Quantifying and Detecting Collective Motion in Crowd Scenes

Xuelong Li, *Fellow, IEEE* , Mulin Chen, and Qi Wang, *Senior Member, IEEE*

***Abstract*—People in crowd scenes always exhibit consistent behaviors and form collective motions. The analysis of collective motion has motivated a surge of interest in computer vision. Nevertheless, the effort is hampered by the complex nature of collective motions. Considering the fact that collective motions are formed by individuals, this paper proposes a new framework for both quantifying and detecting collective motion by investigating the spatio-temporal behavior of individuals. The main contributions of this work are threefold: 1) an intention-aware model is built to fully capture the intrinsic dynamics of individuals; 2) a structure-based collectiveness measurement is developed to accurately quantify the collective properties of crowds; 3) a multi-stage clustering strategy is formulated to detect both the local and global behavior consistency in crowd scenes. Experiments on real world data sets show that our method is able to handle crowds with various structures and time-varying dynamics. Especially, the proposed method shows nearly 10% improvement over the competitors in terms of NMI, Purity and RI. Its applicability is illustrated in the context of anomaly detection and semantic scene segmentation.**

***Index Terms*—Crowd analysis, Collectiveness, Manifold learning, Group detection, Clustering**

I. Introduction

Collective motion, which is the primary component that makes up a crowd, is one of the most attractive phenomena in both nature and human society. Individuals in a collective motion tend to share consistent property, which is fundamentally important for analyzing the underlying pattern of crowd behavior. Since collective motion provides a mid-level representation of crowds, it has drawn increasing attentions in the field of computer vision, and involves a wide range of applications, such as crowd tracking [1], [2], [3], crowd counting [4], [5] and action recognition [6], [7], [8], [9]. However, due to the complex spatial distribution and time-varying dynamics in crowd scenes, both the quantification and detection of collective motion are still difficult tasks.

In order to compare different crowd systems quantitatively, several works are conducted on the quantification of collective motions. Particularly, the *collectiveness* descriptor proposed by Zhou et al. [10] is the first scene-independent quantification measurement. In specific, individual-level collectiveness describes an individuals' behavior consistency with others, and scene-level collectiveness indicates the degree of all the individuals acting as a team. As a fundamental descriptor, collectiveness captures the universal characteristic of collective motions, and it has shown its applicability in crowd video classification and crowd modelling [11]. However, the calculation of collectiveness faces two major challenges: (1) it's complicated to compare the long-term behaviors of individuals, whose motion dynamics change with time; (2) in crowd scenes with manifold structures, due to the information propagation between neighbors, individuals in the same collective motion may exhibit various behaviors, as shown in Fig. 1 (a), which increases the difficulty on collectiveness measurement.

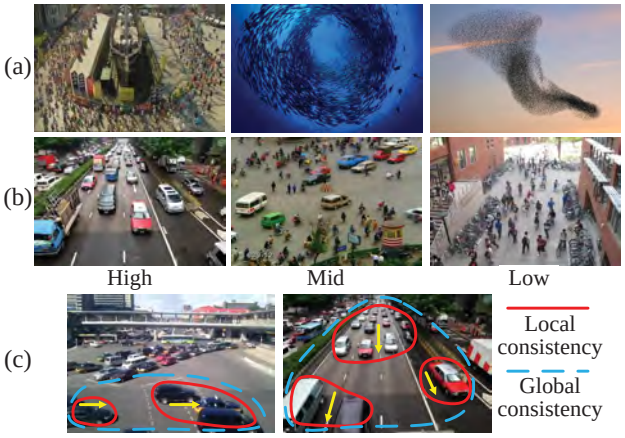


Fig. 1. (a) Collective motions with manifold structures. (b) Crowd scenes with high, mid and low collectiveness. (c) Local and Global consistency in crowd scenes, yellow arrows indicate moving directions.

Collective motion detection aims to cluster the pedestrians according to their motion patterns. Generally speaking, it can be formulated as the clustering of individuals with similar motion patterns. Different clusters convey different semantic behaviors, so collective motion detection could facilitate some semantics-driven tasks, such as crowd activity recognition [12], [13], [14] and scene understanding [15]. Similar to the quantification task, collective motion detection also suffers from the aforementioned two problems. In addition, collective motion involves both local and global behavior consistencies, as shown in Fig. 1 (c). Due to the different arrival time, individuals in the same collective motion may reside far away from each other, which makes the global consistency hard to detect.

The goal of this study is to measure collectiveness precisely and detect collective motions correctly. We put forward a framework, which has the capability to handle complex real-

This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, 61871470, and 61761130079. The authors are with the school of computer science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China. E-mail: xuelong_li@nwpu.edu.cn; chenmulin001@gmail.com;crabwq@gmail.com. Q. Wang is the corresponding author.

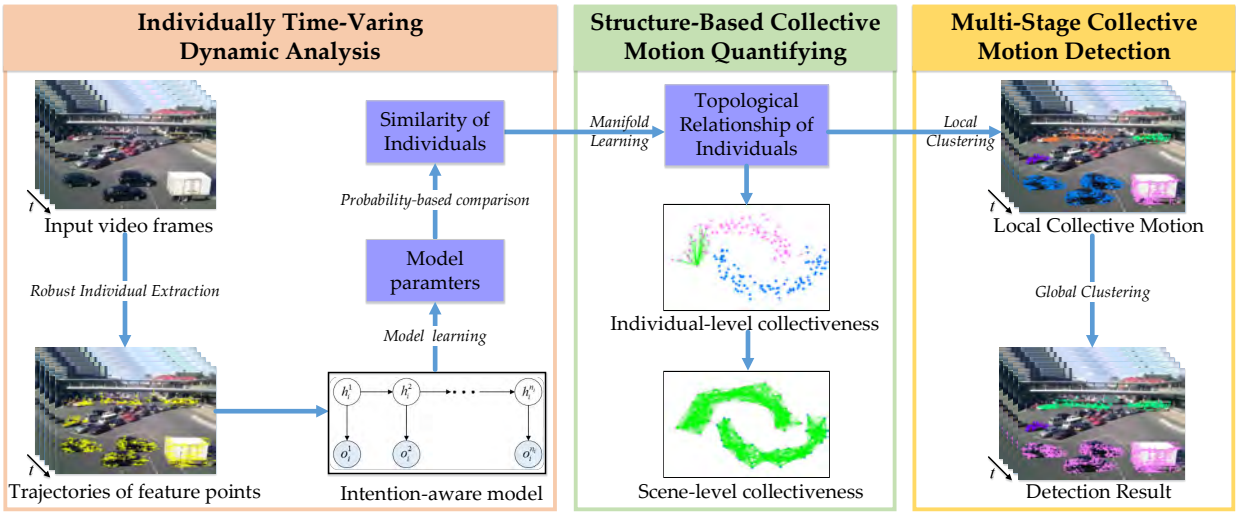


Fig. 2. The pipeline of the proposed framework. First, we extract motions of individuals with robustness, and propose an intention-aware model to analyze time-varying motion dynamics. Then a collectiveness measurement, which investigates the topological relationship between individuals, is utilized to measure collectiveness. Finally, a multi-stage clustering strategy is developed to detect collective motions in crowd scenes.

world crowd systems. Firstly, individuals are identified and represented by feature points. Secondly, the individuals' trajectories are modelled and compared. After that, the topological relationship between the individuals is learned, and the collectiveness is calculated. Finally, based on the learned topological relationship, collective motion detection is performed with a multi-stage clustering method. The pipeline of the proposed framework is shown in Fig.2.

We summarize our contributions as follows.

1) An intention-aware model and a probability-based approach are proposed to deeply exploit and compare the time-varying motion dynamics of individuals. The trajectory of each individual is modelled, and individuals are compared according to their intrinsic motion patterns.
2) A structure-based collectiveness measurement is developed to characterize collective motions with various spatial structures. By exploring the propagation of local similarity, the proposed method is more suitable to reveal the real crowd condition, and measure individual-/scene-level collectiveness accurately.
3) A multi-stage clustering strategy is designed to detect both the local and global consistencies in crowd scenes. During the multi-stage clustering procedure, our method can perceive the global message in the scene and get a whole view of the crowd, which is the weak side of many traditional algorithms.

Compared to the conference version of this research [16], this paper is considerably improved by providing more technical details, more experimental evaluations and applications. Some related issues are also discussed. The rest of this paper is organized as follows. Section II reviews the works on the quantification and detection of collective motion. Section III introduces the individually time-varying dynamic analysis approach. Section IV proposes the structure-based collectiveness measurement. Section V describes the multi-stage collective motion detection method. Section VI presents the extensive experiments to verify the superiority of the proposed framework, and Section VII shows its potential applications. The conclusion and future work follow in Section VIII.

II. Related Work

During the past decade, crowd analysis has captivated many researchers due to the increasing demands on surveillance applications. Scientific studies [17], [18], [19] pointed out that crowds are formed by individuals with similar motion patterns. Collective motion reveals the underlying principles of crowd behaviors and gives a mid-level understanding of crowd phenomenon. Here we briefly review the previous works toward this topic.

A. Collective Motion Quantification

The quantification of collective motion has been long ignored in computer vision until the collectiveness descriptor [10] was proposed. Zhou et al. [10] regarded collectiveness as a bottom feature, and measured it by exploring the relationship between individuals. They built an adjacent graph for the individuals according to their spatial locations and motion directions, and then calculate collectiveness by accumulating the weight along all the paths between individuals. Based on [10], Ren et al. [20] introduced an exponent generating function to modify the accumulating operation. However, both Zhou et al. [10] and Ren et al. [20] rely on a subjective assumption that the relationship between individuals decreases exponentially with the path length, which may not be true for real-world crowds. Wu et al. [21] estimated collectiveness with a density-based clustering method [22]. Li et al. [23] designed a point selection strategy to better extract individuals, and utilized the manifold ranking method to exploit the relationship between individuals. Due to the difficulty of long-term motion exploration, all the above methods perform calculation on each frame separately. So they are limited to capture the time-varying dynamics of

individuals. Shao et al. [11] used long-term trajectory as study object. They first detected collective motions by employing the coherent filtering method [24], and found the anchor trajectory to calculate the transition priori. According the prior, the fitting errors of individuals are averaged to compute collectiveness. This method emphasizes the temporal aspect, but it neglects the interaction among individuals, so it can not deal with the crowds with various structures.

B. Collective Motion Detection

According to the type of clues, existing works on collective motion detection can be roughly classified into two categories: 1) fixed particle-based techniques; 2) feature point-based techniques.

As for the first category, a grid of particles is preliminary overlaid on the scene, and then collective motions are detected by analyzing the optical flow of particles. Brox et al. [25] utilized a nonlinear diffusion method to enhance optical flow, and then detected collective motion by approximating the flow distribution. Base on Lyapunov exponent field and Lagrangian particle dynamics, Ali and Shah [26] proposed a mathematical framework to segment collective flow in crowd scenes. Wu and Wong [27] sought the salient optical flow in crowds, and designed a local-translation domain segmentation model to partition the flow into collective motions. Yuan et al. [7] devised a structural context descriptor to character the optical flow of particles, and detected the collective motion with a potential energy function. Lin et al. [15] employed the thermal diffusion theory to process the optical flow of particles, and then discovered coherent flow by spectral clustering. These methods need to model the motion dynamic of each particle. However, there are always thousands of particles in each scene, which makes the algorithms time-consuming. Moreover, they fail to deal with the complex motion patterns since the flow of particles can not profile the crowd motion precisely.

For the second category, feature points in crowd scenes are extracted to represent individuals, and the detection task is accomplished by exploring their movements. Zhou et al. [10] introduced a manifold learning method to measure the topological relationship of individuals, based on which a collective merging method was employed to detect coherent motion. Wu et al. [21] modified the density-based clustering method [22], and designed a merging strategy to characterize the behavior consistency in crowds. Li et al. [28] devised a context descriptor to reveal the structural property of points, and proposed a multi-view clustering method to fuse the features from different aspects. The above three methods neglect the temporal smoothness, so their performance fluctuate on different frames. Ge et al. [29] detected collective motions with a bottom-up hierarchical clustering method, which depends on the similarity of individuals' trajectories. Zhou et al. [24] found the invariant neighbors of each individual, and combined those with high velocity correlations into the same collective motion. Shao et al. [11] refined the results of Zhou et al. [24] by removing the individuals that do not fit the transition prior. These trajectory-based methods perform relatively better, however, they neglect the consistency between non-neighbors. In addition, they just focus on the individuals within a local region, so the global consistency is ignored.

III. INDIVIDUALLY TIME-VARYING DYNAMIC ANALYSIS

Individuals with similar destinations tend to walk together, and their frequent interactions give rise to the emergence of collective behaviors [17]. Thus, the correlation between individuals is the key to understanding collective motions. Choi et al. [30] modelled the interaction between pedestrians directly, which shows good performance for crowds with multiple pedestrians. However, for large-scale crowds with hundreds or thousands of pedestrians, it's almost infeasible to extract them accurately. So we employ feature points as study objects alternatively. In this section, an individually time-varying dynamic analysis approach is proposed to exploit and compare the time-series movements of individuals. It contains three steps: robust individual extraction, intention-aware hidden state model and probability-based similarity calculation.

A. Robust Individual Extraction

The extraction of individuals is fundamental for the analysis of collective motions. Deep learning-based detection [31], [32], [33] and tracking [34], [35], [36] methods have shown promising performance in recent years, but for crowd scenes which may contain thousands of individuals (see Fig. 1 (a)), these methods are difficult to be performed because manual labels are extremely expensive and almost impossible. Moreover, it is hard to extract the individuals accurately due to the variance of perspectives. Therefore, feature points are utilized in this work as an alternative to represent individuals. This processing avoids the exhaustedly detection and tracking of each individual, and is able to profile the crowd dynamic.

Firstly, we detect feature points by using the generalized Kandae-Lucas-Tomasi (gKLT) [10] detector. It is achieved by finding the minimum Hessian matrix eigenvalue within a sliding window and provides stable candidates for tracking. gKLT is used to detect the points because it finds the points with a relatively uniform distribution. Secondly, Robust Local Optical Flow (RLOF) [37] algorithm is used to track feature points since it is capable of handling the interruption of background noises [38], [39]. Finally, to tackle tracking drifting, a forward-backward refinement strategy [40] is utilized, which uses the resulting position of a feature point as input to the same tracking method, and discards it if the reverse tracking does not result in its initial position.

By incorporating these techniques, we can acquire the feature points' trajectories and velocities along time-series with robustness. To maximize clarity, feature points are written as individuals hereafter. Note that, although optical flow is used for tracking, but the proposed method does not rely on dense particles, so its efficiency is guaranteed.

B. Intention-Aware Hidden State Model

Behavior analysis in crowd is challenging due to the time-varying motions of individuals. According to Mehran et al.

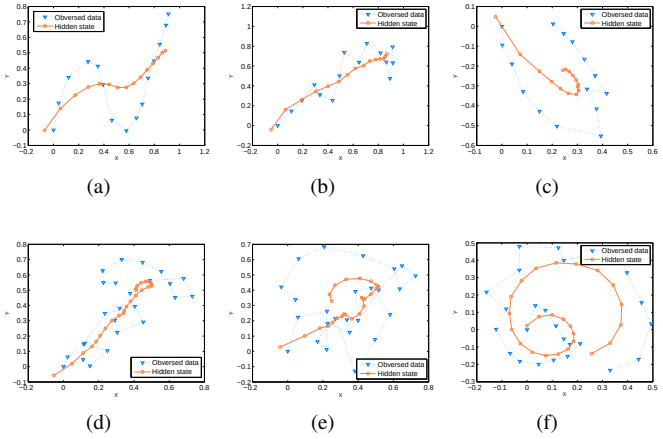


Fig. 3. Hidden states learned from the observed data. The coordinates of the points record the first two dimensions of the observed data and the hidden state without the bias term.

[41], each individual in crowd scenes has its own moving intention. Intuitively, we believe that the intention drives an individual's movement. Therefore, an individual's intrinsic motion pattern can be exploited by inferring its moving intention.

Considering the moving intention as a hidden factor, we built a Linear Dynamic System (LDS) model [42] for each individual separately. LDS models the observed data with a hidden state variable, which is in accordance with our assumption that the individual's movement is considered to be intention-directed. In addition, time-series dependency is assumed on the hidden state variables to reveal the continuity of an individual's moving intentions. Let $\mathbf{o}_i^t = [x_i(t), y_i(t), 1]^T$ be the observed data of individual i at time t, where $[x_i(t), y_i(t)]$ is the spatial location and 1 is a bias term. Then the model is defined with the form of

$$\begin{aligned} \mathbf{h}_i^t &= \mathbf{A}_i \mathbf{h}_i^{t-1} + \mathcal{N}(\mathbf{0}, \mathbf{Q}_i), \\ \mathbf{o}_i^t &= \mathbf{h}_i^t + \mathcal{N}(\mathbf{0}, \mathbf{R}_i), \\ \mathbf{h}_i^1 &\sim \mathcal{N}(\mu_i, \mathbf{F}_i), \end{aligned} \tag{1}$$

where $\mathbf{h}_i^t \in \mathbb{R}^{3\times 1}$ is the hidden variable that encodes the motion dynamic. $\mathbf{A}_i \in \mathbb{R}^{3\times 3}$ is a transition matrix that evolves the hidden variable. $\mathcal{N}$ is a three-dimensional Gaussian distribution, $\mathbf{Q}_i$, $\mathbf{R}_i$ and $\mathbf{F}_i \in \mathbb{R}^{3\times 3}$ are covariances, and $\mu_i \in \mathbb{R}^{3\times 1}$ is the mean. Denoting $\mathbf{\Theta}_i = \{\mathbf{A}_i, \mathbf{Q}_i, \mathbf{R}_i, \mu_i, \mathbf{F}_i\}$ as the set of model parameters, the motion pattern of i can be captured captured once $\mathbf{\Theta}_i$ is learnt. The details about model inference is given in Section III-D.

C. Probability-Based Similarity Calculation

In this part, the intrinsic dynamic similarity of individuals is calculated. For this purpose, we first investigate the spatial relationship of individuals. For each frame, kNN method is utilized to find the neighbor relationship of individuals. Two individuals are regarded as neighbors if they keep neighbor relationship on more than three frames.

Afterwards, the motion similarity of individuals is measured by comparing their intrinsic intentions. To reduce the computation complexity, we only calculate the similarities between neighbors. Non-neighbor interaction will be taken into account in the next section. According to Eq. (1), the log-likelihood of the observed data under specific model parameters is

$$\log(p(\mathbf{o}_i^{1:n_i}|\mathbf{\Theta}_i)) = \sum_{t=1}^{n_i} \log(p(\mathbf{o}_i^t|\mathbf{o}_i^{1:t-1}, \mathbf{\Theta}_i)), \tag{2}$$

where n_i is the length of i's trajectory. The above log-likelihood can be solved by a modified Kalman smoother [43], [44], which is suitable to optimize the LDS model [42]. The log-likelihood can be interpreted as the probability of i's time-series movements under specific moving intention. Consequently, for a pair of neighbor individuals i and j, if i's observed data $o_i^{1:n_i}$ has a high likelihood under j's model parameters, they are considered to share similar motion patterns. So we define the similarity of i and j as

$$S_{ij} = \min[\frac{p(\mathbf{o}_j^{1:n_j}|\mathbf{\Theta}_i)}{p(\mathbf{o}_i^{1:n_i}|\mathbf{\Theta}_j)}, \frac{p(\mathbf{o}_i^{1:n_i}|\mathbf{\Theta}_i)}{p(\mathbf{o}_j^{1:n_j}|\mathbf{\Theta}_j)}], \tag{3}$$

where $\min(\cdot)$ encourages that the individuals have a high probability to be generated under each other's model. Through the above procedures, both spatial and temporal information are sufficiently incorporated into the similarity calculation, so our method is able to compare the spatio-temporal movements of individuals. As shown in Fig. 3, the learned hidden state reveals the motion dynamic of the observed data steadily, even for the data with complex shapes, such as Fig. 3 (d)-(f).

D. Model Initialization and Inference

In the proposed intention-aware model, given the observed data $\{\mathbf{o}_i^{1:n_i}\}$, we would like to find the model parameters $\mathbf{\Theta}_i = \{\mathbf{A}_i, \mathbf{Q}_i, \mathbf{R}_i, \mu_i, \mathbf{F}_i\}$ that best fit the data, which can be achieved by maximizing the log-likelihood of observations,

$$\mathbf{\Theta}_i^* = \arg\max_{\mathbf{\Theta}_i} \log p(\mathbf{o}_i^{1:n_i}; \mathbf{\Theta}_i). \tag{4}$$

Since a hidden state variable is introduced in the model to represent the intention, EM algorithm [44], [42] can be employed to solve Eq. (4). Given the initial values of the model parameters, EM algorithm iteratively estimates missing information and updates the current parameters. Each iteration contains

$$\mathbf{E-step:} \tag{5}$$
$$\vartheta(\mathbf{\Theta}_i, \hat{\mathbf{\Theta}}_i) = \mathrm{E}_{\mathbf{h}_i^{1:n_i}|\mathbf{o}_i^{1:n_i}, \hat{\mathbf{\Theta}}_i^*}[\log p(\mathbf{o}_i^{1:n_i}, \mathbf{h}_i^{1:n_i}; \mathbf{\Theta}_i)],$$
$$\mathbf{M-step:} \tag{6}$$
$$\hat{\mathbf{\Theta}}_i^* = \arg\max_{\mathbf{\Theta}_i} \vartheta(\mathbf{\Theta}_i; \hat{\mathbf{\Theta}}_i),$$

where $p(\mathbf{o}_i^{1:n}, \mathbf{h}_i^{1:n}; \Theta_i)$ is the overall joint distribution of the observations and hidden states parameterized by $\mathbf{\Theta_i}$, and $\hat{\mathbf{\Theta}}_i$ is the current estimation of $\mathbf{\Theta}_i$.

Initialization. Before performing EM algorithm, the model parameters should be initialized. For an individual i, its Gaussian mean μ_i is set as $[0\ 0\ 0]^T$, the covariance matrices $\mathbf{Q}_i$, $\mathbf{R}_i$ and $\mathbf{F}_i$ are initialized as $[1\ 0\ 0;\ 0\ 1\ 0;\ 0\ 0\ 0]$,

$[0.1\ 0\ 0;\ 0\ 0.1\ 0;\ 0\ 0\ 0]$ and $[1\ 0\ 0;\ 0\ 1\ 0;\ 0\ 0\ 1]$. Note that, in $\mathbf{Q}_i$ and $\mathbf{R}_i$, the last element is set as 0 to fix the bias term in $\mathbf{h}_i^t$ and $\mathbf{o}_i^t$. To initialize the transition matrix $\mathbf{A}_i$, a suboptimal learning strategy [45] is utilized. Given $\mathbf{R}_i^t$ and $\mathbf{o}_i^t$, the series of hidden variables $\mathbf{h}_i^{1:n_i}$ can be obtained. Intuitively, $\mathbf{A}_i$ should minimize the transition error of $\mathbf{h}_i^{1:n_i}$, so we have

$$\mathbf{A}_i^* = \arg\min_{\mathbf{A}_i} ||\mathbf{h}_i^{2:n_i} - \mathbf{A}_i \mathbf{h}_i^{1:n_i-1}||_2^2. \tag{7}$$

So $\mathbf{A}_i$ is initialized as $\mathbf{h}_i^{2:n_i}(\mathbf{h}_i^{1:n_i-1})^T$, which is the suboptimal solution of problem (7).

Expectation-step. In this stage, the expectation of $p(\mathbf{o}_i^{1:n_i}, \mathbf{h}_i^{1:n_i}; \mathbf{\Theta}_i)$ is estimated, as in Eq. (5). Given the current model parameters, according to Eq. (1), the joint distribution $p(\mathbf{o}_i^{1:n_i}, \mathbf{h}_i^{1:n_i}; \mathbf{\Theta}_i)$ can be denoted as

$$\begin{aligned}
&p(\mathbf{o}_i^{1:n_i}, \mathbf{h}_i^{1:n_i}; \mathbf{\Theta}_i) \\
=&\prod_{t=1}^{n_i} p(\mathbf{o}_i^t, \mathbf{h}_i^t; \mathbf{\Theta}_i) \\
=&p(\mathbf{h}_i^1; \mu_i, \mathbf{F}_i) \prod_{t=2}^{n_i} p(\mathbf{o}_i^t|\mathbf{h}_i^t; \mathbf{R}_i) p(\mathbf{h}_i^t|\mathbf{h}_i^{t-1}; \mathbf{A}_i, \mathbf{Q}_i) \\
=&\mathcal{N}(\mathbf{h}_i^1|\mu_i, \mathbf{F}_i) \prod_{t=2}^{n_i} \mathcal{N}(\mathbf{o}_i^t|\mathbf{h}_i^t, \mathbf{R}_i) \mathcal{N}(\mathbf{h}_i^t|\mathbf{A}_i \mathbf{h}_i^{t-1}, \mathbf{Q}_i).
\end{aligned} \tag{8}$$

With modified Kalman smoother [44], we can get the following conditional expectations

$$\begin{aligned}
\hat{\mathbf{h}}_i^t &= \mathrm{E}_{\mathbf{h}_i^{1:n}|\mathbf{o}_i^{1:n}}(\mathbf{h}_i^t), \\
\hat{\mathbf{P}}_i^{t,t} &= \mathrm{E}_{\mathbf{h}_i^{1:n}|\mathbf{o}_i^{1:n}}[\mathbf{h}_i^t (\mathbf{h}_i^t)^T], \\
\hat{\mathbf{P}}_i^{t,t-1} &= \mathrm{E}_{\mathbf{h}_i^{1:n}|\mathbf{o}_i^{1:n}}[\mathbf{h}_i^t (\mathbf{h}_i^{t-1})^T],
\end{aligned} \tag{9}$$

then Eq. (5) can be rewritten as

$$\begin{aligned}
&\vartheta(\mathbf{\Theta}_i, \hat{\mathbf{\Theta}}_i) \\
=&-\frac{1}{2}\sum_{t=1}^{n_i} tr(\mathbf{R}_i^{-1}[\mathbf{o}_i^t(\mathbf{o}_i^t)^T - \mathbf{o}_i^t(\hat{\mathbf{h}}_i^t)^T - \hat{\mathbf{h}}_i^t(\mathbf{o}_i^t)^T + \hat{\mathbf{P}}_i^{t,t}]) \\
&-\frac{1}{2}\sum_{t=2}^{n_i} tr(\mathbf{Q}_i^{-1}[\hat{\mathbf{P}}_i^{t,t} - \hat{\mathbf{P}}_i^{t,t-1}\mathbf{A}^T - \mathbf{A}_i(\hat{\mathbf{P}}_i^{t,t-1})^T]) \\
&-\frac{1}{2} tr(\mathbf{F}_i^{-1}[\hat{\mathbf{P}}_i^{1,1} - \hat{\mathbf{h}}_i^1 \mu_i^T - \mu_i(\hat{\mathbf{h}}_i^1)^T + \mu_i \mu_i^T]) \\
&-\frac{n_i}{2}\log|\mathbf{R}_i| - \frac{n_i - 1}{2}\log|\mathbf{Q}_i| - \frac{1}{2}\log|\mathbf{F}_i| \\
&-\frac{1}{2}\sum_{t=2}^{n_i} tr(\mathbf{Q}_i^{-1}\mathbf{A}_i \hat{\mathbf{P}}_i^{t-1,t-1}\mathbf{A}_i^T),
\end{aligned} \tag{10}$$

where $tr(\cdot)$ indicates the trace operator.

Maximization-Step. In this stage, new model parameters $\mathbf{\Theta}_i^* = \{\mathbf{A}_i^*, \mathbf{Q}_i^*, \mathbf{R}_i^*, \mu_i^*, \mathbf{F}_i^*\}$ are obtained by maximizing ϑ. Differentiating Eq. (10) with respect to each parameter and setting it to 0, we get the optimal parameters in the current step,

$$\begin{aligned}
\mathbf{A}_i^* &= \sum_{t=2}^{n_i} \hat{\mathbf{P}}_i^{t,t} (\sum_{t=2}^{n} \hat{\mathbf{P}}_i^{t-1,t-1})^{-1}, \\
\mathbf{Q}_i^* &= \frac{1}{n_i - 1}[\sum_{t=2}^{n} \hat{\mathbf{P}}_i^{t,t} - \mathbf{A}_i^* (\sum_{t=2}^{n_i} \hat{\mathbf{P}}_i^{t,t-1})^T], \\
\mathbf{R}_i^* &= \frac{1}{n}[\sum_{t=1}^{n_i} \mathbf{o}_i^t (\mathbf{o}_i^t)^T - \sum_{t=1}^{n_i} \mathbf{o}_i^t (\hat{\mathbf{h}}_i^t)^T], \\
\mathbf{F}_i^* &= \hat{\mathbf{P}}_i^{1,1} - \mu_i^* (\mu_i^*)^T, \\
\mu_i^* &= \hat{\mathbf{h}}_i^1.
\end{aligned} \tag{11}$$

E. Discussion

In this section we propose an intention-aware approach to characterize the connection between individuals. Its major difference from previous studies is that it has the capability to compare the individuals' spatio-temporal behaviors. Existing works [10], [24], [11], [21], [29] always measure the individuals' similarity by computing their instantaneous velocity correlation on each frame. Thus, these methods fail to give a holistic insight to the behavior consistency in crowds. In our method, the time-series observed data is modelled with LDS, and the similarity is measured with the learnt model parameter. So the proposed method is naturally appropriate for handling time-series data.

However, a problem still exists. For individuals without neighboring relationship, their similarities are set as 0. But this is not true for real-world occasions. Due to the information propagation through neighbors, individuals without neighbor relationship may also keep high consistency [46]. That's why a manifold learning method is followed in the next section to learn the consistency between individuals.

IV. STRUCTURE-BASED COLLECTIVE MOTION QUANTIFICATION

With the individuals' similarities, the collectiveness is measured on both individual- and scene-level in this section. In the previous step, only the neighbors' similarities is calculated. However, the far away individuals may also keep high consistency since local similarity propagates through the paths between them, especially for the crowds with manifold structures, as shown in Fig. 1 (a). According to Ballerini et al. [46], the interaction among individuals depends on their similarities across paths, which is also termed as topological relevance in machine learning. So a manifold learning method is proposed to capture the topological relationship.

A. Methodology

To facilitate explanation, Fig. 4 visualizes a manifold structure formed by set of moving particles. The green and red points have different velocities and reside far away. However, they are connected together by consecutive neighbors, so their path similarity is high. So we first map the local similarity to the topological space and then measure collectiveness according to the individuals' topological relevance.

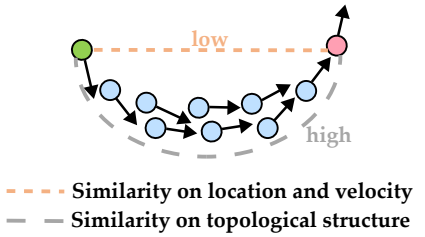


Fig. 4. Illustration of topological relationship. The red and green point show low similarity on spatial location and moving direction, but they are consistent from the topologic perspective. Best viewed in color.

Motivated by the observation that similarity propagates through paths, we put forward an assumption: if two individuals are similar, their topological relevance to any other individual should also be similar. By transmitting topological relationship through similar individuals, the consistency of faraway individuals can be captured. Supposing the topological relationship between individual r and i is Z_{ri}, then the optimal topological relevance matrix $\mathbf{Z}^* \in \mathbb{R}^{N\times N}$ is learnt by minimizing the following function

$$\min_{\mathbf{Z}} \sum_{r=1}^{N} [\frac{1}{2} \sum_{i,j=1}^{N} W_{ij}(Z_{ri} - Z_{rj})^2 + \alpha \sum_{i=1}^{N} (Z_{ri} - I_{ri})^2], \quad (12)$$

where N is the total number of individuals. The weight matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ is set as $(\mathbf{S}+\mathbf{S}^T)/2$ to keep the symmetry, where $\mathbf{S}$ is the similarity graph learned by Eq. (3). $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix. In Eq. (12), the first term guarantees that Z_{ri} should be close to Z_{rj} if i and j are similar, which implies the proposed assumption. The second term prevents that all the elements in $\mathbf{Z}$ are equal. The parameter α captures the trade-off between the two constraints.

From Eq. (12), we can see that problem (12) is independent between different r, so the problem can be solved for each r separately:

$$\min_{\mathbf{Z}_r} \frac{1}{2} \sum_{i,j=1}^{N} W_{ij}(Z_{ri} - Z_{rj})^2 + \alpha \sum_{i=1}^{N} (Z_{ri} - I_{ri})^2, \quad (13)$$

where $\mathbf{Z}_r$ is the r-th row of $\mathbf{Z}$. Taking the derivative of Eq. (13) w.r.t. $\mathbf{Z}_r$, and setting it to 0, we have

$$\mathbf{L}\mathbf{Z}_r^T + \alpha(\mathbf{Z}_r^T - \mathbf{I}_r) = 0, \quad (14)$$

where $\mathbf{L} \in \mathbb{R}^{N\times N}$ is the Laplacian matrix of $\mathbf{W}$, and $\mathbf{I}_r$ is the r-th row of $\mathbf{I}$. Since $(\mathbf{I} + \mathbf{L}/\alpha)$ is invertible, the optimal relevance vector $\mathbf{Z}_r^*$ is

$$\mathbf{Z}_r^* = \mathbf{I}_r(\mathbf{I} + \mathbf{L}/\alpha)^{-1}. \quad (15)$$

Fortunately, $\mathbf{Z}_r^*$ is exact the r-th row of matrix $(\mathbf{I} + \mathbf{L}/\alpha)^{-1}$, so the optimal topological relationship matrix $\mathbf{Z}^*$ is

$$\mathbf{Z}^* = (\mathbf{I} + \mathbf{L}/\alpha)^{-1}. \quad (16)$$

With the topological relationship matrix Z^*, we define the individual-level collectiveness of i as the sum of its relevance with all the other individuals

$$\phi(i) = [\mathbf{Z}^*\mathbf{1}]_i, \quad (17)$$

where $\mathbf{1}$ is the column vector with all elements as 1, and $[\cdot]_i$ means the i-th element of a vector. The scene-level collectiveness is defined as the mean of all the individual collectiveness

$$\Phi = \frac{1}{N}\mathbf{1}^T\mathbf{Z}^*\mathbf{1}. \quad (18)$$

By exploiting the propagation of local similarity, individuals' topological relevance is measured reasonably. So our method is suitable to handle the complex interaction among individuals, and capable of quantifying crowds with manifold structures.

B. Discussion

The objective function of our method is of the similar form with traditional label propagation methods [47], [48]. However, they are different in nature. The label propagation methods learn either features or labels from the labelled data, while the proposed method searches a topological relevance matrix with the weight matrix, which is quite different. In addition, traditional methods require a set of labelled data, so they are semi-supervised. Instead, in our objective, all the elements in the target matrix is unknown, making the proposed method totally unsupervised. Thus, the proposed manifold learning method has certain innovation.

V. Multi-Stage Collective Motion Detection

With all the above quantitative definitions, we can target on the problem of detecting collective motions in crowd scenes. The basic idea is based on the topological relationship between individuals. There exists some works on this topic, they mainly have two obvious limitations: (1) they are not able to handle time-varying dynamics of collective motions owing to the insufficient use of spatio-temporal information; (2) they neglect the global consistency of individuals' behaviors. Motivated by these deficiencies, we introduce a multi-stage clustering method gradually exploring the local and global consistency.

A. Local Clustering

The topological relationship is utilized to cluster individuals in an intuitive way, which finds the locally consistent individuals by simply thresholding the values on $\mathbf{Z}^*$. Especially, supposing th_1 is the threshold (th_1 is 0.5 in our experiments), if $\mathbf{Z}_{ij}^* > th_1$ and $\mathbf{Z}_{jk}^* > th_1$, then the three individuals will be merged into the same sub-cluster even when $\mathbf{Z}_{ik}^* < th_1$. Fig. 5 illustrates that the local clustering processing detects local consistency accurately, but fails to cluster the coherent individuals within the global scope. So a further global refinement is devised to process the obtained sub-clusters.

B. Global Clustering

Since the sub-clusters can not capture the global consistency, we propose to merge them according to their spatial locations and motions. First, the consistency of sub-clusters is measured.

(a) Local clustering result

(b) Global clustering result

Fig. 5. Results of local and global clustering. The global clustering combines the continuous sub-clusters precisely.

Suppose the center position of individual i's trajectory is p_i, and its average velocity is $\overrightarrow{v_i}$. Thus the location and motion of a sub-cluster c are denoted as

$$\begin{aligned} \mathbf{p}_c &= \frac{1}{N_c}\sum_{i\in c}\mathbf{p}_i \\ \mathbf{v}_c &= \frac{1}{N_c}\sum_{i\in c}\mathbf{v}_i, \end{aligned} \tag{19}$$

where N_c is the total number of individuals within c. Then the coherency of sub-clusters is measured according to the following observations. For sub-clusters c_1 and c_2, if c_1 resides along c_2's motion direction, then c_2 is likely to appear on c_1's position after several frames. So c_1 and c_2 may exhibit coherent behavior. Moreover, sub-clusters belonging to the same collective motion often have close spatial locations and similar motion directions. Therefore the consistency between sub-clusters is defined as

$$\begin{aligned} \mathbf{Con}(c_1,c_2) =& (1+\cos(\mathbf{v}_{c_1}+\mathbf{v}_{c_2},\mathbf{p}_{c_1}-\mathbf{p}_{c_2})) \\ &\times (1+\cos(\mathbf{v}_{c_1},\mathbf{v}_{c_2})) \\ &\times \exp(-\frac{2}{\max(w,h)}||\mathbf{p}_{c_1}-\mathbf{p}_{c_2}||_2^2), \end{aligned} \tag{20}$$

where $\cos()$ computes the cosine similarity, w and h are the width and height of the current frame. The first term complies with the first observation, and the other two imply the second observation. Similar to the local clustering stage, two sub-clusters are considered to be consistent if their consistency is greater than the threshold th_2 (th_2 is 0.5 in the experiments). By merging the consistent sub-clusters iteratively, we can get the final collective motions. Note that, to remove the interference of merging order, only the sub-clusters with the highest consistency are combined in each iteration.

The multi-stage clustering method is able to detect the both local and global collective motions in crowd scenes. Because clustering method employs the spatial-temporal topological relationship of individuals, our collective motion detection method can achieve stable performance.

C. Discussion

This section arouses the following question. Since both the proposed manifold learning method and the global clustering processing pull the far away individuals together into a collective motion, what's the difference between them? Here we discuss this confusion. The manifold learning method mainly deals with the individuals that exhibit different behaviors and linked by consecutive neighbors. For two far away individuals, their topological relationship will be low if they are not connected by neighbors. However, the individuals in the same collective motion may step into the scene at different times, so there may be no neighbors between them, as shown in Fig. 5 (a). For those individuals, it's necessary to introduce the global clustering step. Thus, the manifold learning method focuses on the behavior divergence, while the global clustering strategy handles the individuals with different arriving time. They play different roles on the detection of collective motion, and both of them are important for the proposed framework. The whole procedure is outlined in Algorithm 1.

Algorithm 1 The proposed framework

Input: Input video, parameters k, α, thresholds th_1 and th_2.
Output: Individual-level collectiveness $\{\phi(i)\}$, scene-level collectiveness Φ, clusters of collective motion.

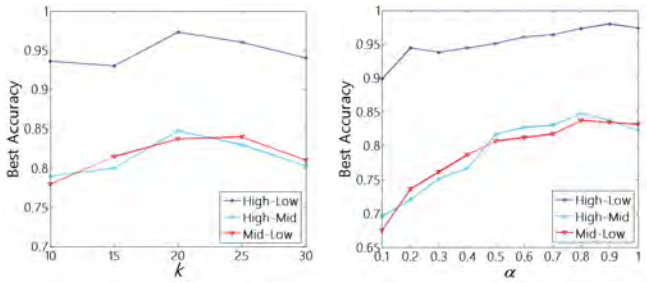
Stage: Individual-based time-varying dynamic analysis
1: Detect and track feature points.
2: **for** each individual i **do**
3: Define observed data $\{\mathbf{o}_i^t = [x_i(t), y_i(t), 1]^T, t \in (1, n_i)\}$;
4: Learn model parameters $\mathbf{\Theta}_i$ by Eq. (5) and (6).
5: **end for**
6: Calculate individuals' similarity matrix $\mathbf{S}$ by Eq.3. (Section III)

Stage: Structure-based collective motion quantification
7: Compute topological relationship matrix $\mathbf{Z}$ with $\mathbf{S}$ by Eq.16.
8: Calculate $\{\phi(i)\}$ with $\mathbf{Z}$ by Eq.17.
9: Calculate Φ with $\mathbf{Z}$ by Eq.18. (Section IV)

Stage: Multi-stage collective motion detection
10: Merge individuals into sub-clusters by thresholding $\mathbf{Z}$ with th_1. (Section V-A)
11: **repeat**
12: Combine consistent sub-clusters with th_2 by Eq.19;
13: **until** no consistent sub-clusters
14: Get final clusters of collective motions. (Section V-B)

VI. Experiments

In this section, the proposed framework is evaluated on two tasks: collectiveness measurement and collective motion detection. Throughout the experiments, we make all the competitors use their respective optimal parameters to ensure a fair comparison.



(a) Selection of k (b) Selection of α

Fig. 6. (a) The curve of best accuracies with varying parameter k. k is varied from 10 to 30 with a 5 spacing. (b) The curve of best accuracies with varying parameter α. α is varied from 0.1 to 1 with a 0.1. It can be seen that the performance is relatively good with $k = 20$ and $\alpha = 0.8$.

A. Parameter Selection

Firstly, experiments are conducted to decide the best configuration of parameters k (used in kNN procedure) and α (the balance parameter). With different k and α, the scene-level collectiveness is measured on the crowd videos in Collective Motion Database. According to the obtained collectiveness, we perform binary video classification of high-low, high-mid, and mid-low categories (the detailed experimental setting is introduced in Section VI-B). Then the best classification accuracy across varying threshold is employed as the criterion for parameter selection. The parameters are trained on the first 30 frames in 100 randomly selected videos, and all the remaining frames are further employed to evaluate the collectiveness measurement in Section VI-B.

Parameter k influences the performance greatly since it determines the size of the neighborhood. A small k leads to the underestimation of collectiveness and makes a collective motion divided into several parts. Meanwhile, a large k combines the far away individuals together, and brings additional noises to the final result. Fig. 6 (a) shows the curve of the best accuracies with varying k, and we can see that the performance is better with k equal to 20. So k is selected as 20 in this work.

Additionally, the manifold learning parameter α is also crucial for the overall performance. It directly affects the calculation of topological relevance, which is the basis of the collective motion quantification and detection. So it's essential to find the best value of α. The corresponding curve is shown in Fig. 6 (b), accordingly α is chosen as 0.8.

Then the selected value of k and α are used in all the following experiments.

B. Collectiveness Measurement Evaluation

In order to verify the performance of the proposed collectiveness measurement, we measure the scene-level collectiveness on real-world crowd videos, and compare its consistency with labelled human perception.

Dataset. Collective Motion Database is employed here, which consists of 413 crowd videos captured from 62 different scenes with various structures. Each video clip contains 100 frames, and is labelled manually as low, medium and high by 10 subjects according to the behavior consistency. By majority voting, the videos are partitioned into three categories. In this work, the collectiveness Φ is measured for each video. Then we threshold Φ to perform binary classification of high-low, high-mid and mid-low categories. With all the possible thresholds, we can obtain a set of classification precisions, recalls and F-measures [49]. The averaged precision, recall and F-measure are used as evaluation criteria.

(a) High-Low

	Our	CT	MCC
precision	**0.92**	0.88	0.81
recall	**0.71**	0.60	0.58
F-measure	**0.75**	0.58	0.51

(b) High-Mid

	Our	CT	MCC
precision	**0.87**	0.79	0.76
recall	**0.70**	0.55	0.57
F-measure	**0.69**	0.52	0.48

(c) Mid-Low

	Our	CT	MCC
precision	**0.83**	0.73	0.74
recall	**0.72**	0.49	0.47
F-measure	**0.65**	0.44	0.40

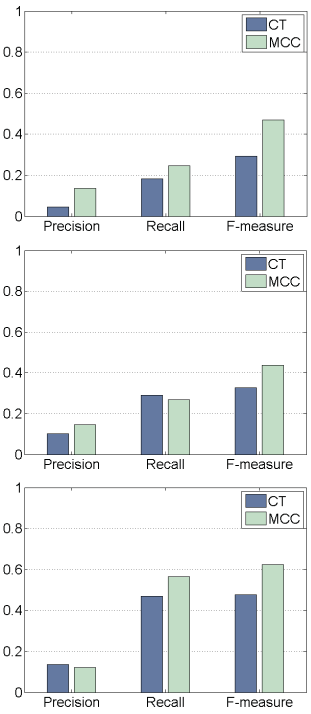


Fig. 7. The left of (a-c) are the averaged performance of classifying high-low, high-mid, and mid-low collecitveness videos by our method, CT and MCC. The right of (a-c) are the relative improvements of our method compared with CT and MCC. The bold face shows the best result.

Performance Evaluation. Two state-of-the-art methods are taken for comparison, they are Collective Transition (CT) [11] and Measuring Crowd Collectiveness (MCC) [10], are taken for comparison. The classification results are shown in Fig. 7, and the bar charts visualize the comparative improvements of our method compared with CT and MCC. The proposed method achieves the highest averaged precision, recall and F-measure in all situations, which means that it produces more accurate collectiveness than CT and MCC. CT learns a collective transition prior for the crowd motion, and computes collectiveness by accumulating the fitting error of each individual. So it captures the temporal information. However, the ignorance of structural property makes it unable to handle the crowds with complex structures. MCC builds an adjacency graph for individuals, and leans their topological similarity. But it measures the collectiveness for each frame separately, so it can not perceive the time-varying motion dynamic of individuals. In our method, the above problems are settled by manifold learning and intention-aware modelling. So it shows superiority over CT and MCC. Fig. 8 shows some representative results. The collectiveness score is the sum of the rating of 10 subjects.

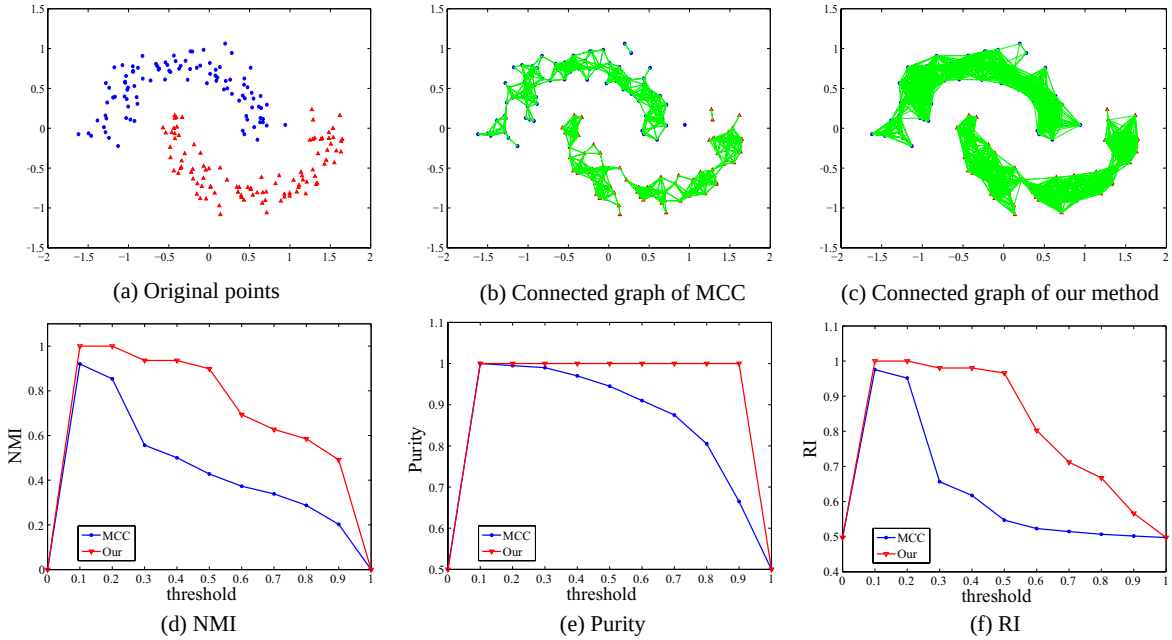


Fig. 9. Clustering results on the two-moon toy dataset by MCC and the proposed methods. In (a)-(c), different colors indicate different clusters, and green lines indicate the connection between points. Best viewed in color.

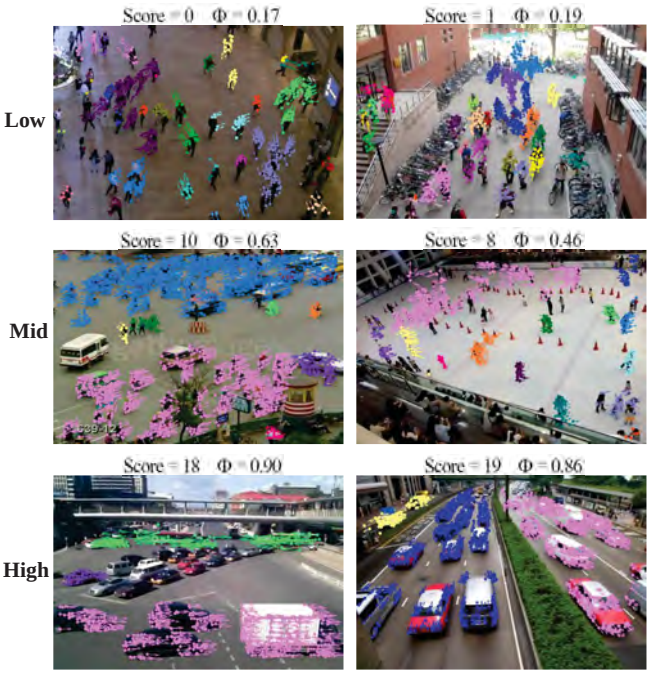


Fig. 8. Representative classified crowds with their ground truth scores (from 0 to 20) and measured scene-level collectiveness Φ (from 0 to 1). Φ keeps consistency with the ground truth score.

TABLE I
PERFORMANCE COMPARISON OF RMCC AND MCC. BEST RESULTS ARE IN BOLD FACE.

	High-Low		High-Mid		Mid-Low	
	RMCC	MCC	RMCC	MCC	RMCC	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

C. Manifold Learning Evaluation

Here we evaluate the proposed manifold learning method by comparing it with the one in MCC.

First, we replace the manifold learning method in MCC with ours, and compare the replaced MCC with the original one on measuring collectiveness. The comparison of performance is shown in Table I. Although the precision is lower in mid-low case, the replaced MCC (named as RMCC) achieves better performance compared with the original MCC. Given the adjacent graph, the manifold learning method in MCC computes the topological similarity of two individuals by accumulating the weight along all paths between them. However, the crowd information propagates through neighbors [46], not all the paths. Moreover, it assumes that the topological relevance decreases exponentially with the length of path, which seems arbitrary. On the contrary, our manifold learning method complies with the information propagation theory by emphasizing the neighbor relationship, and the basic assumption is reasonable. So it's more suitable to measure the topological relationship of individuals.

In addition, experiments are conducted on a toy dataset. In this test, two clusters of data points are generated in the two-moon pattern, as shown in Fig. 9 (a), points in each moon form a cluster. For the points, the affinity matrix $\mathbf{W}$ is constructed with the Gaussian kernel according to the Euclidean distances of points. According to the affinity matrix, the topological relationship matrix can be learnt. Then we threshold the topological matrix and combine the points with high relevance iteratively, and the final clusters can be obtained. Fig. 9 (d)-(e) show the clustering performance of the proposed manifold learning method and the one in MCC. Compared to MCC, the proposed method achieves higher NMI, Purity and RI with varying threshold, which indicates the good performance.

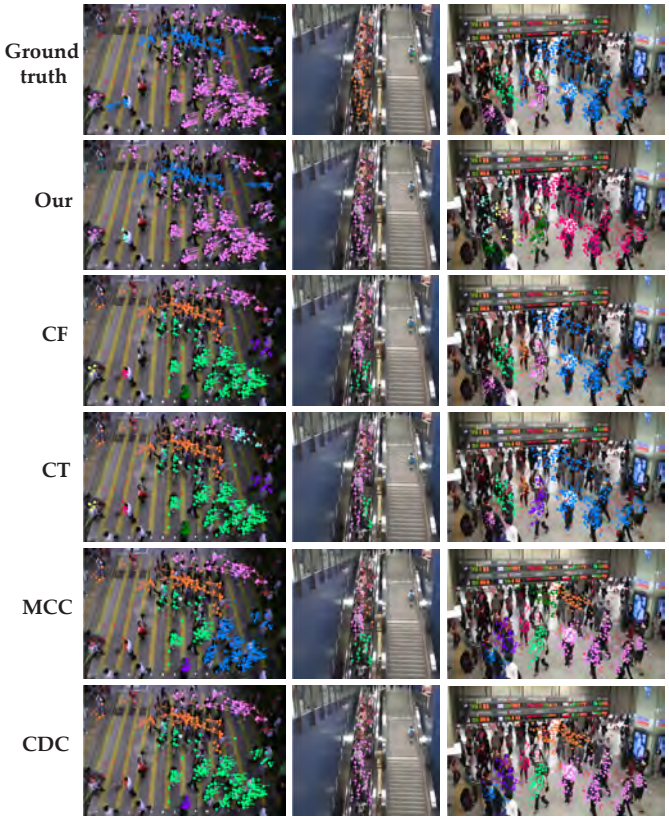


Fig. 10. Representative results of collective motion detection. Scatters with different colors indicate different detected collective motions, and the red plus sign indicates outliers. Our result is closer to the ground truth.

TABLE II
QUANTITATIVE COMPARISON OF COLLECTIVE MOTION DETECTION METHODS. THE BEST RESULTS ARE IN BOLD FACE.

	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

Furthermore, we visualize the topological relationship between points. It can be seen in Fig. 9 (d)-(e) that both MCC and our method perform best when threshold is 0.1. So we connect the points with green line if their topological relevance exceeds 0.1. In Fig. 9 (b), some points are not connected into the corresponding moon, which means that MCC fails to partition the points into two clusters correctly. On the other hand, Fig. 9 (c) shows that the proposed manifold learning method successfully connects the points in each moon, and there is not any line between different moons, which demonstrates that all the points are clustered into the correct category. So the proposed manifold learning method is applicable to unsupervised clustering task.

D. Collective Motion Detection Evaluation

To demonstrate the effectiveness of the proposed collective motion detection approach, comparison experiments are conducted on the CUHK Crowd Dataset [11].

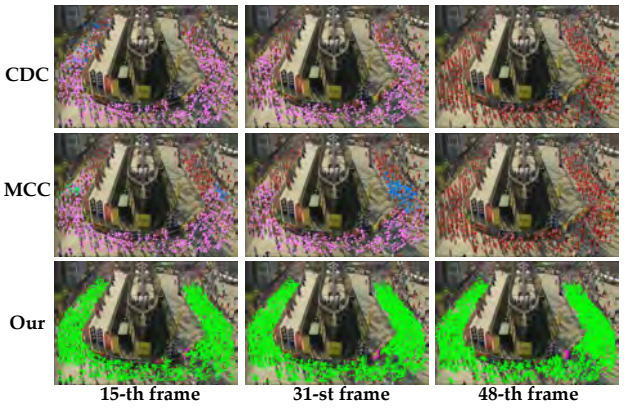


Fig. 11. Comparison of collective motion detection results along time-series. Scatters with different colors indicate different detected collective motions, and the red color indicates outliers.

Dataset. CUHK Crowd Dataset contains 474 crowd video clips captured from various crowd scenes, and 300 of them are labelled with the ground truth for collective motion detection. The ground truth contains the collective motion index of each individual, and individuals outside of any collective motion are labelled as outliers.

Performance Evaluation. The proposed method is compared with Coherent Filtering (CF) [24], Collective Transition (CT) [11], Measuring Crowd Collectiveness (MCC) [10], and Collective Density Clustering (CDC) [21], which represent the state-of-the-art. Since collective motion detection is equivalent to the clustering of individuals, we employ three standard clustering metrics as measurements: Normalized Mutual information (NMI) [50], Purity [51], and Rand Index (RI) [52]. The quantitative comparison of different methods is shown in Table II, and some representative detection results are visualized in Fig. 10. From Table II, we can see that our method achieves the highest NMI, Purity and RI, which indicates its consistency with human perception. Both CF and CT find the invariant surroundings of individuals within a local region, so they can not detect the global collective motion. As shown in the first column in Fig. 10, both CF and CT erroneously split the pedestrians moving in the same direction into sub-clusters. On the contrary, our method detect global consistency accurately with the multi-stage clustering strategy. MCC discovers coherent motion with a collective merging method, which focuses on the neighbors' relationship and neglects the global consistency. So it shares the same deficiency with CF and CT, as shown in the second column in Fig. 10. CDC puts up a good performance well on detecting global collective motion, since it also emphasizes the continuous sub-clusters. However, both CDC and MCC process each frame separately, and omit the temporal information. So they can not sustain their performance along time-series. As shown in Fig. 11, CDC and MCC perform well on the 15th frame, but the performance decreases on the 31st and 48th frames. Particularly, both of them fail on the 48th frame due to the tracking noise. The proposed method maintains good performance on all frames because of its capability to handle the time-varying dynamics.

In addition, we conduct experiments on collective motion

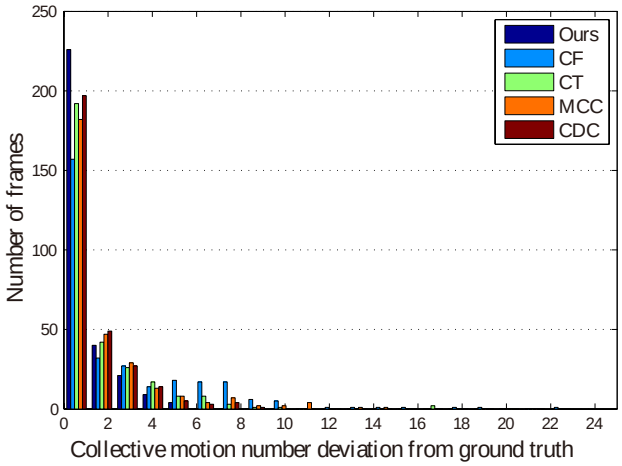


Fig. 12. Histogram of the collective motion number difference comparing with ground truth on the CUHK Crowd Dataset. Our method shows less deviation.

TABLE III
QUANTITATIVE ON COLLECTIVE MOTION NUMBER ESTIMATION. THE BEST RESULTS ARE IN BOLD FACE.

	Our	CF	CT	MCC	CDC
AD	**1.15**	2.45	1.63	2.02	1.59
MSE	**1.32**	3.01	1.83	2.56	1.84

number estimation. The estimation accuracy indicates the capability to detect global collective motion. Fig. 12 shows the distribution of deviation between the detected number and ground truth. Compared with others, our method has less deviation from the ground truth, and its deviation mainly locates in the range of [0,2]. For quantitative evaluation, we calculate the Average Difference (AD) and Mean Square Error (MSE) of each method as follows

$$\begin{aligned} \mathrm{AD} &= \frac{1}{N_{clips}} \sum_{clip} |\ \mathrm{Num}(clip) - \mathrm{Num}_{gt}(clip)|, \\ \mathrm{MSE} &= \sqrt{\frac{\sum\limits_{clip} \left(|\mathrm{Num}(clip) - \mathrm{Num}_{gt}(clip)| - \mathrm{AD}\right)^2}{N_{clips}}}, \end{aligned} \tag{21}$$

where $\mathrm{Num}(clip)$ records the number of detected collective motions in each video clip, $\mathrm{Num}_{gt}(clip)$ is the ground truth, and N_{clips} is the number of video clips in the dataset. The lower AD corresponds to the less deviation from real group number, and the lower MSE indicates a higher stability of group detection. Table III denotes the AD and MSE of each method. The AD and MSE of the proposed method are the lowest. CDC also obtains relatively good results, due its global clustering procedure. The performance of CF is unsatisfactory because it can not distinguish groups with subtle difference. The proposed method has the ability to capture the global consistency precisely, so it achieves promising results.

VII. APPLICATIONS

In order to demonstrate the usefulness of the proposed framework, we show its potential contribution on anomaly detection and semantic scene segmentation, which are often studied in crowd surveillance.

Fig. 13. (a) Crowd Scenes with abnormal pedestrians. (b) Anomaly detection results, green scatters indicate abnormal pedestrians and arrows indicate moving directions. Our method correctly identifies the abnormal pedestrians in the crowd scenes.

A. Anomaly Detection in Crowd Scenes

The objective of anomaly detection in crowd scenes is to discover and locate individuals with abnormal behaviors. This is critically important for security based applications. Whereas, both the extraction of individuals and the classification of behaviors are difficult issues. In the proposed method, individuals are extracted and represented with robust feature points and then classified into different collective motion clusters according to their dynamics. Since different feature points and clusters have distinctive properties, we can use this information as a criterion to identify the anomalies. To be specific, we average the individual-level collectiveness within each cluster, and threshold the obtained value representing the cluster collectiveness. A low cluster collectiveness value indicates individuals in the cluster are inconsistent with others, and they are considered to be abnormal. As visualized in Fig.13, two pedestrians moves against all the others, which can be regarded as an abnormal event. Our method extracts the abnormal pedestrians precisely and classifies them from the normal pedestrians accurately. Thus, the proposed approach is helpful to the anomaly detection task.

B. Semantic Scene Segmentation

Our framework can also be utilized to segment semantic regions in videos containing crowd scenes. Initially, the examined frame is segmented into patches as Fig. 14 (a) illustrates. Then collective motions are detected by our method, and each kind of collective motion is assigned with an index, as shown in Fig. 14 (b). Thirdly, every patch is encoded by an index vector recording the types and times of crossing collective motions.

For instance, suppose there are two kinds of collective motions. If Motion1 passes through patch i for 3 frames, and Motion2 for 5 frames, an index vector $\mathbf{IV}_i$=[3,5] will be used

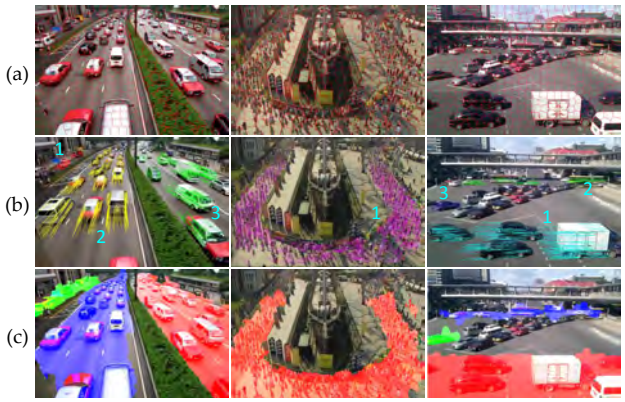


Fig. 14. (a) Segmented patches for the examined frame. (b) Collective motions detected by our methods. Different colors of lines indicate trajectories of different collective motions. (c) Semantic regions after merging patches.

to characterize this information. Similarly, if the two motions pass through patch j for 1 time and 4 times respectively, it is denoted by $\mathbf{IV}_j$ = [1,4]. Consequently, we can define the similarity of patch i and j as

$$\mathbf{S}_{patch}(i,j) = \exp(-\frac{||\mathbf{IV}_i - \mathbf{IV}_j||_2}{N_{frames}}), \tag{22}$$

where N_{frames} means the total number of frames for supporting the segmentation of the examined frame. In this way, patch i and j will have higher similarity if the collective motions passing through them are prone to be parts of the same semantic region. Finally, based on the similarity matrix $\mathbf{S}_{patch}$, clustering method can be employed to merge the patches into semantic regions, as shown in Fig. 14 (c). In our implementation, we employ SLIC [53] to segment the image into 500 patches and spectral clustering [54], [55] to merge patches. Other alternative algorithms are also feasible.

It is worthwhile to mention that collective motion detection has also some other applications, such as crowd management and human-robot interaction. For example, Arror et al. [56] developped a crowd simulation tool to facilate robot navigation. The proposed framework may be also applicable for these practical tasks.

VIII. Conclusion and Future Work

In this work, the quantification and detection of collective motion is studied. Unlike traditional methods, which neglect the temporal dependency of crowd behaviors, we propose to model individuals movements with a hidden-state model, and compare them with a probability-based similarity calculation method. With the obtained similarity, a structure-based collectiveness measurement is developed to investigate individuals' topological relationship, and quantify the behavior consistency on both individual- and scene-level. Finally, a multi-stage clustering strategy is presented to detect collective motion accurately. Through extensive experiments on various real-world crowd videos, we demonstrate the superiority of the proposed method over the state-of-the-art competitors. As the proposed methodology provides a comprehensive understanding of crowds, it may be also applicable in some crowd-related researches, such as anomaly detection and semantic scene segmentation.

To further verify the effectiveness, we plan to extend the proposed method to more practical applications on crowd surveillance, such as crowd event retrieval, activity recognition and video abstraction. Meanwhile, because feature points are too local, it's also desirable to design more discriminative feature to capture the contextual information of crowds. In addition, one limitation of the proposed frameowrk is its computation complexity, so we also would like to speed up the algorithm in the future work.

References

[1] X. Liu, D. Tao, M. Song, L. Zhang, J. Bu, and C. Chen, "Learning to track multiple targets," *IEEE Transaction on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1060–1073, 2015.
[2] F. Zhu, X. Wang, and N. Yu, "Crowd tracking with dynamic evolution of group structures," in *European Conference on Computer Vision*, 2014, pp. 139–154.
[3] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, 2014.
[4] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8198–8207.
[5] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 833–841.
[6] K. Hsiao, K. Xu, J. Calder, and A. Hero, "Multicriteria similarity-based anomaly detection using pareto depth analysis," *IEEE Transaction Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1307–1321, 2016.
[7] Y. Yuan, J. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 45, no. 3, pp. 562–575, 2015.
[8] Y. Ji, Y. Yang, X. Xu, and H. Shen, "One-shot learning based pattern transition map for action early recognition," *Signal Processing*, vol. 143, pp. 364–370, 2018.
[9] Y. Ji, Y. Yang, F. Shen, H. Shen, and X. Li, "A survey of human action analysis in hri applications," *IEEE Transactions on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2019.2912988, 2019.
[10] B. Zhou, X. Tang, H. Zhang, and X. Wang, "Measuring crowd collectiveness," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1586–1599, 2014.
[11] J. Shao, C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2227–2234.
[12] H. Wang and C. O'Sullivan, "Globally continuous and non-markovian crowd activity analysis from videos," in *European Conference on Computer Vision*, 2016, pp. 527–544.
[13] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3488–3496.
[14] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 46–58, 2020.
[15] W. Lin, Y. Mi, W. Wang, J. Wu, J. Wang, and T. Mei, "A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes," *IEEE Transaction on Image Processing*, vol. 25, no. 4, pp. 1674–1687, 2016.
[16] Q. Wang, M. Chen, and X. Li, "Quantifying and detecting collective motion by manifold learning," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4292–4298.
[17] C. Reynolds, "Flocks, herds and schools: A distributed behavioral model," in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, 1987, pp. 25–34.
[18] M. Moussaïd, S. Garnier, G. Theraulaz, and D. Helbing, "Collective information processing and pattern formation in swarms, flocks, and crowds," *topiCS*, vol. 1, no. 3, pp. 469–497, 2009.
[19] R. Hughes, "The flow of human crowds," *Annual Review of Fluid Mechanics*, vol. 35, no. 1, pp. 169–182, 2003.
[20] W. Ren, S. Li, Q. Guo, G. Li, and J. Zhang, "Agglomerative clustering and collectiveness measure via exponent generating function," *CoRR*, vol. abs/1507.08571, 2015.

[21] Y. Wu, Y. Ye, and C. Zhao, "Coherent motion detection with collective density clustering," in *ACM Conference on Multimedia*, 2015, pp. 361–370.
[22] A. Rodriguez and A. Liao, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.
[23] X. Li, M. Chen, and Q. Wang, "Collectiveness via refined topological similarity," *ACM TOMM*, vol. 12, no. 2, 2016.
[24] B. Zhou, X. Tang, and X. Wang, "Coherent filtering: Detecting coherent motions from crowd clutters," in *European Conference on Computer Vision*, 2012, pp. 857–871.
[25] T. Brox, M. Rousson, R. Deriche, and J. Weickert, "Colour, texture, and motion in level set based segmentation and tracking," *Image Vision Comput.*, vol. 28, no. 3, pp. 376–390, 2010.
[26] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6.
[27] S. Wu and H. Wong, "Crowd motion partitioning in a scattered motion field," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 5, pp. 1443–1454, 2012.
[28] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4147–4153.
[29] W. Ge, R. Collins, and B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, 2012.
[30] W. Choi and S. Savarese, "Understanding collective activitiesof people from videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1242–57, 2014.
[31] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 2999–3007.
[32] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6517–6525.
[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, "SSD: single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.
[34] Y. Song, C. Ma, L. Gong, J. Zhang, R. Lau, and M. Yang, "CREST: convolutional residual learning for visual tracking," in *IEEE International Conference on Computer Vision*, 2017, pp. 2574–2583.
[35] H. Fan and H. Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 5487–5495.
[36] S. Schulter, P. Vernaza, W. Choi, and M. Chandraker, "Deep network flow for multi-object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2730–2739.
[37] T. Senst, V. Eiselein, and T. Sikora, "Robust local optical flow for feature tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1377–1387, 2012.
[38] H. Fradi and J. Dugelay, "Towards crowd density-aware video surveillance applications," *Information Fusion*, vol. 24, pp. 3–15, 2015.
[39] H. Fradi, V. Eiselein, J. L. Dugelay, I. Keller, and T. Sikora, "Spatio-temporal crowd density model in a human detection and tracking framework," *Signal Processing Image Communication*, vol. 31, pp. 100–111, 2015.
[40] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.
[41] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 935–942.
[42] A. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 909–926, 2008.
[43] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.
[44] R. Shumway and D. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *Journal of Time*, vol. 3, no. 4, pp. 253–264, 2010.
[45] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.
[46] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, and A. Procaccini, "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, p. 1232, 2007.
[47] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Olkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, pp. 321–328, 2003.
[48] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Transaction Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1279–1289, 2016.
[49] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 43, no. 2, pp. 660–672, 2013.
[50] B. Schlkopf, J. Platt, and T. Hofmann, "A local learning approach for clustering," in *Advances in Neural Information Processing Systems*, 2006, pp. 1529–1536.
[51] C. Aggarwal, "A human-computer interactive method for projected clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 448–460, 2004.
[52] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
[53] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "S-LIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
[54] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
[55] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2001, pp. 849–856.
[56] A. Aroor, S. Epstein, and R. Korpan, "Mengeros: A crowd simulation tool for autonomous robot navigation," in *AAAI Fall Symposia*, 2017, pp. 123–125.