Quantifying and Detecting Collective Motion by Manifold Learning

Qi Wang[1*], **Mulin Chen**[1], **Xuelong Li**[2]
[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
[2]Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an, 710119, Shaanxi, P. R. China
crabwq@gmail.com, chenmulin001@gmail.com, xuelong_li@opt.ac.cn

Abstract

The analysis of *collective motion* has attracted many researchers in artificial intelligence. Though plenty of works have been done on this topic, the achieved performance is still unsatisfying due to the complex nature of collective motions. By investigating the similarity of individuals, this paper proposes a novel framework for both quantifying and detecting collective motions. Our main contributions are threefold: (1) the time-varying dynamics of individuals are deeply investigated to better characterize the individual motion; (2) a structure-based collectiveness measurement is designed to precisely quantify both individual-level and scene-level properties of collective motions; (3) a multi-stage clustering strategy is presented to discover a more comprehensive understanding of the crowd scenes, containing both local and global collective motions. Extensive experimental results on real world data sets show that our method is capable of handling crowd scenes with complicated structures and various dynamics, and demonstrate its superior performance against state-of-the-art competitors.

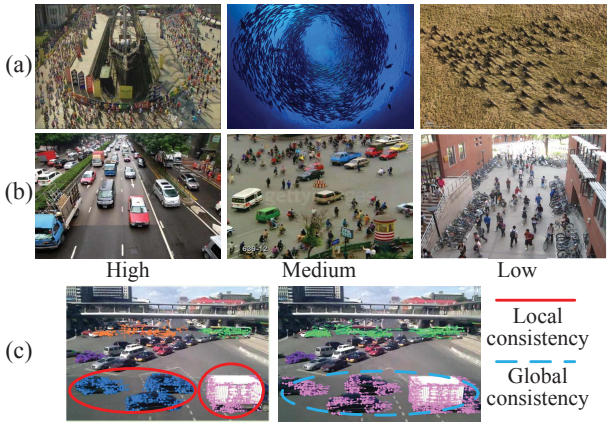


Figure 1: (a) Collective motion in pedestrians, fish shoal and bison herd. (b) Crowd scenes with varying collectiveness. (c) Local and Global consistency in crowd scenes.

Introduction

Collective motion, which is pervasive in crowd systems, has been extensively studied in many disciplines, such as psychology (Wheelan 2005), biologic (Ballerini 2008), physics (Hughes 2003). It exists widely in natural and social scenarios (e.g. Fig. 1(A)), and contains a lot of information about the crowd phenomenon. In the artificial intelligence, collective motion is primarily about human crowds, and involves a lot of applications such as multi-agent navigation (Godoy et al. 2016), crowd tracking (Zhu, Wang, and Yu 2014; Wang, Fang, and Yuan 2014; Fang, Wang, and Yuan 2014), and crowd monitoring (Zhang et al. 2015). However, both the quantification and detection of collective motions are still difficult tasks because of the complex structures and time-varying dynamics in crowd scenes.

Collectiveness is a fundamental descriptor of collective motions firstly proposed by (Zhou et al. 2014) as a quantification measure. Individual-level collectiveness indicates an individual's consistency with others, and scene-level collectiveness measures the degree of individuals' actions as an entirety in a crowd scene, as shown in Fig. 1(B). As a comprehensive feature, collectiveness is practical to quantify collective motions, and has demonstrated its merit in crowd behavior analysis (Shao, Loy, and Wang 2014; Li, Chen, and Wang 2016). Though many efforts have been spent on the quantitative calculation of collectiveness, the achieved performance is still far from ideal. This is because existing works are either limited to utilize temporal information or unable to handle collective motions with complicated spatial structures.

The detection of collective motions is also a hot but challenging issue in the realm of artificial intelligence. Generally speaking, the objective of collective motion detection is to find individuals with high behavior consistency from their time-series observations (Zhou, Tang, and Wang 2012), and the difficulty comes from two aspects. First, because of occlusion and tracking noise, it's not easy to get the accurate time-series observations of individuals. To avoid this problem, some works (Wu, Ye, and Zhao 2015; Zhou et al. 2014) detect collective motions on each frame separately, leading to an inadequate utilization of tempo-

*Qi Wang is the corresponding authors. This work is supported by the National Natural Science Foundation of China under Grant 61379094 and Natural Science Foundation Research Project of Shaanxi Province under Grant 2015JM6264.

ral information. Second, individuals in a collective motion may exhibit both local and global behavior consistency (e.g. Fig. 1(C)). Many previous works (Zhou et al. 2014; Shao, Loy, and Wang 2014; Zhou, Tang, and Wang 2012; Stauffer and Grimson 2000; Hassanein, Hussein, and Gomaa 2016; Xu et al. 2015; Liu et al. 2016) focus on the motion correlation of individuals within a local region and are limited to detect the global consistency.

In this study, we propose a framework, which is able to handle complex real-world crowd systems, to measure collectiveness accurately and detect collective motions precisely. Our contributions are summarized as follows.

1. Time-varying dynamics are deeply explored to better express the intrinsic characters of moving individuals. A hidden state-based model and a probability-based approach are put forward to explore and compare the time-varying motion dynamics of individuals.

2. A structure-based collectiveness measurement is devised to quantify collective motions with variety of spatial distributions. Instead of using the Euclidean structure, a more suitable manifold topological structure is investigated to calculate the individual/scene level collectiveness.

3. A multi-stage clustering strategy is proposed to detect collective motions precisely. This ensures our method's ability to discover collective motions with both local and global consistency along time.

Related Work

In the realm of artificial intelligence, collective motion analysis has attracted many researchers. Among numerous efforts towards this topic, we target on the works for measuring collectiveness and detecting collective motions.

To quantify collective motions, several works are engaged in calculating the collectiveness of crowd systems. Zhou et al. (2014) and Ren et al. (2015) utilized path similarity to measure collectiveness. Wu, Ye, and Zhao (2015) introduced the concept of collective density to measure collectiveness. However, the above three methods share the same problem that they measure collectiveness just by one frame and neglect the temporal correlation. Shao, Loy, and Wang (2014) calculated collectiveness on the basis of group detection, but it's limited to deal with various crowd structures, as well as the first three methods.

There are also many works focusing on detecting collective motions. Ali and Shah (2007) proposed a Lagrangian Particle based approach to segment crowd flows. Wang et al. (2014) detected coherent motion fields by spectral clustering. Wu and Wong (2012) segmented crowd motions by local-translational motion approximation. However, these flow based methods fail when handling crowds with complex patterns. Zhou et al. (2014) and Wu, Ye, and Zhao (2015) performed group detection by utilizing the information of just one frame, so they can't deal with time-varying collective motion. Some trajectories-based methods (Ge, Collins, and Ruback 2012; Zhou, Tang, and Wang 2012; Shao, Loy, and Wang 2014) achieved relatively better performance on group detection, but they are easily influenced by tracking failure and limited to detect global consistency.

Individually Time-Varying Dynamic Analysis

In crowd scenes, complex interaction among individuals makes it difficult to analyze collective motions directly. Therefore, we start by investigating the individuals' motions and their correlations. Due to the complexity of extracting pedestrians from crowd scenes, feature points are taken as the study objects, which can be detected and tracked with a generalized KLT (gKLT) tracker (Zhou et al. 2014). For ease of understanding, feature points are written as individuals in this section. First, a hidden state-based model is designed to model the trajectories of individuals. After that, a probability-based approach is put forward to calculate the consistency of individuals' motion dynamics.

Hidden state-based Model. We assume an individual's behavior is determined by its moving intention, instead of random occurrence, which means the movement of each individual is driven by a hidden intention factor. Accordingly, the behavior of each individual is considered to be dominated by a hidden state-based model. Given such a model, the trajectory of an individual can be generated under its guidance.

Considering the variety of individuals' moving intentions, we build a hidden state-based model for each individual separately to model their trajectories. In each model, a hidden state variable is inferred from an observed data since the moving direction of a pedestrian is supposed to be intention-orientated. In addition, considering the continuity of a pedestrian's moving intention, we assume the time-series dependency of hidden state variables. Denoting point i's spatial location at time t as $o_i^t = [x_i(t), y_i(t)]$, the model can be defined as

$$\begin{aligned} h_i^t &= A_i h_i^{t-1} + \mathcal{N}(0, Q_i) \\ o_i^t &= h_i^t + \mathcal{N}(0, R_i) \\ h_i^t &\sim \mathcal{N}(\mu_i, S_i), \end{aligned} \tag{1}$$

where $h_i^t \in \mathbb{R}^3$ is the hidden state variable that encodes the dynamics. $A_i \in \mathbb{R}^{3\times3}$ is a state transition matrix and $\mathcal{N}$ is a three-dimensional multivariate Gaussian distribution. Q_i, R_i and S_i are covariances, and $\mu_i \in \mathbb{R}^{3\times1}$ is the mean value of Gaussian distribution. Given the observed data of an individual i, the set of all parameters $\Theta_i = \{A_i, Q_i, R_i, \mu_i, S_i\}$ can be learned by Expectation Maximization (EM) algorithm (Chan and Vasconcelos 2008). According to the time-series dependency of hidden state variables, the log-likelihood of the observed data under the system parameters is

$$log(p(o_i^{1:n_i}|\Theta_i)) = \sum_{t=1}^{n_i} log(p(o_i^t|o_i^{1:t-1}, \Theta_i)), \tag{2}$$

which can be effectively estimated with a Kalman smoother (Shumway and Stoffer 1982). And n_i is the length of i's trajectory.

Probability-based Similarity Calculation. To measure two individuals' similarity, their neighbor relationship should be taken into account. First, the kNN method is employed to find the individuals' neighbor relationship on each frame. Then, two individuals as considered as neighbors if they are neighbors for more than three consecutive frames.

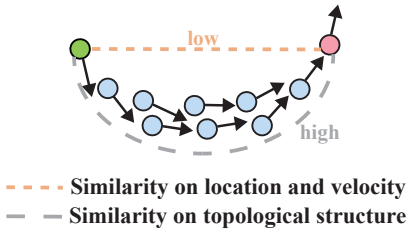


Figure 2: Illustration of topological relevance. The red point and the green point shows low similarity on spatial velocity, but they keep high topological relevance to each other. Best viewed in color.

For a pair of neighbor individuals i and j, if $o_j^{1:n_j}$ has a high log-likelihood to be generated under i's model parameters Θ_i, we can consequently say that the moving intention of j is similar to that of i. So the similarity of i and j is defined as

$$S(i,j) = \min\left(\frac{p(o_j^{1:n_j}|\Theta_i)}{p(o_i^{1:n_i}|\Theta_j)}, \frac{p(o_i^{1:n_i}|\Theta_i)}{p(o_j^{1:n_j}|\Theta_j)}\right), \tag{3}$$

where the $\min$ function restricts that the individuals with high consistency must have high probability to be produced under the model of each other. For individuals without neighbor relationship, their similarities are set to be 0. By jointly combining kNN and the hidden state-based model, both spatial and temporal relationship of individuals are successfully investigated.

Structure-Based Collective Motion Quantification

Generally, individuals in crowd scenes tend to form manifold structures (Yang et al. 2008; 2010; Peng et al. 2015), and interactions between the individuals depend more on their topological relationship than metric distance (Ballerini 2008). Therefore in this section, a manifold learning method is introduced to explore the structures of crowds and calculate collectiveness by learning the topological relationship between individuals.

For two individuals, their spatial similarity may be low, but their topological relevance to each other will be high if they are linked by consecutive neighbors. As shown in Fig. 2, the red and the green points exhibit low similarity on spatial location and velocity, but they are connected in the same collective motion. Thus, if individual i and j keep high consistency, their topological relevance to any other individual is assumed to be similar.

Given the similarity of individuals, we aim to compute the topological relationship between them. Based on the above assumption, the cost function to guide the search of the topological relationship matrix $Z \in \mathbb{R}^{N\times N}$ is defined as

$$Q(Z) = \sum_{r=1}^{N}\left(\frac{1}{2}\sum_{i,j=1}^{N} W_{ij}||Z_{ri} - Z_{rj}||^2 + \alpha\sum_{j=1}^{N}||Z_{rj} - I_{rj}||^2\right), \tag{4}$$

where r, i and j are individual indexes, Z_{ri} indicates the individual i's topological relevance to r, and the adjacency matrix $W \in \mathbb{R}^{N\times N}$ is set as $(S+S^T)/2$. I is a identity matrix, and N is the total number of individuals in the scene.. The smoothness constraint (first term) is designed to satisfy the proposed assumption, and the fitting constraint (second term) prevents all the elements of Z to be equal. And parameter α balances the two terms. Then the optimal relevance vector is

$$Z^* = \min_Z Q(Z). \tag{5}$$

Note that the problem (5) is independent for different r. Thus, we can solve the following problem separately for each r:

$$\min_{Z_r} \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}||Z_{ri} - Z_{rj}||^2 + \alpha\sum_{i=1}^{N}||Z_{ri} - I_{ri}||^2, \tag{6}$$

where Z_r is the r-th column of matrix Z. The optimal solution Z_r^* should satisfy that the derivative of Eq.(6) with respect to Z_r is equal to zero, so we have

$$LZ_r^* + \alpha(Z_r^* - I_r) = 0, \tag{7}$$

where $L \in \mathbb{R}^{N\times N}$ is the Laplacian matrix of W, and I_r is the r-th column of I. Then we get the optimal relevance vector as

$$Z_r^* = (I + L/\alpha)^{-1} I_r, \tag{8}$$

Since I_r is the r-th column of identity matrix I, it's obvious to see that Z_r^* is the r-th column of $(I + L/\alpha)^{-1}$. Thus, the optimal topological relationship matrix Z^*, which satisfies Eq.(5), can be denoted as

$$Z^* = (I + L/\alpha)^{-1}. \tag{9}$$

With all the above derivations, the individual-level collectiveness of i is defined as its topological relationship with all the other individuals

$$\phi(i) = [Z^* e]_i, \tag{10}$$

where $e \in \mathbb{R}^{N\times 1}$ is a column vector with all the elements as 1, $[\cdot]_i$ indicates the i-th element of a vector. The scene-level collectiveness is denoted as the mean value of all the individual-level collectiveness, which can be written as

$$\Phi = \frac{1}{N} e^T Z^* e. \tag{11}$$

By exploring the topological relationship between individuals, the proposed method is suitable to deal crowds with various structure.

Multi-Stage Collective motion Detection

Local Clustering

Based on the topological matrix Z, we borrow an intuitive strategy to discover the local consistency, which simply thresholds the element values of Z^*. Specifically, if $Z^*(i,j) > z$ and $Z^*(j,k) > z$ (z is set to be 0.5), the three individuals are combined in to one cluster even if $Z^*(i,k) < z$. The local clustering strategy performs well on detecting local consistency in crowd scenes, but fails to discover global consistency, as shown in Fig. 3. That's why we develop a further global clustering refinement.

(a) Local clustering result (b) Global clustering result

Figure 3: Results of local and global clustering. After global clustering, coherent sub-clusters are precisely combined.

Global Clustering

For the purpose of merging sub-clusters, it's essential to measure the consistency according to their locations and motions. Considering an individual i with n_i-length trajectory $\{[x_i(1), y_i(1)], \cdots, [x_i(n_i), y_i(n_i)]\}$, its center position is denoted as $p_i = [\frac{1}{n_i}\sum_{t=1}^{n_i} x_i(t), \frac{1}{n_i}\sum_{t=1}^{n_i} y_i(t)]$ and its average motion is $\overrightarrow{m}_i = \frac{1}{l}\sum_{t=1}^{n_i} \overrightarrow{M}_i(t)$. Thus, for a sub-cluster C, its location and motion are defined as

$$P(C) = \frac{1}{N_C}\sum_{i\in C} p_i \tag{12}$$

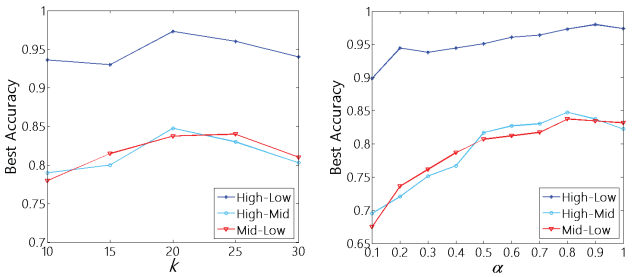
$$\overrightarrow{Mot}(C) = \frac{1}{N_C}\sum_{i\in C} \overrightarrow{m}_i, \tag{13}$$

where N_C is the number of individuals belonging to C. We assume two sub-clusters are likely to belong to the same collective motion if one resides along the other's moving direction. Besides, sub-clusters with close positions and similar motions may keep high consistency. Based on these two assumptions, the consistency between a pair of sub-clusters is defined as

$$\begin{aligned} Con(C_1, C_2) = (1 + cos(\frac{\overrightarrow{Mot}(C_1)+\overrightarrow{Mot}(C_2)}{2}, \overrightarrow{P(C_1) - P(C_2)})) \\ \times(1 + cos(\overrightarrow{Mot}(C_1), \overrightarrow{Mot}(C_2))) \\ \times e^{-\frac{2}{max(w,h)}||\overrightarrow{P(C_1)-P(C_2)}||^2}, \end{aligned} \tag{14}$$

where w and h are the width and height of the frame respectively. In Eq.14, the first term is designed according to the first assumption, and the remaining two terms comply with the second assumption. If $Cons(C_1, C_2) > c$ (c is a threshold chosen as 0.6), C_1 and C_2 are considered to be consistent and then merged into a new sub-cluster. By conducting this procedure iteratively until there are no consistent sub-clusters, the final clusters are obtained, which is also the result of collective motion detection. Since the order in which sub-cluster pairs are visited will influence the final result, we just merge those with the highest consistency in each iteration.

The multi-stage clustering method has the ability to discover both local and global consistency. The incorporating of spatial-temporal topological relationship makes our method sustain its performance along time-series.



(a) Selection of k (b) Selection of α

Figure 4: Video classification performance with varying k and α. k is varied from 10 to 30 with a 5 spacing, and α is varied from 0.1 to 1 with a 0.1 spacing.

Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the proposed method on two aspects: collectiveness measurement and collective motion detection.

Selection of Parameters

In the beginning, there are several parameters to be set first. For the hidden state-based model, μ is set as $[0\ 0\ 0]^T$ and the state transition matrix A is initialized by the suboptimal learning method (Chan and Vasconcelos 2008). The covariances Q, R, S are initialized as [1 0 0; 0 1 0; 0 0 0], [0.1 0 0; 0 0.1 0; 0 0 0], and [1 0 0; 0 1 0; 0 0 1]. As for the selection of kNN parameter k and the manifold learning parameter α, we have conducted the parametric experiments to determine their choices. Under different k and α values, collectiveness of video clips in the Collective Motion Database is calculated. Then it is used to perform binary classification of high-low, high-medium, and medium-low categories (the details of this setup will be explained in Section). The obtained best accuracies is used as the criterion of choosing parameters. In this training stage, the 100 video clips of the dataset are selected randomly, and 30 frames in each selected clips are used to train the parameters. All the rest frames are used as testing set in the following section.

An appropriate choice of the kNN parameter is essential for a good result. When k is too small, the computed collectiveness is inclined to be underestimated and the collective motions will be divided into parts. Whereas, if k is too large, individuals will be connected with those from far away, which brings noise to the result. From Fig. 4(A), it can be seen that the proposed method achieves relatively better performance when k is 20. Thus, $k = 20$ is the best choice.

In addition, the manifold learning method is important for exploring the topological relationship of individuals, which directly influences both the collectiveness measurement and collective motion detection. And the value of α affects how close the individuals are connected from the aspect of topologic. Therefore, it's necessary to find the best choice of α. As shown in Fig. 4(B), we finally choose $\alpha = 0.8$ in this work.

	High-Low			High-Low			High-Low		
	Our	CT	MCC	Our	CT	MCC	Our	CT	MCC
Precision	**0.92**	0.88	0.81	**0.87**	0.79	0.76	**0.83**	0.73	0.74
Recall	**0.71**	0.60	0.58	**0.70**	0.55	0.57	**0.72**	0.49	0.47
F-measure	**0.75**	0.58	0.51	**0.69**	0.52	0.48	**0.65**	0.44	0.40

Table 1: Performance of our method, CT and MCC on video classification. Best results are in bold face.

	High-Low		High-Low		High-Low	
	RM	MCC	RM	MCC	RM	MCC
Precision	**0.84**	0.81	**0.81**	0.76	0.72	**0.74**
Recall	**0.61**	0.58	**0.63**	0.57	**0.62**	0.47
F-measure	**0.57**	0.51	**0.59**	0.48	**0.51**	0.40

Table 2: Performance of MCC after and before replacing the manifold learning with ours. Replaced MCC is written as RM for short. Best results are in bold face.

Collectiveness Measurement Evaluation

To demonstrate the effectiveness of the proposed collectiveness measurement method, we compute scene-level collectiveness on Collective Motion Database (Zhou et al. 2014) and compare its consistency with the human perception.

Data Set. Collective Motion Database contains 413 crowd video clips (100 frames per clip) captured from 62 different scenes with various densities and structures. Each video clip is labeled with a ground truth score, which indicates the degree of behavior consistency in the crowd scene. And the clips are sorted in to high, medium, and low collectiveness according to their scores.

Performance Evaluation. We calculate the scene-level collectiveness Φ for each video, and perform binary classification of high-low, high-medium, and medium-low categories according to Φ. In order to show the effectiveness of the proposed method, Collective Transition (CT) (Shao, Loy, and Wang 2014) and Measuring Crowd Collectiveness (MCC) (Zhou et al. 2014) representing the state-of-the-art are taken for comparison. The precision-recall-F measure is

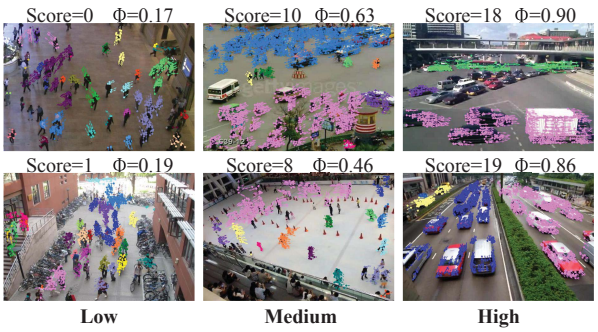


Figure 5: Representative classified scenes with their measured scene-level collectiveness Φ (from 0 to 1) and ground truth scores (from 0 to 20). It can be seen that Φ keeps consistency with the ground truth score.

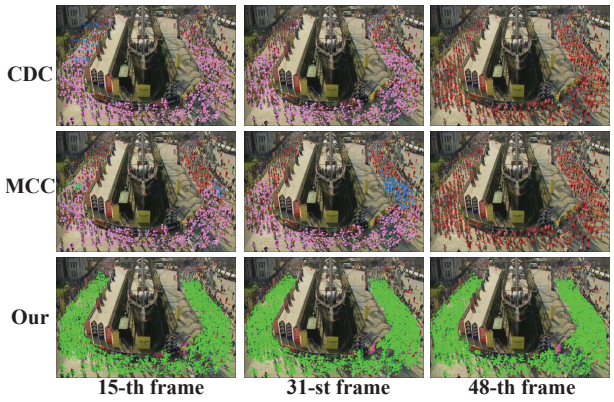


Figure 7: Comparison of collective motion detection results along time-series. Scatters with different colors indicate different detected collective motions, and the red color indicates outliers. Compared to CDC and MCC, our method maintains better performance along time-series and detects less outliers.

employed for evaluation. The averaged results are visualized in Table 1. It is manifest that our method always achieves higher precision, recall and F-measure than CT and MCC. CT learns a transition matrix of a detected group, and uses the fitting errors of trajectories to measure collectiveness. It neglects the structures of crowds. Based on the individuals' topological relationship, MCC measures collectiveness on each frame without utilizing temporal information. It's thus not able to quantify collective motions along time-series. On the contrary, our method addresses these problems by proposing a structure-based collectiveness measurement, and exploring the time-varying dynamics of individuals. Consequently, the proposed one outperforms CT and MCC. Some representative results are shown in Fig. 5.

The proposed manifold learning method in Section is also compared with that in MCC. We replace the manifold learning method in MCC with ours, and compare their classification performance. Experimental results are shown in Table 2. Despite the lower precision in Mid-Low classification, the replaced MCC shows superior performance than MCC. The manifold learning method of MCC investigates topological relationship by accumulating similarities along all paths between each pair of individuals. Some useless paths are therefore included. Our method performs better because it emphasizes the neighbor relationship between individuals, which complies with the theory that collective motions are formed by the information propagation between neighbors (Ballerini 2008). All these experiments indicate the proposed collective calculation is more suitable for the real situations.

Collective Motion Detection Evaluation

To validate the superiority of our collective motion detection approach, we conduct experiments on CUHK Crowd Dataset (Shao, Loy, and Wang 2014) and compare it with state-of-the-art competitors. The parameters for detection

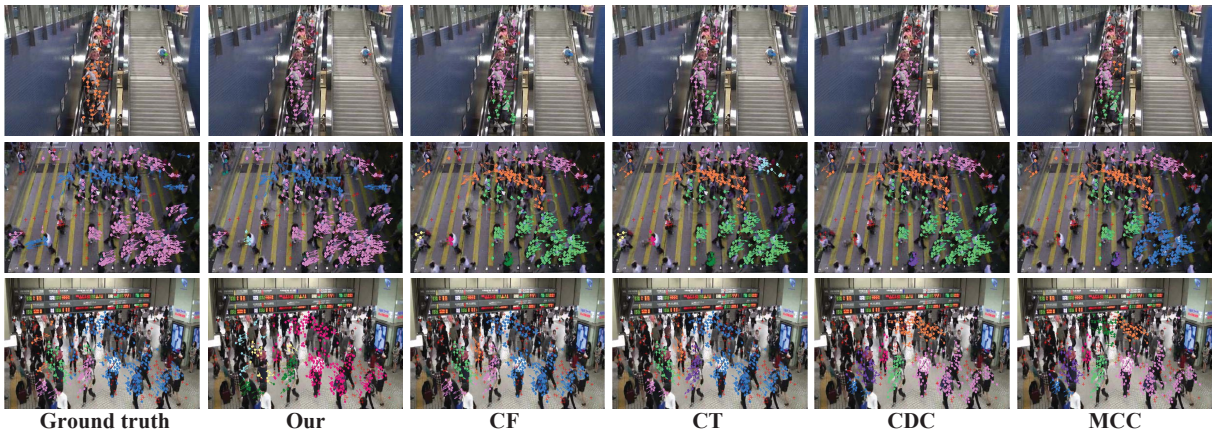


Figure 6: Representative comparison results of collective motion detection. Scatters with different colors indicate different detected collective motions, and the red color indicates outliers. Our result is closer to the ground truth and can detect less mislabeled outliers than the competitors.

Comparison of Collective Motion Detection

	Our	CF	CT	CDC	MCC
NMI	**0.60**	0.42	0.48	0.39	0.40
Purity	**0.86**	0.73	0.78	0.74	0.85
RI	**0.87**	0.78	0.83	0.73	0.74

Table 3: Quantitative comparison of collective motion detection methods. The best results are in bold face.

are the same as those used in collectiveness calculation.

Data Set. CUHK Crowd Dataset provides 474 crowd videos for group detection, which are captured from real-world crowd scenes with a variety of crowdness. It records the labels of collective motions that each individual belongs to, and individuals not belonging to any collective motion are annotated as outliers.

Performance Evaluation. The detection results of the proposed method are compared with four state-of-the-art algorithms, namely, Coherent Filtering (CF) (Zhou, Tang, and Wang 2012), Collective Transition (CT) (Shao, Loy, and Wang 2014), Collective Density Clustering (CDC) (Wu, Ye, and Zhao 2015), and Measuring Crowd Collectiveness (MCC) (Zhou et al. 2014).

The detection of collective motions can be considered as the clustering of individuals in crowd scenes. So we evaluate the results of different methods by adopting three widely used clustering metrics: Normalized Mutual information (NMI) (Wu and Schölkopf 2006; Peng et al. 2016), Purity (Aggarwal 2004), and Rand Index (RI) (Rand 1971). Quantitative comparison is shown in Table 3. It is clear that our method achieves the highest NMI, Purity and RI, which validates the superiority of the proposed collective motion detection algorithm. Some representative detection results are shown in Fig. 6. Since CF and CT detect collective motions by locally clustering the trajectories of individuals, both of them are limited to detect global consistency. This can be observed in the second row in Fig. 6, where CF and CT mistakenly split a cluster of pedestrians moving in the same direction into two clusters. Instead, our method is more capable of discovering global consistency because of the multi-stage clustering method. MCC employs a manifold learning technique to detect collective motions, but shares the same shortcoming with CF and CT, as shown in the first row in Fig. 7. CDC detects coherent motions by measuring crowd density in crowd scenes. Nevertheless, both CDC and MCC detect collective motions frame by frame separately, and neglect the temporal smoothness. Thus they can't maintain a stable performance along time-series. As Fig. 7 visualizes, CDC and MCC perform well at the 15th frame, but they can't maintain performance at the 31st and the 48th frame. Especially, at the 48th frame, both CDC and MCC fail to detect the actural collective motion because of tracking failure. Our method achieves stable performance on all frames because of its successful exploration of time-varying dynamics.

Conclusion and Future Work

In this paper, we study the problem of quantifying and detecting collective motions in crowd scenes. The time-varying dynamics of individuals are sufficiently explored by a hidden state-based model. Then a structure-based collectiveness measurement is developed to quantify collective motions and a multi-stage clustering strategy is introduced to detect collective motions in crowd scenes. Experiments on various real-world videos validate that our method yields substantial boosts over state-of-the-art competitors.

In the future work, we would like to extend our method to more applications in artificial intelligence, such as activity recognition and video description. It's also desirable to apply our method in crowd behavior simulation.

References

Aggarwal, C. C. 2004. A human-computer interactive method for projected clustering. *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence 16(4):448–460.

Ali, S., and Shah, M. 2007. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–6.

Ballerini, M. 2008. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the national academy of sciences* 105(4):1232–1237.

Chan, A. B., and Vasconcelos, N. 2008. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(5):909–926.

Fang, J.; Wang, Q.; and Yuan, Y. 2014. Part-based online tracking with geometry constraint and attention selection. *IEEE Trans. Circuits Syst. Video Techn.* 24(5):854–864.

Ge, W.; Collins, R. T.; and Ruback, B. 2012. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(5):1003–1016.

Godoy, J. E.; Karamouzas, I.; Guy, S. J.; and Gini, M. L. 2016. Implicit coordination in crowded multi-agent navigation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2487–2493.

Hassanein, A. S.; Hussein, M. E.; and Gomaa, W. 2016. Semantic analysis for crowded scenes based on non-parametric tracklet clustering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3389–3395.

Hughes, R. L. 2003. The flow of human crowds. *Annual Review of Fluid Mechanics* 35(1):169–182.

Li, X.; Chen, M.; and Wang, Q. 2016. Measuring collectiveness via refined topological similarity. *ACM TOMM* 12(2).

Liu, W.; Zha, Z.; Wang, Y.; Lu, K.; and Tao, D. 2016. p-laplacian regularized sparse coding for human activity recognition. *IEEE Transaction on Industrial Electronics* 63(8):5120–5129.

Peng, X.; Lu, J.; Zhang, Y.; and Yan, R. 2015. Automatic subspace learning via principal coefficients embedding. *IEEE Transactions on Cybernetics* 1–14.

Peng, X.; Xiao, S.; Feng, J.; Yau, W.; and Yi, Z. 2016. Deep subspace clustering with sparsity prior. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligenc*, 1925–1931.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336):846–850.

Ren, W.; Li, S.; Guo, Q.; Li, G.; and Zhang, J. 2015. Agglomerative clustering and collectiveness measure via exponent generating function. *CoRR* abs/1507.08571.

Shao, J.; Loy, C. C.; and Wang, X. 2014. Scene-independent group profiling in crowd. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2227–2234.

Shumway, R. H., and Stoffer, D. S. 1982. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time* 3(4):25–264.

Stauffer, C., and Grimson, W. E. L. 2000. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(8):747–757.

Wang, W.; Lin, W.; Chen, Y.; Wu, J.; Wang, J.; and Sheng, B. 2014. Finding coherent motions and semantic regions in crowd scenes: A diffusion and clustering approach. In *European Conference on Computer Vision*, 756–771.

Wang, Q.; Fang, J.; and Yuan, Y. 2014. Multi-cue based tracking. *Neurocomputing* 131:227–236.

Wheelan, S. A. 2005. *The handbook of group research and practice*. SAGE Publications.

Wu, M., and Schölkopf, B. 2006. A local learning approach for clustering. In *Advances in Neural Information Processing Systems*, 1529–1536.

Wu, S., and Wong, H. 2012. Crowd motion partitioning in a scattered motion field. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 42(5):1443–1454.

Wu, Y.; Ye, Y.; and Zhao, C. 2015. Coherent motion detection with collective density clustering. In *ACM Conference on Multimedia*, 361–370.

Xu, H.; Zhou, Y.; Lin, W.; and Zha, H. 2015. Unsupervised trajectory clustering via adaptive multi-kernel-based shrinkage. In *IEEE International Conference on Computer Vision*, 4328–4336.

Yang, Y.; Zhuang, Y.; Wu, F.; and Pan, Y. 2008. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transaction on Multimedia* 10(3):437–446.

Yang, Y.; Nie, F.; Xiang, S.; Zhuang, Y.; and Wang, W. 2010. Local and global regressive mapping for manifold learning with out-of-sample extrapolation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 649–654.

Zhang, X.; Yang, S.; Tang, Y. Y.; and Zhang, W. 2015. Crowd motion monitoring with thermodynamics-inspired feature. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 4300–4302.

Zhou, B.; Tang, X.; Zhang, H.; and Wang, X. 2014. Measuring crowd collectiveness. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(8):1586–1599.

Zhou, B.; Tang, X.; and Wang, X. 2012. Coherent filtering: Detecting coherent motions from crowd clutters. In *European Conference on Computer Vision*, 857–871.

Zhu, F.; Wang, X.; and Yu, N. 2014. Crowd tracking with dynamic evolution of group structures. In *European Conference on Computer Vision*, 139–154.