, j have the neighbor relations, otherwise, w_{ij} is set as 0. Trajectory's neighbors are found through a k-near-neighbors (KNN) method. We reduce the dimensions of trajectories with a optimization procedure in [16]. Through the optimization, trajectory data will be projected to a lower dimensional manifold embedding space where the cluster information (neighboring relations) is reserved.

In the lower dimensional manifold embedding space, a k-means algorithm is adopted to perform initial trajectory clustering and obtain initial trajectory cluster labels.

3.3. Fine Clustering

Using initial labels and low level features as inputs, the middle level features could be extracted by the proposed CLDA, as in section 2.2. In the procedure, topic probabilities of trajectory trees in the topic space are computed. Given K topics, each tree has K corresponding probabilities. The topic label of the largest probability is assigned to the tree. After this step, each trajectory tree has a topic label. The topic labels

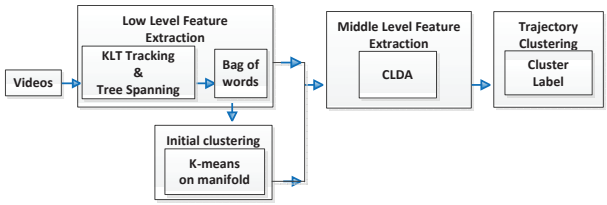


Fig. 3. Flowchart of the proposed trajectory clustering approach.

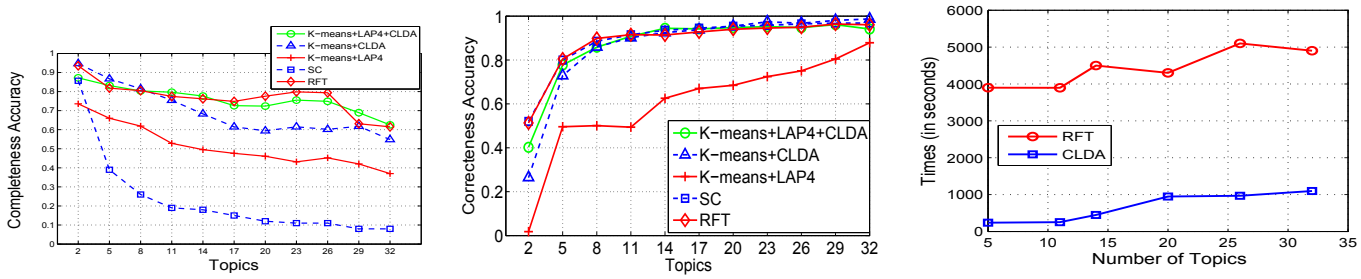
will be input to the information entropy clustering algorithm [9], which computes potential information entropy of trees that a trajectory segment belongs to, and then determines the cluster label of the trajectory segment.

4. EXPERIMENTS

Experiments are conducted on the video dataset collected from the crowded New York's Grand Central station [9]. The video has a resolution of 720*480 pixels, and has a temporal length of 1800 seconds. On the video, a KLT tracking algorithm is used to calculate 47866 trajectories, which have the average temporal length of 133 frames. According to the low level feature (visual word) calculation procedure in section 3, the codebook size is 72*48*5, so the low level feature dimensionality is 17280.

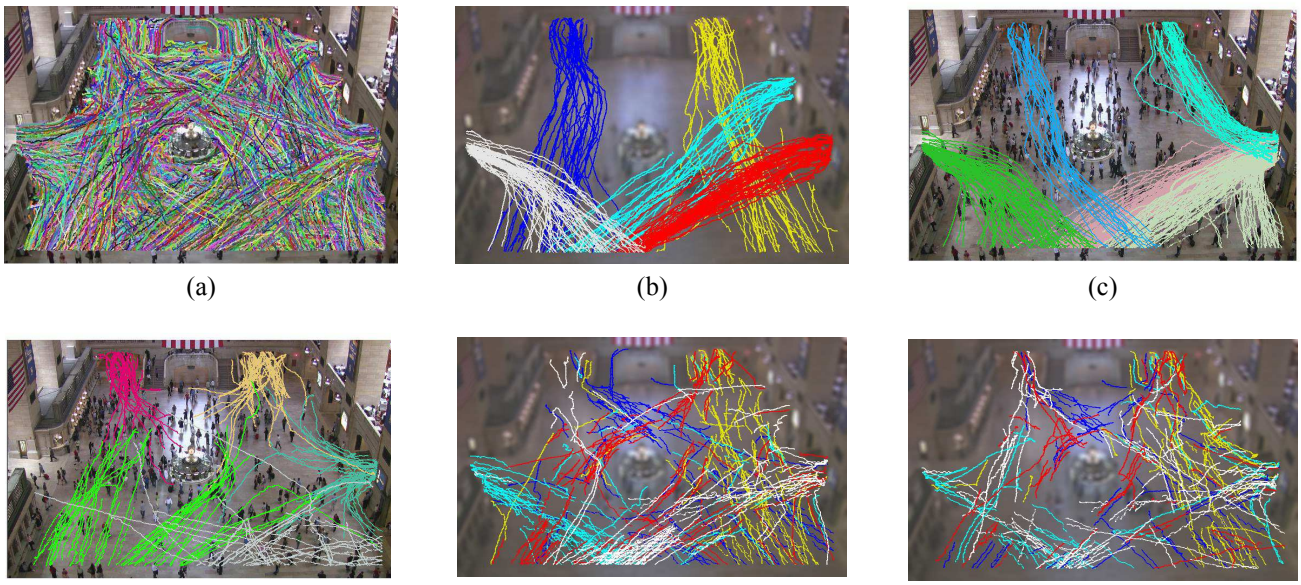
Correctness and completeness [9], [17] are used to evaluate the trajectory clustering performance. Completeness measures accurately the trajectories from the same clusters are clustered together. Correctness measures the trajectories from the different clusters are divided. If all trajectories are clustered into one single cluster, the completeness is 100% and the correctness is 0%, and vice versa. The labeled trajectories in [9] are used as ground truth, in which 1507 pairs for completeness and 2000 pairs for correctness. In Fig.4(a) and Fig.4(b), we compare our approach to the Spectral Clustering (SC) approach [7] and the Random Field Topic (RFT) based approach [9]. When implementing the approach SC, We use a linear interpolation to align the trajectories and measure the similarities with the Euclidean distance.

Fig.4(a) shows the comparison of completeness performance. "LAP4" denotes the 4 dimensional manifold embedding. It can be seen that the "SC" approach that does not use dimension reduction or LDA for middle level feature extraction has a low performance. This is because in crowded video scenes trajectories are overlapped with each other and are fragmented. Therefore, direct clustering in low level feature space is diff cult. Without manifold embedding, the approach "k-means+CLDA" has a lower completeness performance than our approach ("k-means+LAP4+CLDA") with manifold embedding. It can also be seen that our approach (without any scene prior) has slightly better performance than the RFT model (with scene priors) [9]. This demonstrates the effectiveness of our proposed cluster specif c LDA model.



(a) (b) (c)

Fig. 4. Performance and comparison of trajectory clustering approaches. (a) Completeness performance, (b) Correctness performance, (c) Model learning time.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Comparison of trajectory clustering results. (a)original trajectory segments, (b) "k-means+LAP4+CLDA" approach, (c) "RFT" approach, (d) "SC" approach, (e) "K-means+CLDA" approach, (f) "K-means+LAP4" approach.

Fig.4(b) shows comparison of correctness performance from five approaches. Except the "k-means+LAP4" approach, other four approaches have similar correctness performance. The correctness performance of the "k-means+CLDA" approach is high because CLDA still can calculate the cluster labels through statistical inference. However, in Fig.4(a), the completeness performance of the "k-means+CLDA" approach is low. This is for the reason that trajectory data in the high dimensional low level features space is very sparse, which makes it difficult to directly perform clustering using a k-means clustering algorithm.

Fig.4(c) compares the learning time for RFT and the CLDA model under different topics. It can be seen that the learning procedure of our proposed CLDA model is much faster than that of the RFT model.

In Fig.5, we visualize some trajectory clusters by different approaches. It can be seen that our approach can calculate clear trajectory paths and clusters.

5. CONCLUSION

We have proposed a cluster specific latent Dirichlet allocation (CLDA) to learn trajectory middle level features for trajectory inference and clustering. Using manifold based clustering as initialization, the proposed CLDA could be used to perform trajectory inference and calculate trajectory middle level features (topics) without using scene priors. We also proposed a trajectory clustering approach based on the CLDA model, and validated the effectiveness of the approach and compared it with recent approaches. Experiments and comparisons show that our approach has a comparable performance to the scene priors based RFT model. In addition, our approach has a higher learning speed.

6. ACKNOWLEDGMENT

This work is supported in Part by National Basic Research Program of China (973 Program) with Nos. 2011CB706900, 2010CB731800, and National Science Foundation of China with Nos. 61039003, 61271433 and 61202323.

7. REFERENCES

[1] Brendan Tran Morris and Mohan M Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, 2008.

[2] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, 2004.

[3] Brendan Morris and Mohan Trivedi, "Learning trajectory patterns by clustering: Experimental studies and comparative evaluation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 312–319, 2009.

[4] Alper Yilmaz, Omar Javed, and Mubarak Shah, "Object tracking: A survey," *Acm Computing Surveys*, vol. 38, no. 4, pp. 13, 2006.

[5] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering.," *In Advances in Neural Information Proccessing Systems*, vol. 14, pp. 585–591, 2001.

[6] Xiaogang Wang, Kinh Tieu, and Eric Grimson, "Learning semantic scene models by trajectory analysis," *Proc. European Conf. Computer Vision*, pp. 110–123, 2006.

[7] Weiming Hu, Dan Xie, Zhouyu Fu, Wenrong Zeng, and Steve Maybank, "Semantic-based surveillance video retrieval," *IEEE Trans. Image Processing*, vol. 16, no. 4, pp. 1168–1181, 2007.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *the Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang, "Random f eld topic model for semantic region analysis in crowded scenes from tracklets," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3441–3448, 2011.

[10] Xiaogang Wang, Xiaoxu Ma, and W Eric L Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[11] Chong Wang, David Blei, and Fei-Fei Li, "Simultaneous image classif cation and annotation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1903–1910, 2009.

[12] Li Fei-Fei and Pietro Perona, "A bayesian hierarchical model for learning natural scene categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 524–531, 2005.

[13] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 248–256, 2009.

[14] Yang Wang, Payam Sabzmeydani, and Greg Mori, "Semi-latent dirichlet allocation: A hierarchical model for human action recognition," *Human Motion–Understanding, Modeling, Capture and Animation*, pp. 240–254, 2007.

[15] Tomasi Carlo and Takeo Kanade, "Detection and tracking of point features," *Int'l Journal of Computer*, 1991.

[16] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[17] Bart Moberts, Anna Vilanova, and Jarke J van Wijk, "Evaluation of f ber clustering methods for diffusion tensor imaging," *Visualization, IEEE*, pp. 65–72, 2005.