SCIENCE CHINA
Information Sciences

• MOOP •

November 2017, Vol. 60 113101:1–113101:7
doi: 10.1007/s11432-017-9237-1

Patch-based topic model for group detection

Mulin CHEN[1], Qi WANG[1,2*] & Xuelong LI[3]

[1]*School of Computer Science and Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an* 710072*, China;*
[2]*Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an* 710072*, China;*
[3]*Center for Optical Imagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an* 710119*, China*

Received July 12, 2017; accepted September 19, 2017; published online October 10, 2017

Abstract Pedestrians in crowd scenes tend to connect with each other and form coherent groups. In order to investigate the collective behaviors in crowds, plenty of studies have been conducted on group detection. However, most of the existing methods are limited to discover the underlying semantic priors of individuals. By segmenting the crowd image into patches, this paper proposes the Patch-based Topic Model (PTM) for group detection. The main contributions of this study are threefold: (1) the crowd dynamics are represented by patch-level descriptor, which provides a macroscopic-level representation; (2) the semantic topic label of each patch are inferred by integrating the Latent Dirichlet Allocation (LDA) model and the Markov Random Fields (MRF); (3) the optimal group number is determined automatically with an intro-class distance evaluation criterion. Experimental results on real-world crowd videos demonstrate the superior performance of the proposed method over the state-of-the-arts.

Keywords group detection, collective behavior, crowd analysis, latent topic

Citation Chen M L, Wang Q, Li X L. Patch-based topic model for group detection. Sci China Inf Sci, 2017, 60(11): 113101, doi: 10.1007/s11432-017-9237-1

1 Introduction

In crowd scenes, people usually interact with the surroundings and group together. Individuals within the same group exhibit collective behaviors, and share similar motion patterns. Since groups provides a mid-level understanding about the crowd phenomenon, group detection has been an active research area in computer vision, and involves a lot of practical applications, such as crowd counting [1], crowd tracking [2] and anomaly detection [3]. Though many approaches have been proposed in recent years, group detection remains to be a difficult task due to the complex nature of collective behaviors.

One limitation shared by existing studies is the locality of study object. To analyze the crowd behavior, it is fundamental to extract the individuals in crowd scenes. Since detection and tracking algorithms are inapplicable in crowd scenes, previous studies mostly treat particles [4,5] or feature points [6–11] as study objects, and model their velocities directly to detect groups. But both the particles and feature points are too microcosmic to reflect the global crowd motion. Moreover, their velocities may fluctuate dramatically

* Corresponding author (email: crabwq@nwpu.edu.cn)

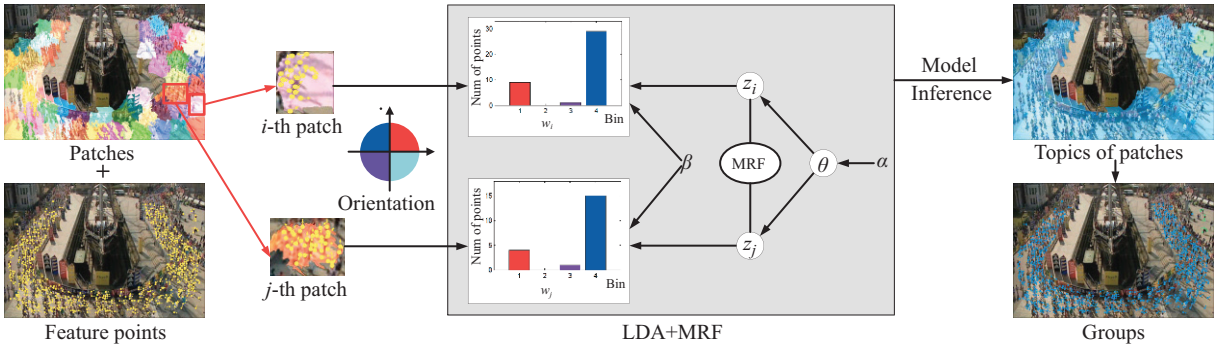

Figure 1 (Color online) Pipeline of the proposed PTM. First, the patch-level descriptor is constructed with the distribution of the feature points over the orientation space. Then, the obtained descriptor is fed into the LDA model, and an MRF prior is imposed on the hidden priors to enforce the spatial coherence. After model inference, the semantic motion prior within each patch is learned. Finally, the feature points are combined according to the prior of the corresponding patch. Scatters with different colors indicate different detected groups.

because of the locality property. So it is essential to perceive the motion dynamics at the macroscopic level.

Another difficulty in group detection is the exploration of underlying semantic motion prior. Instead of random occurrence, the movement of each individual is driven by a motion prior, which can also be interpreted as the moving intention [10]. In the same group, individuals have similar behaviors, which implies the fact that they share the same semantic prior. Thus, the investigation of the underlying prior could facilitate group detection. However, many existing studies [4–8, 11] only emphasize the observed movements of individuals, and neglect the underlying prior.

In addition, the decision of group number is also a barrier to detect groups. Generally speaking, group detection can be considered as the clustering of individuals. But unlike standard clustering problem, the desired group number is unknown in the crowd analysis literature. Some methods [6–8, 10] simply combine the individuals whose similarities are larger than a fixed threshold. But it is impractical to choose a threshold that suitable for crowds with various densities and structures. So it is necessary to decide the group number automatically.

In this paper, a new group detection method, namely Patch-based Topic Model (PTM) is developed to tackle the above issues. Our main contributions are summarized as follows.

(1) A patch-level descriptor is developed to represent crowd motion from the macroscopic aspect. Thus, the proposed method is robust to the fluctuation of feature points.

(2) The semantic priors of crowd motions are deeply exploited by the Latent Dirichlet Allocation (LDA) [12] model, and the Markov Random Field (MRF) is introduced to enforce the spatial coherence.

(3) The optimal group number is decided automatically with an intra-class distance evaluation criterion. Then the proposed method is able to handle crowds without various densities.

The rest of this paper is organized as follows. Section 2 introduces the proposed PTM method. Section 3 presents the experimental results. The conclusion and future work are given in Section 4.

2 PTM for group detection

In this section, the PTM for group detection is introduced. First, the crowd image is segmented into patches, and the feature points in the crowd scenes are extracted. Then, LDA model and MRF are jointly combined to learn the semantic motion prior of each patch. With the learned priors, each feature points can be assigned with a group label according to the patch it belongs to. Finally, the intra-class distance (ID) is utilized to decide the optimal group number and produce the final groups. The pipeline of the proposed model is shown in Figure 1.

2.1 Patch-level descriptor

Due to the difficulty to identify pedestrians in crowd scenes, existing approaches mostly treat particles or feature points as individuals directly. However, as mentioned above, the local particles or feature points cannot reveal the real crowd motion. So we propose to represent the crowd movements at the patch-level.

First, image segmentation technique is employed to divide the crowd scene into patches. In this work, the SLIC algorithm is used since it can produce compact image patches efficiently. And each crowd scene is segmented into 100 patches. Then, feature points are detected and tracked with a generalized Kandae-Lucas-Tomasi (gKLT) tracker [8]. And we remove the patches where no feature points are detected. Finally, the orientation space is divided into four directions, and the patch-level descriptor of each patch is defined as the distribution of its corresponding feature points over the divided orientation space, as illustrated in Figure 1.

2.2 Latent topic model

To learn the motion prior within each segmented patch, LDA is utilized. LDA is a classical topic model to discover the hidden topics from text corpus. Given the corpus and the desired topic number, LDA learns the topic distribution over document according to the words of each document. In the past decades, LDA has shown dominant performance in natural language processing [13], and image segmentation [14,15]. In this study, we treat the patches as documents, and the crowd image as corpus. The patch-level descriptor is regarded as the words. Then we aim to learn the topic of each patch, which can be understood as the semantic motion prior.

Here we describe the generative process of LDA, as shown in Figure 1. Supposing there are N documents in the corpus, the topic number is K and the size of the vocabulary is W (in this work, W is 4 since the orientation space is divided into four directions), $\alpha \in \mathbb{R}^{K\times 1}$, $\beta \in \mathbb{R}^{K\times W}$ are hyper-parameters for Dirichlet distribution and multinomial distribution respectively, $\theta \in \mathbb{R}^{K\times 1}$ is a multinomial vector, z_i is the topic of document i, and w_i is the word sequence of i. For a corpus, LDA first draws the distribution of θ_{z_i} according to the Dirichlet distribution parameterized by α. Then for each document i, the hidden topic z_i is produced with a multinomial distribution parameterized by θ_{z_i}. Finally, the word sequence w_i is chosen from the vocabulary with a multinomial distribution parameterized by β and conditioned on the topic z_i, i.e., $p(w_i|z_i,\beta)$. The joint distribution of $\theta, w = \{w_i\}_{i=1}^N, z = \{z_i\}_{i=1}^N$, α and β is

$$\begin{aligned} &p(\theta, w, z|\alpha, \beta) = p(\theta|\alpha)p(z|\theta)p(w|z,\beta),\\ &p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K (\theta_k)^{\alpha_k - 1},\\ &p(z|\theta) = \prod_{i=1}^N \theta_{z_i},\\ &p(w|z,\beta) = \prod_{i=1}^N \beta_{z_i,w_i}, \end{aligned} \tag{1}$$

where $\Gamma()$ is the Gamma distribution.

The above traditional LDA is suitable for handling text data. But in the field of image processing, the spatial correlation between the patches should also be emphasized. Inspired by the development on image segmentation [14, 15], we jointly incorporate MRF into LDA to enforce the spatial coherence. Instead of drawing z_i from the multinomial distribution directly, a MRF prior is placed on the hidden topics:

$$p(z|\theta,\sigma) \propto \exp\left(\sum_{i=1}^N \log\theta_{z_i} + \sigma \sum_{i\sim j} \delta(z_i = z_j)\right), \tag{2}$$

where $i\sim j$ means that patch i is adjacent to patch j, δ is the indicator function, σ controls the balance of the fitting term (first term) and the smooth term (second term). The smooth term encourages the adjacent patches to share the same topic. Thus the spatial coherence is achieved. And the joint distribution is

rewritten as

$$p(\theta, w, z|\alpha, \beta, \sigma) \propto \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} (\theta_k)^{\alpha_k - 1} \exp\left(\sum_{i=1}^{N} \log \theta_{z_i} + \sigma \sum_{i \sim j} \delta(z_i = z_j) \right) \prod_{i=1}^{N} \beta_{z_i, w_i}. \quad (3)$$

And the optimal topics are learned with

$$z^* = \arg\max_{z} p(w|z). \quad (4)$$

The optimal conjunction of α, θ, β, z and σ can be efficiently solved by the variational inference method [14]. Readers can refer to [14] for the details.

With the obtained topics, each feature point can be assigned with a prior label according to the patch it belongs to, as shown in Figure 1. Then each pair of points are clustered into the same group if the corresponding patches share the same topic.

2.3 Decision of optimal groups

The above model captures the latent topic of each patch, which is also the motion prior. However, it needs the number of topics to be given, which is difficult in real-world applications. So we utilize the ID evaluation criterion [16] to determine the optimal topic number, then the group number can be also decided automatically.

First, we set the topic number K to be eight discrete values $(1, 2, \ldots, 8)$, and learn the latent topics with different K. After that, the feature points are clustered into K groups. Then we build a similarity graph S on the feature points as follows:

$$S(m, n) = \begin{cases} 1, & \text{if } \cos(\boldsymbol{v}_m, \boldsymbol{v}_n) \times \mathrm{e}^{-\lambda(\boldsymbol{p}_m - \boldsymbol{p}_n)^2} > \beta, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

where $\boldsymbol{v}_m$ and $\boldsymbol{p}_m$ are the velocity and spatial location of point m, cos() indicates the cosine similarity. λ is a parameter, and β is a threshold.

With the similarity graph, ID is denoted as

$$\mathrm{ID} = \frac{1}{\sum_{c=1}^{C} \frac{N_c(N_c-1)}{2}} \sum_{c=1}^{C} \sum_{m,n \in \text{Group } c} 1 - S(m, n), \quad (6)$$

where C is the total number of detected groups, N_c is the number of feature points in group c. A small value of ID indicates that the points within each group are with consistent velocities and small spatial distances. Then the topic number K with the smallest ID is chosen as the optimal K, and the corresponding groups are the final groups. The details of our approach is described in Algorithm 1.

Algorithm 1 Algorithm of the proposed method

Input: Image patches, feature points, parameter λ, threshold β

```
for K = 1, 2, ..., 8
    Compute the patch-level descriptor;
    Learn the topic of each patch with K;
    Combine the patches with the same topic and obtain groups;
    Calculate the ID of groups;
end
Set the groups with smallest ID as the final result.
```

Output: Detected groups

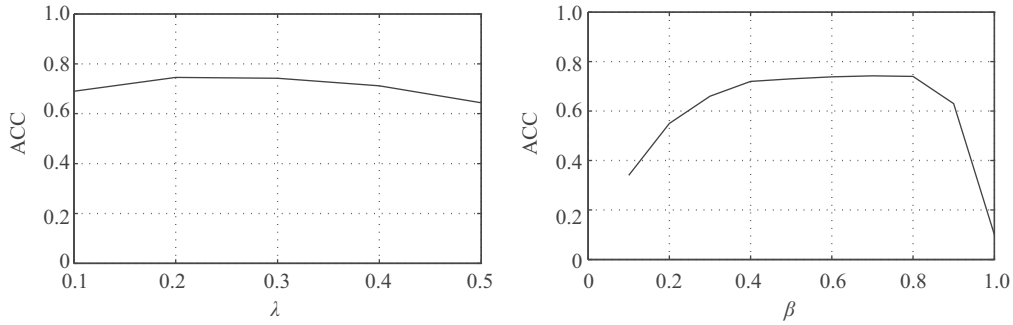

Figure 2 ACC curves of our method under different λ and β values.

3 Experiments

In this section, extensive experiments are conducted on a real-world crowd dataset to evaluate the proposed PTM. Throughout the experiments, all the competitors take their optimal parameters. The methods are compared on two aspects: group detection and group number estimation.

Dataset. Experiments are conducted on CUHK Crowd Dataset [7]. The dataset consists of 474 crowd videos from real-world scenes. It gives the spatial locations and velocities of feature points, and the group label of each point is also annotated by human observers. For selecting the best parameters, we randomly choose 100 videos, and use the 30 frames of each selected video as the training set. And all the rest frames are taken as testing test.

Selection of parameters. There are two parameters, including λ and β, in our method. Figure 2 plots the group detection accuracy (ACC) [9] of our method under different λ and β. λ governs the scale of the detected group. When λ is large, a united group may be mistakenly divided into parts. And when λ is too small, the far away points will be similar, and some noise will be included. As shown in Figure 2, the performance is relatively good when λ is 0.2, so we set $\lambda = 0.2$ in the experiments. In addition, the threshold β also controls the similarity of points. When β is too large, there will be no similar points, leading to the over-segmentation. And when it is too small, all the points will be included in the same group. We finally choose β as 0.8 in this work.

Competitors. To evaluate the performance of the proposed PTM, three state-of-the-art methods are taken for comparison, they are Coherent Filetering (CF) [6], Collective Transition (CT) [7] and Measuring Crowd Collectiveness (MCC) [8].

Performance on group detection. Group detection is performed on every video, and the average results are reported. Three widely used measurements, the ACC [9], Purity and Rand Index (RI) [10] are taken to evaluate the performance quantitatively. The experimental results are exhibited in Table 1. It is manifest that the proposed PTM has the highest ACC, Purity and RI, which indicates the best performance. CF finds the invariant neighbors of a feature point, and clusters them into the same group if their motion correlation is large. CT refines the results of CF by removing the points with large transition errors. MCC exploits the topological similarities of points, and combines those with high similarity. The above three methods take the feature points as study objects directly, so they are sensitive to the fluctuation of feature points. In addition, all of them neglect the underlying motion prior in crowd motion, and cannot decide the optimal group number automatically according to the crowd densities. The proposed PTM deeply investigates the semantic motion prior within each patches, and utilizes the intro-class distance criterion to decide the desired number of groups. So it does not has the above drawbacks and shows satisfying performance. Examples of the ground truth and the results of different methods are shown in Figure 3. The results of PTM are close to the ground truth.

Performance on group number estimation. In this part, we evaluate the performance the proposed PTM on group number estimation. Two standard metrics, averaging difference (AD) and variance (VAR) [11] are used as measurements. A lower value of AD indicates the less deviation from the ground truth, and a lower VAR means a higher stability on the group number estimation. The AD

Table 1 Results of group detection methods. The best results are in bold face

	PTM	CF	CT	MCC
ACC	**0.7824**	0.7034	0.7523	0.6810
Purity	**0.8748**	0.7331	0.7762	0.8472
RI	**0.8534**	0.7821	0.8343	0.7442

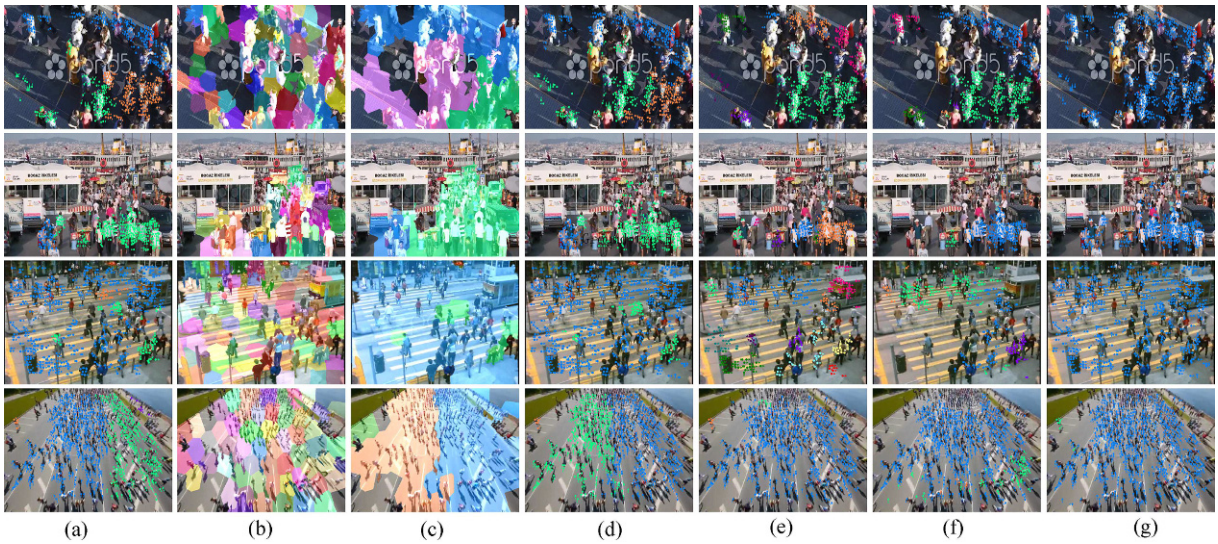

Figure 3 (Color online) Representative results of group detection. (a) Ground truth; (b) segmented patches; (c) topics learned by the proposed model. Patches with the same topic are visualized with the same color. (d)–(g) group detection results of the proposed PTM, CF, CT and MCC. Scatters with different colors indicate different detected groups. It can be seen that our method achieves the consistent results with the ground truth.

Table 2 Quantitative comparison on group number estimation. The best results are in bold face

	PTM	CF	CT	MCC
AD	**1.21**	2.45	1.63	1.59
VAR	**1.42**	3.01	1.83	1.874

and VAR of different methods are shown in Table 2. It can be seen that the proposed PTM achieves the lowest VAR. CF fails because it cannot capture the subtle difference of points' movements. Both CT and MCC threshold the points' similarity to detect groups, but it is unreliable to find a suitable threshold for the crowds with various densities and structures. The proposed PTM finds the optimal group number according to the ID of detected groups, so it is able to estimate the group number correctly.

4 Conclusion

In this study, a new PTM is put forward for group detection. The feature points are firstly assembled into patches, which represent the crowd motion at the macroscopic level. Then LDA and MRF are jointly integrated to explore the semantic motion prior within each patch, based on which the coherent points are identified. Finally, the intra-class distance evaluation criterion is used to find the optimal groups. Extensive experiments on real-world videos shows that our method achieves comparative performance against the state-of-the-arts.

In the future work, we plan to utilize semantic motion prior into some specific applications on crowd surveillance, such as anomaly detection and activity recognition. Moreover, it is also desirable to design more powerful descriptors to quantify the complicated crowd behaviors.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2017YFB1002202), National Natural Science Foundation of China (Grant Nos. 61773316, 61379094),

Fundamental Research Funds for the Central Universities (Grant No. 3102017AX010), and Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

Conflict of interest The authors declare that they have no conflict of interest.

Supporting information The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhang Y Y, Zhou D, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 589–597
2 Wang Q, Fang J W, Yuan Y. Multi-cue based tracking. Neurocomputing, 2014, 131: 227–236
3 Yuan Y, Fang J W, Wang Q. Online anomaly detection in crowd scenes via structure analysis. IEEE Trans Syst Man Cybernet, 2015, 45: 562–575
4 Ali S, Shah M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007. 1–6
5 Lin W Y, Mi Y, Wang W Y, et al. A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. IEEE Trans Image Process, 2016, 25: 1674–1687
6 Zhou B L, Tang X O, Wang X G. Coherent filtering: detecting coherent motions from crowd clutters. In: Proceedings of European Conference on Computer Vision, Florence, 2012. 857–871
7 Shao J, Loy C C, Wang X G. Scene-independent group profiling in crowd. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 2227–2234
8 Zhou B L, Tang X O, Zhang H P, et al. Measuring crowd collectiveness. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 1586–1599
9 Li X L, Chen M L, Nie F P, et al. A multiview-based parameter free framework for group detection. In: Proceedings of AAAI Conference on Artificial Intelligence, San Francisco, 2017. 4147–4153
10 Wang Q, Chen M L, Li X L. Quantifying and detecting collective motion by manifold learning. In: Proceedings of AAAI Conference on Artificial Intelligence, San Francisco, 2017. 4292–4298
11 Chen M L, Wang Q, Li X L. Anchor-based group detection in crowd scenes. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, New Orleans, 2017. 1378–1382
12 Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022
13 Lu H Y, Xie L Y, Kang N, et al. Don't forget the quantifiable relationship between words: using recurrent neural network for short text topic discovery. In: Proceedings of AAAI Conference on Artificial Intelligence, San Francisco, 2017. 1193–1198
14 Zhao B, Li F-F, Xing E P. Image segmentation with topic random field. In: Proceedings of European Conference on Computer Vision, Heraklion, 2010. 785–798
15 Zhou B L, Wang X G, Tang X O. Random field topic model for semantic region analysis in crowded scenes from tracklets. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, 2011. 3441–3448
16 Lu Z, Yang X K, Lin W Y, et al. Inferring user image-search goals under the implicit guidance of users. IEEE Trans Circ Syst Video Tech, 2014, 24: 394–406