Detecting Coherent Groups in Crowd Scenes by Multiview Clustering

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

Abstract—Detecting coherent groups is fundamentally important for crowd behavior analysis. In the past few decades, plenty of works have been conducted on this topic, but most of them have limitations due to the insufficient utilization of crowd properties and the arbitrary processing of individuals. In this study, a Multiview-based Parameter Free framework (MPF) is proposed. Based on the L1-norm and L2-norm, we design two versions of the multiview clustering method, which is the main part of the proposed framework. This paper presents the contributions on three aspects: (1) a new structural context descriptor is designed to characterize the structural properties of individuals in crowd scenes; (2) a self-weighted multiview clustering method is proposed to cluster feature points by incorporating their orientation and context similarities; (3) a novel framework is introduced for group detection, which is able to determine the group number automatically without any parameter or threshold to be tuned. The effectiveness of the proposed framework is evaluated on real-world crowd videos, and the experimental results show its promising performance on group detection. In addition, the proposed multiview clustering method is also evaluated on a synthetic dataset and several standard benchmarks, and its superiority over the state-of-the-art competitors is demonstrated.

Index Terms—Crowd analysis, group detection, context descriptor, multiview clustering, graph clustering

1 Introduction

People in crowd scenes tend to connect with the surroundings and form coherent groups. Within each group, the pedestrians share similar motion patterns and exhibit collective behaviors. As the primary components that make up a crowd, groups convey sufficient information about the crowd phenomenon, and provide a mid-level representation of the semantic behaviors. So group detection has become an attractive research area in the field of video surveillance, and been applied into a wide range of video applications, such as event recognition [1], [2], [3], [4], crowd tracking [5], [6], [7], [8], crowd counting [9], [10], [11] and semantic scene segmentation [12]. Though tremendous efforts [12], [13], [14], [15], [16], [17], [18], [19], [20] toward group detection have been made in the past years, there is still room for improvement.

The major difficulty in group detection is that the study object is too microcosmic. Due to the severe occlusion in crowd scenes, many state-of-the-art methods detect and track feature points to avoid identifying pedestrians directly, and then combine those points with similar motions into the same group. However, there are always many points on one pedestrian and the velocities of these points may have big differences. For example, the points on a pedestrian's head may move in the opposite direction to those ones on the feet. This phenomenon is named as *motion deviation* in this paper. Due to motion deviation, the velocities of feature points are too microcosmic to reflect the real movements of pedestrians accurately. Moreover, due to the locality property, the velocity of a feature point may fluctuate dramatically between consecutive frames. Instead of extracting feature points, Solera et al. [21] detected small groups based on the results of pedestrian detector. However, they consider each pedestrian to be a point, and then the influence of velocity fluctuation also exists. Thus, it's necessary to develop a stable descriptor to perceive the pedestrians' motion patterns from the macroscopic view.

In addition, the lack of prior is also a barrier for group detection. Group detection aims to cluster the individuals with similar behaviors. However, unlike standard clustering tasks, the definition of crowd group is relatively subjective, and it is hard to obtain the prior for each crowd scene, such as the desired cluster number [22] and the preference about anchors [23]. Thus, many clustering methods cannot be used in group detection, and some previous works cluster the feature points by thresholding the adjacent graph [12], [16], [17], [18], [19], [20]. This strategy is dominant in group detection, since it doesn't need the prior about the group number and achieves manifest performance on some occasions. However, it's unrealistic to find a threshold that is suitable for all crowds because the crowd density varies across scenes. In addition, these arbitrary clustering approaches neglect the intrinsic correlation inside the adjacent graph. To be specific, if the graph is built with exactly c connected components, the points should be clustered into c groups. But existing works are limited in detecting the groups according to the graph structure.

In this paper, a Multiview-based Parameter Free framework (MPF) is proposed to mitigate the impacts of the above problems. Multiview clustering, which partitions the data by integrating different features, is used for group detection. First, feature points are extracted and considered to be the individuals in crowd scenes. And the orientation and context graphs are built to perceive the individuals' relationship

• *Q. Wang, M. Chen, F. Nie, and X. Li are with the School of Computer Science and with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China. E-mails: crabwq@gmail.com, chenmulin@mail.nwpu.edu.cn, feipingnie@gmail.com, xuelong_li@ieee.org. X. Li and F. Nie are the corresponding authors.*

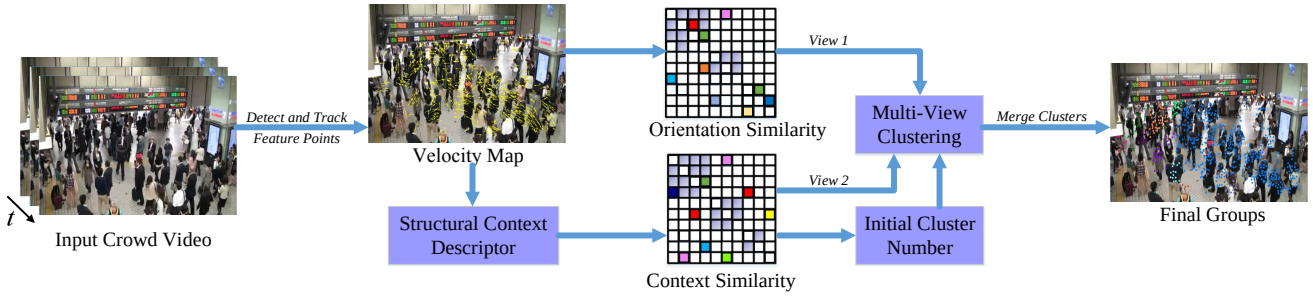


Fig. 1. The pipeline of the proposed framework. First, an orientation graph is built according to the feature points' orientation similarities. Then, a structural context descriptor is proposed to describe the structures of points. Third, the graphs are integrated by a novel self-weighted multiview clustering method. Finally, a merging approach is designed to combine the coherent subgroups.

on both the microcosmic and macroscopic views. Second, with the orientation and context graphs, the subgroups are obtained by a Self-weighted Multiview Clustering (SMC) method. Two versions of SMC are developed based upon the L1-norm and L2-norm. Finally, a tightness-based merging strategy is designed to combine the similar subgroups into final groups. The pipeline of the proposed approach is illustrated in Fig. 1. Our main contributions are summarized as follows.

1. A structural context descriptor is designed to express the structures of feature points. The proposed context descriptor can represent pedestrians' motion dynamics from the macroscopic view and is robust to motion deviation.
2. A self-weighted multiview clustering method is developed to simultaneously integrate the orientation and context relationship of points. Unlike existing multiview clustering approaches, the proposed method doesn't resort to any hyperparameter, and this property makes it applicable for various clustering tasks.
3. A novel group detection framework is proposed, which has salient advantages: (1) the incorporation of features on multiple views; (2) the automatic decision of group number without involving any arbitrary threshold; (3) the capability of handling crowds with varying densities.

Compared to the conference version of this research [24], this paper is substantially improved by introducing more technical parts and providing more experimental evaluations. To be specific: (1) Section 4 proposes the L1-norm version of the Self-weighted Multiview Clustering method, which yields a new clustering objective, and a new optimization algorithm is derived to solve the objective; (2) Section 6.1 evaluates the proposed method on group number estimation; (3) Section 6.2 introduces the experiments on a synthetic dataset to demonstrate the robustness of SMC to data noise. And the converge study of the multiview clustering methods is also added.

The remainder of this paper is organized as follows. Section 2 reviews the previous works on group detection and multiview clustering. Section 3 introduces the details to construct the orientation and context graphs. Section 4 presents the multiview clustering method to cluster the points into subgroups. Section 5 put forward the tightness-based merging strategy to combine the coherent subgroups. In Section 6, experiments are conducted to validate the proposed method. Conclusions and future works are made in Section 7.

2 Related Work

In this section, we first briefly review the related works on group detection, and then some existing multiview clustering methods are introduced.

2.1 Group Detection

Detecting the groups in crowd scenes is a hot topic in video surveillance. According to the type of study object, previous methods can be roughly divided into two classes: 1) fixed particle-based approaches; 2) feature point-based approaches.

For the first kind of methods, they overlay a grid of particles on the crowd scene, and investigate the particles' optical flow by particle advection. Ali and Shah [13] utilized the Lyapunov exponent field to model the particles' motion patterns, and then segmented the coherent flow. Wu and Wong [14] proposed a local-translation domain segmentation model to discover the collective motion. Hu et al. [25] learned the motion patterns in crowds by utilizing the instantaneous motions. Mehran et al. [1] employed the social force model to find the abnormal groups. Mehran et al. [26] designed the streakline descriptor to quantify the flow of particles. Yuan et al. [2] profiled the crowd flow with a potential energy function. Lin et al. [12] used the thermal diffusion method to enhance the flow of particles, and segmented the coherent particles by spectral clustering. The major deficiency of the above methods is that they are time-consuming, since each scene contains several thousands of particles.

As for the second category, the trajectories of points are taken as study objects. Ge et al. [15] detected the small groups of pedestrians by hierarchical clustering. Zhou et al. [16] found the stable neighbors of each point and combined those with similar velocities. Zhou et al. [18] first measured collectiveness of each point by graph learning, and then detected collective motions by thresholding the collectiveness. Shao et al. [17] presented the transition error to refine the groups obtained by Zhou et al. [16]. Wang et

al. [20] introduced an intention-based model to compare the similarity of points, and then combined the similar points into groups. Chen et al. [27] employed the manifold ranking method for group detection. Zhang et al. [28] investigated the spacing interactions of trajectories, and introduced a group sparsity constraint to characterize the coherent motion patterns. Solera et al. [21] employed detection and tracking methods to extract the pedestrians, and then considered each pedestrian as a point and investigated their trajectories.

A drawback shared by all the above methods is that the particles and points are too microcosmic to reveal the real conditions of pedestrians. And most of them involve arbitrary thresholds, so they are impractical for crowd systems with various densities.

2.2 Multiview Clustering

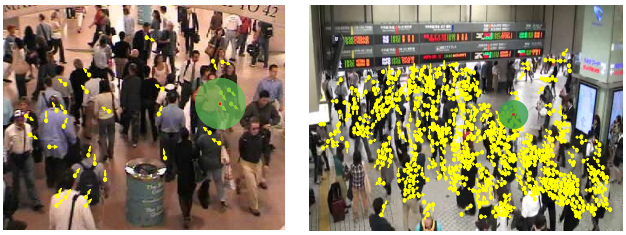
Multiview clustering aims to obtain the consensus clustering results across multiple views, and has inspired a surge of interests [29], [30], [31], [32], [33], [34], [35], [36], [37], [38] in machine learning.

Kumar et al. [29] extended the co-regularization strategy into the spectral clustering scheme to achieve the clustering goal, and let all the views share the same weight. Cai et al. [30] learned a commonly shared Laplacian graph from the multiview features, and proposed a non-negative relaxation to improve the robustness. Xia et al. [31] handled different features in different ways, and found a low-dimensional projection to approximate all the features. Xia et al. [32] learned the transition probability matrix of each view, and put the matrices into a Markov chain to enforce the smoothness. Wahid et al. [33] introduced the formulations of crossover, mutation and tuning steps, and conducted multiview clustering with an evolutionary process. Li et al. [34] integrated the heterogeneous features with local manifold fusion, and approximated the graphs with bipartite graphs to improve the efficiency. So this method can deal with the large-scale problem. Zhang et al. [35] regraded the input graphs as a tensor, and imposed a low-rank constraint to exploit the complementary information from different views. Li et al. [36] combined the local kernel alignment technique into the multiview clustering framework to preserve the local data structure. Liu et al. [37] considered the correlation among different views, and designed a matrix-induced regularization to emphasize the diversity of information sources. Instead of assuming the optimal graph to be a linear combination of the input graphs, Liu et al. [38] proposed an optimal neighborhood clustering method to learn the optimal graph within the neighborhood of the original graphs, which enhances the representability of the optimal graph.

All of these methods resort to additional parameters, which affects the performance directly and restricts the applicability to process various kinds of data.

3 Graph Construction

In this section, the orientation and context graphs are constructed to capture the correlations of individuals. First, due to the difficulty of detection and tracking in crowd scenes, feature points are regarded as individuals. In order to reveal the orientation similarity of points, the orientation graph is built adaptively based on the crowd density. Then, a novel context descriptor is designed to profile the structural properties of points, and a context graph is constructed to capture the points' relationship from the macroscopic view.



(A) Low frame rate (B) High frame rate

Fig. 2. Crowd frames with low and high frame rates. Yellow points indicate the feature points, and the green circles indicate the neighborhood of the corresponding points (red color).

3.1 Adaptive Orientation Description

To capture the underlying moving principle inside a crowd scene, it's essential to identify the pedestrians. Due to the serious occlusion and noise in crowds, it's impractical to extract pedestrians directly. So we alternatively take feature points as study objects. The generalized Kandae-Lucas-Tomasi (gKLT) tracker [18] is used to extract the feature points, since it performs detection and tracking jointly with efficient computation. As pointed out by behaviorists [39], instead of keeping connections with all the others, individuals in crowds only interact with their local neighbors. Thus, we need to find the neighbor relationship between the points.

Most of the existing methods [15], [16], [17], [18], [20] find the neighbors of each point by using kNN method, which involves a parameter k. However, the value of k may influence the overall performance greatly. And it's infeasible to choose a suitable k for the crowds with varying densities. So we propose to find the neighbors of each point adaptively according to the crowd density. Considering a frame with n points, the spatial position of a point i $(i = 1, 2, ..., n)$ is denoted as (p_i^x, p_i^y), and its orientation is denoted as $\overrightarrow{ori_i} = (ori_i^x, ori_i^y)$. Then the spatial distance between points i and j is computed as

$$D(i,j) = \sqrt{(p_i^x - p_j^x)^2 + (p_i^y - p_j^y)^2}. \quad (1)$$

Suppose there exists a variable r, and points i and j are considered as neighbors if their distance $D(i,j)$ is smaller than r. Then the orientation graph G_m can be calculated as

$$G_m(i,j) = \begin{cases} max(\frac{\overrightarrow{ori_i} \cdot \overrightarrow{ori_j}}{|\overrightarrow{ori_i}| \times |\overrightarrow{ori_j}|}, 0), & \text{if } D(i,j) < r \\ 0, & \text{else} \end{cases}, \quad (2)$$

where the $max()$ function prevents the similarity from being negative. And the orientation similarity is 0 for the points without neighboring relationship.

In Eq. (2), the value of r is crucial for the computation. In this work, r is empirically set as the n-th smallest element in all pairs of the distance D. Throughout experiments, we

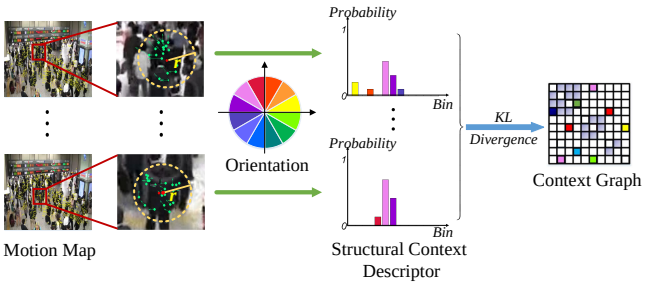


Fig. 3. Details about the construction of context graph.

find this setting is appropriate. Specifically, when n is fixed, a higher point density corresponds to a smaller r, which complies with the fact that the neighbors should reside within a small radius if the points are with a high density. In addition, existing tracking methods [40], [41] are limited in dealing with the large variation between consecutive frames. Thus, for videos with a low frame rate, there may be only a few feature points although the crowd density is high, as shown in Fig. 2 (A). In these occasions, the incorporation of n prevents the value of r from being too large.

3.2 Structural Context Description

After the above stage, the orientation similarity of points is captured. However, due to the motion deviation, the velocities of the points are too microcosmic to reveal the crowd condition. So we propose to profile the feature points from the macroscopic view. As aforementioned, points in crowd scenes keep close relationship with their surroundings, so the structural property of a point can be represented by its neighbors. To this end, a novel Structural Context (SC) descriptor is formulated to express the structure of each point.

For each point i, its neighbor set C is obtained by including the points within the radius r, as mentioned before. Then, we divide the orientation space into 12 bins, as shown in Fig. 3. Thus, the SC of i is defined as a vector with 12 elements, with its m-th element denoted as

$$SC_i(m) = p(\overrightarrow{ori_a} \in bin_m | a \in C), \tag{3}$$

where $p(\cdot)$ indicates the probability, and bin_m is the m-th orientation bin.

SC is exactly the distribution of neighbors' orientations over the divided orientation space, so it can reveal the structural properties of points. Our assumption behind this descriptor is that when a point's velocity fluctuates, the neighbors' velocities can assist to reveal its real condition. Given the SC of each point, a context graph can be constructed as below

$$G_c(i,j) = exp\{-\frac{1}{2}[\mathrm{KL}(SC_i||SC_j) + \mathrm{KL}(SC_j||SC_i)]\}, \tag{4}$$

where $\mathrm{KL}(SC_i||SC_j) = \sum_{m=1}^{12} SC_i(m)\log\frac{SC_i(m)}{SC_j(m)}$ is the Kullback Leibler (KL) divergence between the SC_i and SC_j [42]. Thus, the context graph is capable to describe the similarity of points' structures.

4 Self-weighted Multiview Clustering

In this section, we design two versions of the Self-weighted Multivew Clustering method to cluster the points into subgroups. The optimization algorithms for the proposed two objectives are also introduced. The proposed method learns the weights of different views adaptively according to the graph structure, so it's parameter free.

4.1 Multiview Clustering Formulations

Generally speaking, group detection can be interpreted as the clustering of points. In this part, both the orientation and context graphs are integrated to cluster the points. We first briefly review the Constrained Laplacian Rank (CLR) method [22], which conducts clustering task based on a single-view graph. Supposing there are n samples to be classified into c clusters, the objective of CLR is

$$\min_{\sum_j S_{ij}=1, S_{ij}\geqslant 0, rank(L_S)=n-c} ||S-G||_F^2, \tag{5}$$

where $S \in \mathbb{R}^{n\times n}$ is a target graph with exactly c connected components, and $||\cdot||_F$ is the Frobenius Norm. $G \in \mathbb{R}^{n\times n}$ is the input graph, which indicates the similarity of samples. And $L_S = D_S - (S^T+S)/2 \in \mathbb{R}^{n\times n}$ is the Laplacian matrix of S. The rank constraint $rank(L_S) = n-c$ guarantees that S contains exact c connected components, corresponding to the desired c clusters. Therefore, the clustering objective can be achieved as long as the optimal S is obtained. The superiority of CLR can be summarized from two aspects: 1) it performs well even when the input graph is constructed with low quality; 2) unlike other spectral-based clustering methods [30], [32], [43], it doesn't need any post-processing.

To investigate the data correlation captured from different aspects, we extend CLR to the multiview clustering scheme. Denoting n and n_v as the number of samples and views respectively, the graphs corresponding to the n_v views are written as $G^{(1)}, G^{(2)}, ..., G^{(n_v)} \in \mathbb{R}^{n\times n}$. Different from problem (5), we aim to find a S that approximates each of the graphs, so the optimization problem is

$$\begin{aligned} &J_{L2} = \min_{w^{(v)},S} ||S - \sum_{v=1}^{n_v} w^{(v)} G^{(v)}||_F^2 \\ &s.t. w^{(v)} \geqslant 0, \sum\nolimits_v w^{(v)} = 1, S_{ij} \geqslant 0, \\ &\sum\nolimits_j S_{ij} = 1, rank(L_S) = n-c, \end{aligned} \tag{6}$$

where scalar variable $w^{(v)}$ is the weight of the graph $G^{(v)}$.

In addition, since L1-norm distance is more robust to noise compared with L2-norm distance, we further define a L1-norm objective:

$$\begin{aligned} &J_{L1} = \min_{w^{(v)},S} ||S - \sum_{v=1}^{n_v} w^{(v)} G^{(v)}||_1 \\ &s.t. w^{(v)} \geqslant 0, \sum\nolimits_v w^{(v)} = 1, S_{ij} \geqslant 0, \\ &\sum\nolimits_j S_{ij} = 1, rank(L_S) = n-c, \end{aligned} \tag{7}$$

where $||\cdot||_1$ is the L1-norm.

4.2 Optimization Algorithms

Without prior knowledge, an intuitive idea is assigning the equal weight to each graph, just as [29]. However, this

strategy ignores the diversity of different views and tends to be gravely affected when some views perform badly. Thus, we aim to approximate the graphs with different confidences. For this purpose, two self-conducted weight learning algorithms are proposed to solve problem (7) and (6) respectively.

Optimization Algorithm for Solving J_{L2} in Eq. (6)

Eq. (7) is difficult to solve because the constraint $rank(L_S) = n - c$ is a nonlinear constraint. According to Nie et al. [22], enforcing the rank constraint is equivalent to solving the following problem

$$\min_{F\in\mathbb{R}^{n\times c},F^TF=I} Tr(F^TL_SF), \tag{8}$$

where $Tr()$ is the trace operator, and $\mathbb{R}^{n\times n}$ is the identity matrix. Therefore, problem (6) can be transformed into the following problem

$$\begin{aligned}&\min_{w^{(v)},S,F} ||S-\sum_{v=1}^{n_v} w^{(v)}G^{(v)}||_F^2+\lambda Tr(F^TL_SF)\\ &s.t. w^{(v)}\geqslant 0, \sum\nolimits_v w^{(v)}=1, S_{ij}\geqslant 0,\\ &\sum\nolimits_j S_{ij}=1, F\in\mathbb{R}^{n\times c}, F^TF=I,\end{aligned} \tag{9}$$

where the parameter λ can be tuned automatically in a heuristic way according to the number of connected components in L_S [22]. Then we propose to optimize $w^{(v)}$, S and F iteratively.

When $w^{(v)}$ and S are fixed, problem (9) becomes problem (8). According to the spectral theory [43], the optimal F is formed by the c eigenvectors of L_S corresponding to its k smallest eigenvalues.

When w and F are fixed, problem (9) becomes the following problem

$$\min_{\sum_j S_{ij}=1,S_{ij}\geqslant 0} ||S-\sum_{v=1}^{n_v} w^{(v)}G^{(v)}||_F^2+\lambda Tr(F^TL_SF), \tag{10}$$

and the details of optimization can be referred to [22].

When S and F are fixed, the problem seems complicated to solve because it can't be directly decoupled into rows. So we transform problem (4) into a different form, which is a crucial step for the optimization. The target graph S is first converted into a column vector $a\in\mathbb{R}^{n^2\times 1}$, and the input graphs $G^{(1)}, G^{(2)}, ..., G^{(n_v)}$ are also converted into $B^{(1)}, B^{(2)}, ..., B^{(n_v)}\in\mathbb{R}^{n^2\times 1}$. Denoting a matrix $B\in\mathbb{R}^{n^2\times n_v}$ with its v-th column equal to $B^{(v)}$, and denoting a vector $w=[w^{(1)}, w^{(2)}, ..., w^{(n_v)}]^T\in\mathbb{R}^{n_v\times 1}$, Eq. (9) naturally becomes a vector form problem

$$\min_{w\mathbf{1}=1,w\geqslant 0} ||a-Bw||_2^2, \tag{11}$$

which is much easier to solve, and $\mathbf{1}\in\mathbb{R}^{n_v\times 1}$ is the vector with all the elements equal to 1. Spreading the terms in Eq. (11), the problems becomes

$$\min_{w\mathbf{1}=1,w\geqslant 0} \frac{1}{2}w^TB^TBw-w^TB^Ta. \tag{12}$$

The above function is a standard quadratic programming (QP) problem, which can be readily solved by an iterative algorithm [22].

The detailed algorithm for solving problem (6) is provided in Algorithm 1.

Optimization Algorithm for Solving J_{L1} in Eq. (7)

Similar to the above transformation, problem (7) is equivalent to the following problem

$$\begin{aligned}&\min_{w^{(v)},S,F} ||S-\sum_{v=1}^{n_v} w^{(v)}G^{(v)}||_1+\lambda Tr(F^TL_SF)\\ &s.t. w^{(v)}\geqslant 0, \sum\nolimits_v w^{(v)}=1, S_{ij}\geqslant 0,\\ &\sum\nolimits_j S_{ij}=1, F\in\mathbb{R}^{n\times c}, F^TF=I,\end{aligned} \tag{13}$$

During the optimization, F is updated as in Eq. (8). And when w and F are fixed, problem (13) becomes the following problem

$$\min_{\sum_j S_{ij}=1,S_{ij}\geqslant 0} ||S-\sum_{v=1}^{n_v} w^{(v)}G^{(v)}||_1+\lambda Tr(F^TL_SF), \tag{14}$$

which can be optimized by the L1-norm solution in [22].

When S and F are fixed, problem (13) is transformed into

$$\begin{aligned}&\min_{w^{(v)}} ||S-\sum_{v=1}^{n_v} w^{(v)}G^{(v)}||_1\\ &s.t. w^{(v)}\geqslant 0, \sum\nolimits_v w^{(v)}=1.\end{aligned} \tag{15}$$

Denoting a, B and w as the same definition in the L2-norm solution, the above problem is simplified into the following

$$\min_{w\mathbf{1}=1,w\geqslant 0} ||a-Bw||_1. \tag{16}$$

With the iterative reweighted method [44], the above problem can be optimized by solving the following one iteratively:

$$\min_{w\mathbf{1}=1,w\geqslant 0} Tr(a-Bw)^TU(a-Bw), \tag{17}$$

where $U\in\mathbb{R}^{n\times n}$ is a diagonal matrix with its i-th diagonal element as $\frac{1}{2|a_i-B_i\widetilde{w}|}$, and $\widetilde{w}$ is the current solution of w. Nie et al. [44] have proved that the iterative method will finally converge to the optimal solution of problem (16). Problem (17) can be spread into the following form

$$\min_{w\mathbf{1}=1,w\geqslant 0} \frac{1}{2}w^TB^TUBw-w^TB^TUa, \tag{18}$$

which is with the similar form to problem (12). Algorithm 2 provides the detailed algorithm to solve problem (7).

With the suggested optimization algorithm, given an initial w, the closed form solution of problem (6) and (7) can be computed by updating S, F and w alternately until convergence. Different from existing multiview clustering algorithms [29], [30], [31], [32], [33], [34], [37], [38], the proposed method is totally self-weighted, and doesn't resort to any hyper parameter. This property is promising because we do not need to tune those additional parameters when handling various crowds. And the convergence study of the optimization algorithms will be given in Section 6.2.

4.3 Discussion

In the group detection task, there are two views to be learned, so the weight vector is initialized as $[\frac{1}{2}, \frac{1}{2}]^T$. The cluster number c is set as the number of strongly connected components in the context graph, which can be efficiently

Algorithm 1 Algorithm to solve problem (6)

Input: Data graphs $G^{(1)}, ..., G^{(2)}$, cluster number c, number of views n_v.

1: Initialize $w^{(1)}, ..., w^{(n_v)}$ as $\frac{1}{n_v}$.
2: Initialize S as $\sum_{v=1}^{n_v} w^{(v)} G^{(v)}$.
3: **repeat**
4: Update F by solving Eq. (8).
5: Update S by solving problem (10).
6: Convert S into a column vector a, convert $G^{(1)}, ..., G^{(2)}$ into a matrix B, denot a vector $w = [w^{(1)}, ..., w^{(n_v)}]$, update w by solving problem (12).
7: **until** Converge

Algorithm 2 Algorithm to solve problem (7)

Input: Data graphs $G^{(1)}, ..., G^{(2)}$, cluster number c, number of views n_v.

1: Initialize $w^{(1)}, ..., w^{(n_v)}$ as $\frac{1}{n_v}$.
2: Initialize S as $\sum_{v=1}^{n_v} w^{(v)} G^{(v)}$.
3: **repeat**
4: Update F by solving Eq. (8).
5: Update S by solving problem (14).
6: Convert S into a column vector a, convert $G^{(1)}, ..., G^{(2)}$ into a matrix B, denot a vector $w = [w^{(1)}, ..., w^{(n_v)}]$, update w by solving problem (18).
7: **until** Converge

computed by the Depth First Search method [45]. Then, graphs on both the orientation and context views are integrated to learn the target graph S by solving problem (6) or (7). The obtained S contains exact c connected components, so it can also be considered as an indicator matrix, where $S_{ij} > 0$ indicates that points i and j belong to the same cluster. Since S assigns a cluster index to each point, the clustering procedure is accomplished immediately when the optimal solution of problem (6) or (7) is acquired. However, in crowd scenes, not all the points in one group keep close connections with each other, and they are actually united in a weakly connected component. When calculating c, a weakly connected component may be split into several strongly connected ones, leading to an overestimation of cluster number. Thus, it's necessary to merge the obtained subgroups that actually belong to the same group.

5 Tightness-based Merging

To combine the coherent subgroups acquired by the previous stage, a tightness-based cluster merging strategy is put forward. Denoting the learned weights of the orientation and context graph as w_m and w_c respectively, an integrated graph is presented as

$$G = w_m G_m + w_c G_c. \tag{19}$$

The graph G approximates both the orientation and context graph of points. The reason that we don't use the learned target graph S is that due to the rank constraint, the similarity in S is 0 for points clustered into different subgroups. So S is unsuitable to decide whether two subgroups are consistent.

(A) Subgroups (B) Final Groups

Fig. 4. Comparison of groups (A) before and (B) after merging. Scatters with different colors indicate different detected groups, and arrows indicate motion orientations.

Inspired by the theory that connection of pedestrians leads to the emergence of groups [39], a tightness measure is put forward to capture the intra-relationship of subgroups. We assume there exists an anchor point within each subgroup, which could reflect the motion pattern of the corresponding subgroup. Then the tightness of a subgroup is considered to be the behavior consistency between the anchor point and the others.

With the weight graph G, we find the anchor within each subgroup. First, the collectiveness is calculated for each point, which describes the consistency between the corresponding point and all the others in the same subgroup. Denoting a subgroup as sub_α, the collectiveness of a point i within sub_α is

$$\phi_i = \sum_{j \in sub_\alpha} G(i, j). \tag{20}$$

The anchor point is assumed to be consistent with others and surrounded by many neighbors. Denoting the anchor of sub_α as q, we can locate it according to its collectiveness and number of neighbors,

$$q = \max_{i \in sub_\alpha} (\phi_i + \delta_i), \tag{21}$$

where δ_i records the number of i's neighbors. Thus, the tightness T of sub_α is the collectiveness of its anchor point q,

$$T(sub_\alpha) = \phi_q. \tag{22}$$

With all the above quantitative definitions, we can target on the merging of subgroups. If the merging of two subgroups will produce a higher tightness, then the subgroups are supposed to be coherent. Two subgroups sub_α and sub_β are consistent if

$$T(sub_\alpha + sub_\beta) > max[T(sub_\alpha), T(sub_\beta)]. \tag{23}$$

Through experiments, we have found that the anchor always resides in the center of a subgroup. A higher tightness means that the center of the merged group has a higher collectiveness than the original anchors, so the subgroups are coherent. By merging consistent subgroups iteratively, the final groups are obtained. Since the sequence of merging will affect the result, we only combine the pair of subgroups with the highest value of $T(sub_\alpha + sub_\beta)$ in each iteration.

Benefiting from the merging operation, local coherent motions are automatically combined into global motions, as shown in Fig. 4. The merging procedure stops when no coherent subgroups are qualified to be combined, so

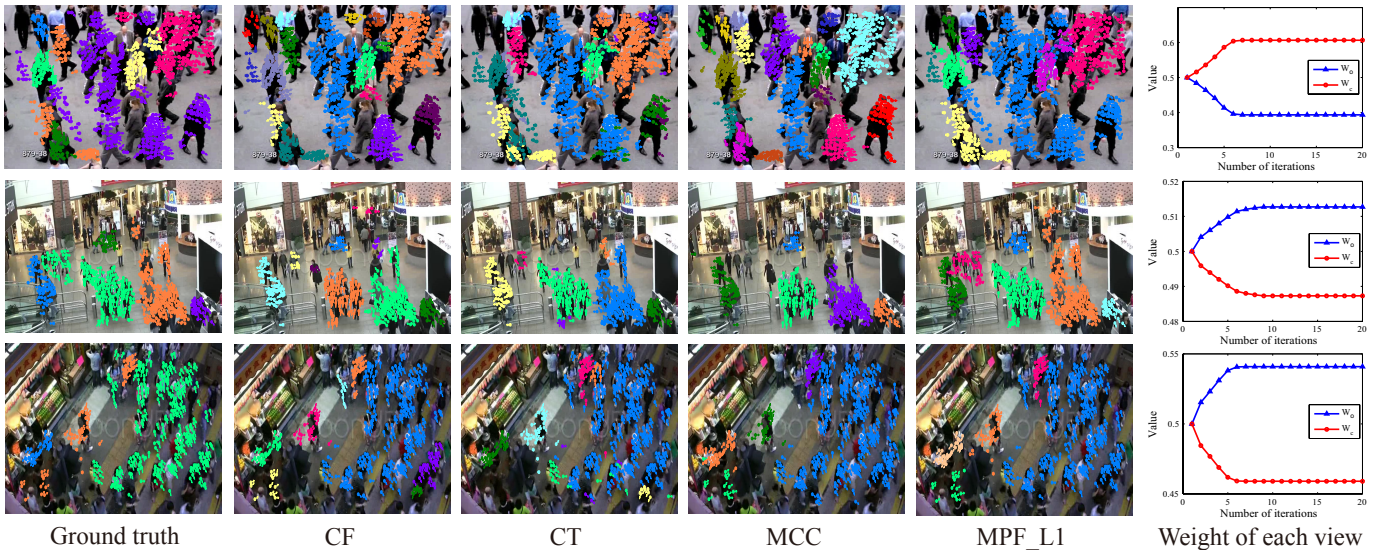
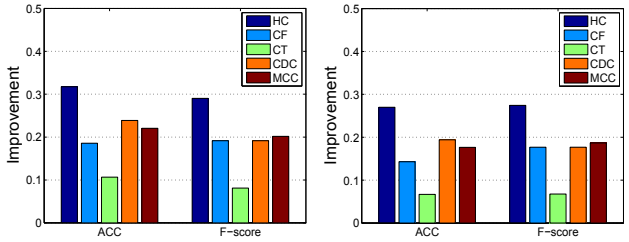


Fig. 5. Representative results of different group detection methods, and the changing trends of w_o and w_c with the number of iterations. Scatters with different colors indicate different detected groups, and arrows indicate motion orientations. The results of MPF_L1 are consistent with the ground truth.

TABLE 1
Quantitative comparison on group detection. Best results are in bold face, and the second-best results are underlined.

	HC	CF	CT	CDC	MCC	MPF_L1	MPF_L2
ACC	0.63	0.70	0.75	0.67	0.68	**0.83**	<u>0.80</u>
F-score	0.62	0.67	0.74	0.67	0.67	**0.80**	<u>0.79</u>



(A) MPF_L1 (B) MPF_L2

Fig. 6. The relative improvements of (A) MPF_L1 and (B) MPF_L2 compared with HC, CF, CT, CDC and MCC.

it provides a more principled termination criterion than choosing an arbitrary threshold manually [16], [17], [18], [20].

6 Experiments

In this section, the proposed group detection framework is evaluated on group detection and group number estimation. And the effectiveness of the proposed multiview clustering method is also demonstrated. Throughout all the experiments, we let the competitors use their respective optimal parameters.

6.1 Evaluation of the Multivew-based Parameter Free Framework

In this work, the CUHK Crowd Dataset [17] is used to verify the proposed framework's performance on group detection and group number estimation. Four state-of-the-art group detection methods are chosen for comparison. The group detection framework with L1-norm and L2-norm clustering objectives are termed as MPF_L1 and MPF_L2 respectively. Two widely used metrics, the accuracy (ACC) [46] and F-score [32] are taken as measurements to evaluate the methods quantitatively.

Dataset: CUHK Crowd Dataset consists of 474 crowd videos, where the frame rate varies from 20 to 30 fps. And the crowd densities and perspective scales are different. Group label for each feature point is annotated by human observers. We perform group detection on every video and average the obtained ACC and F-score as experimental results.

Competitors: To demonstrate the effectiveness of the proposed group detection framework, five state-of-the-art methods, Hierarchical Clustering (HC) [15], Coherent Filtering (CF) [16], Collective Transition (CT) [17], Measuring Crowd Collectiveness (MCC) [18] and Coherent Density Clustering (CDC) [19], are taken for comparison.

Performance on Group Detection

The group detection results of different methods are shown in Table 1. And the improvements of MPF_L1 and MPF_L2 over the competitors are visualized in Fig. 6. It's manifest that the proposed MPF_L1 and MPF_L2 achieve the highest two ACC and F-score, which indicates that they perform better than the competitors. HC clusters the trajectories of points hierarchically according to their Hausdorff distances. CF and CT detect groups by extracting the invariant neighbors of each point. MCC detects collective motions by thresholding the collectiveness of points. CDC employs a density-based clustering approach to cluster points. All the above methods only utilize the orientation feature, and

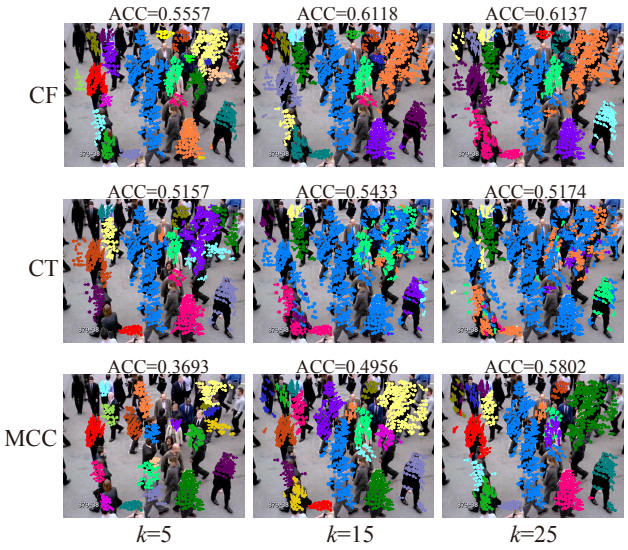


Fig. 7. Performance of CF, CT and MCC with varying k. Scatters with different colors indicate different detected groups, and arrows indicate motion orientations.

TABLE 2
Comparison of MPF_L2 and its variants. Best results are in bold face.

	MPF_L2	r=15	r=25	r=35	View-m	View-c
ACC	**0.80**	0.71	0.78	0.71	0.72	0.70
F-score	**0.79**	0.72	0.78	0.72	0.73	0.69

neglect the structural properties of points. So they tend to be affected by motion deviation and the fluctuation of points' velocities. The proposed MPF_L1 and MPF_L2 jointly incorporate the orientation and context features with a multiview clustering method, so they are able to perceive the motion patterns more accurately, and detect groups more correctly. Fig. 5 shows some representative results of different group detection methods. We can see that the results of MPF_L1 is more consistent with the ground truth. Note that, the performance of MPF_L2 is slightly inferior to that of MPF_L1 because the Frobenius norm squares the residue error of each element, so MPF_L2 is prone to be affected by the outliers in the data graphs, which may be caused by velocity fluctuation.

The proposed MPF_L1 and MPF_L2 share the promising property that no parameter or threshold is involved. To better illustrate the importance of this property, we compare the results of CF, CT and MCC with varying parameter. The above three methods are chosen because they all involve a kNN processing. Fig. 7 shows the clustering results of CF, CT and MCC on a video clip where k is set as 5, 15 and 25. The corresponding ACC is also exhibited. The results show that the performance of these three methods is sensitive to the value of k. For crowd motions with various densities, it's hard to chose an appropriate k that satisfies all occasions. Though CDC doesn't need the kNN procedure, it has several additional thresholds to be tuned, so it's not applicable as well. The proposed MPF_L1 and MPF_L2 avoid this problem naturally because it's totally parameter free.

Fig. 8. Representative results of MPF_L2 on MPT-20x100 dataset. Each point corresponds to a pedestrian (points are on the bottom of pedestrians). Scatters with different colors indicate different detected groups, and arrows indicate motion orientations.

We also show the performance of utilizing orientation view and context view separately, denoted as View-o and View-c. Here the L2-norm clustering objective is used. As exhibited in Table 2, View-o achieves better results than View-c. This result doesn't mean that context feature fails on all videos. We visualize the changing trend of w_o and w_c versus the number of iterations in Fig. 5. For the scene in the first row, the value of w_c increases while w_o decreases. On this occasion, context feature captures the pedestrians' movements better because that there are many points on the same pedestrian, and the motion deviation is serious. Meanwhile, for the videos captured from the overlooking perspective (second and third rows in Fig. 5), orientation feature performs better since the pedestrians are small and their velocities can be approximated by those of feature points. Besides, Table 2 shows MPF_L2 is better than View-m and View-c, so we conclude that the proposed Structural Context descriptor (SC) assists the orientation aspect, and the combination of them is reasonable. Moreover, when there is only one view, the clustering method exactly reduces to the original CLR. And the superior performance of MPF_L2 demonstrates that the proposed multiview clustering method is more useful than CLR because it is able to integrate the features captured from multiple views.

In addition, MPF sets the relationship threshold r as the the n-th smallest element in D. To demonstrate the validity of this adaptive setting, we fix r as 15, 25 and 35, and show the corresponding performance in Table 2. When r equals to 15 and 35, the performance drops dramatically. This is because a small value of r makes a group divided into parts, while a large r brings some noise. The performance is relatively well when r is 25, but it's not so good as MPF_L2 because a fixed r can't be suitable for crowd videos with various densities and frame rates. Therefore, the adaptive decision of parameter r does improve the overall performance of the MPF_L2.

MPF performs well when clustering feature points, and we further show its ability to handle the pedestrians directly. Here we run the proposed method on the MPT-20x100 dataset [21], which contains 20 crowd videos and provides the trajectory of each pedestrian. Some representative results of MPF_L2 on MPT-20x100 dataset are shown in Fig. 8. It can be seen that the proposed method correctly connects the pedestrians with similar motion patterns. The number of pedestrians is much less than that of feature points, but MPF still performs well because it is parameter-free and can deal

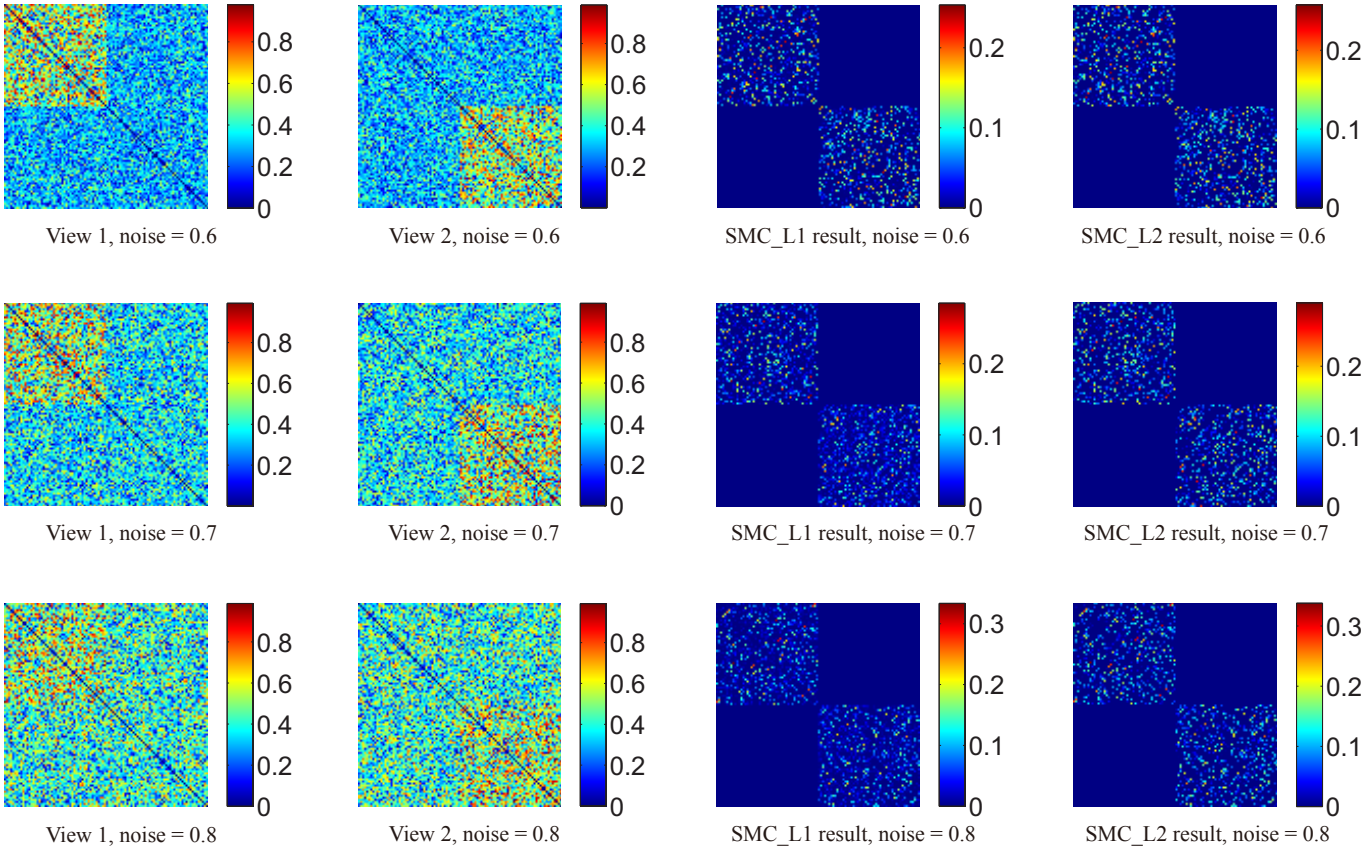


Fig. 9. Clustering results on the block diagonal synthetic data by SMC_L1 and SMC_L2 methods.

TABLE 3
Performance on group number estimation. Best results are in bold face, and the second-best results are underlined.

	HC	CF	CT	CDC	MCC	MPF_L1	MPF_L2
AD	2.83	2.45	1.63	1.59	2.02	<u>1.16</u>	**1.12**
VAR	3.41	3.01	1.83	1.84	2.56	**1.27**	<u>1.35</u>

with the trajectories with various densities. So our method can also deal with the pedestrians' trajectories directly.

Performance on Group Number Estimation

The proposed group detection methods decide the group number automatically without any arbitrary processing. Here we validate their performance on group number estimation. Two widely used metrics, namely Average Difference (AD) and Variance (VAR) [19] are used to evaluate the performance. A lower AD means a lower deviation from the ground truth, and a low value of VAR indicates a stable performance. The experimental results are shown in Table 3. The AD and VAR of MPF_L1 and MPF_L2 are lower than the other methods. HC, CF and MCC fail because they roughly combine the points with similar velocities into the same group, and can't recognize the subtle differences of points' motion patterns. CT refines the results of CF by introducing the transition error, so it performs better than CF. CDC detects groups with a multi-stage clustering strategy, which is able to detect both the local and global coherent motion, so it also performs well. MPF_L1 and MPF_L2 first cluster the points into subgroups, and then combine the coherent subgroups according to their tightness, so they are capable of deciding the group number precisely and achieve the best results.

Computational Efficiency

In Section 2, we have mentioned that the particle-based approaches are time-consuming. Here we demonstrate this statement by comparing the proposed methods with two classical particle-based methods, including Lagrangian Coherent Structures (LCS) [13] and Streakline [26]. Denoting the three frames in Fig. 5 as Frame 1, Frame 2 and Frame 3 respectively (from top row to bottom row), we run MPF_L1, MPF_L2, LCS and Streakline on these frames and show the time costs in Table 4. When calculating the time cost, we ignore the feature extraction stage and just consider the computation of group detection. As shown in Table 4, the efficiency of the particle-based methods has great relation with the frame size. Larger frame size brings more particles to be processed. On the other hand, the time costs of both MPF_L1 and MPF_L2 depend on the number of feature points. Since the amount of feature points is much less than that of particles, the proposed methods are more efficient than the particle-based algorithms.

6.2 Evaluation of Self-weighted Multiview Clustering method

In this part, experiments are conducted on a synthetic dataset and four real-world datasets to demonstrate the

TABLE 4
Efficiency comparison with particle-based methods. Best results are in bold face.

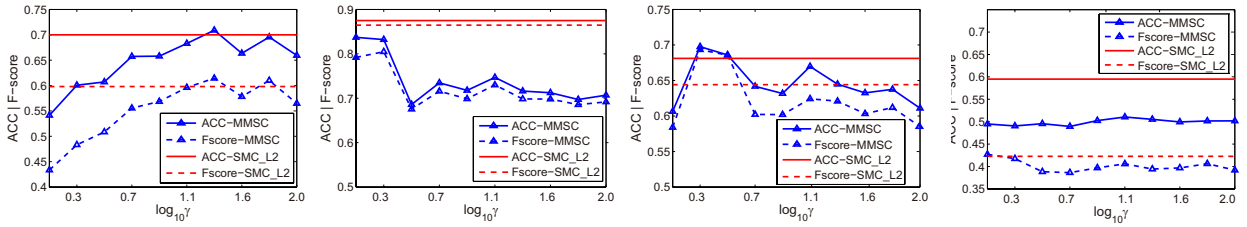
Frames	Point number	Frame size	LCS	Streakline	MPF_L1	MPF_L2
Frame 1	1029	480×640	13.93s	62.84s	0.93s	**0.91**s
Frame 2	597	480×856	19.14s	80.63s	**0.36**s	**0.36**s
Frame 3	460	480×856	18.53s	79.17s	0.25s	**0.24**s

TABLE 5
ACC (mean ± standard deviation%) on multiview datasets. Best results are in bold face.

Dataset	Co-reg	RMSC	MMSC	AMGL	IVA	SMC_L1	SMC_L2
MSRC-v1	70.00±5.50	67.41±4.91	71.00±4.46	72.30±4.59	73.43±4.72	**78.10**	70.00
Digits	79.39±4.98	77.40±5.23	83.75±9.04	72.91±9.23	85.41±5.29	**87.65**	87.50
Caltech101-7	42.52±3.04	58.55±2.79	69.76±3.96	56.33±5.71	68.51±3.74	**80.73**	68.11
Caltech101-20	48.18±4.16	51.22±3.03	51.04±3.81	44.12±4.67	50.32±4.28	58.63	**59.51**

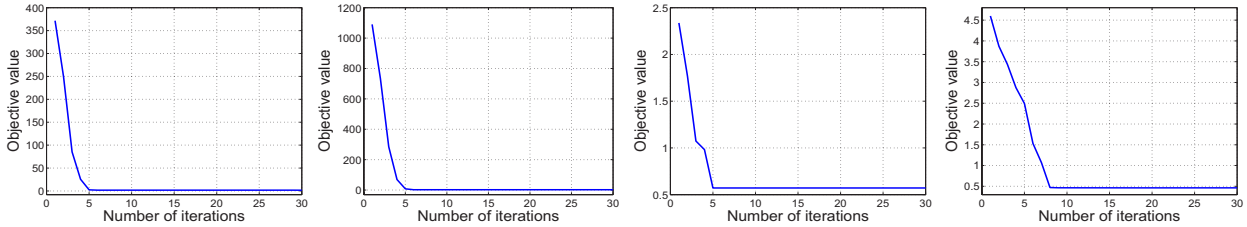
TABLE 6
F-score (mean ± standard deviation%) on multiview datasets. Best results are in bold face.

Dataset	Co-reg	RMSC	MMSC	AMGL	IVA	SMC_L1	SMC_L2
MSRC-v1	59.05±5.02	59.37±3.49	61.44±5.42	61.52±2.40	62.25±3.01	**71.06**	59.81
Digits	71.93±2.37	68.98±3.78	79.20±8.36	71.65±8.59	84.32±6.59	86.31	**86.46**
Caltech101-7	44.50±2.93	55.66±1.44	69.34±4.64	57.96±3.78	69.10±2.73	**76.93**	64.11
Caltech101-20	39.12±3.49	46.21±2.33	40.59±4.10	37.25±3.88	46.59±3.10	**48.33**	42.27



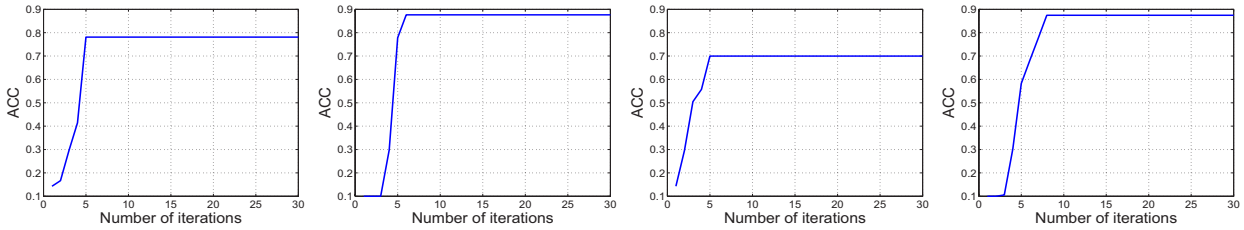
(A) MSRC-v1 (B) Digits (C) Caltech101-7 (D) Caltech101-20

Fig. 10. Performance comparison of MMSC and SMC_L2 on four datasets. We can see that MMSC is sensitive to the value of γ, while SMC_L2 sustains good performance on different datasets.



(A) SMC_L1 on MSRC-v1 (B) SMC_L1 on Digits (C) SMC_L2 on MSRC-v1 (D) SMC_L2 on Digits

Fig. 11. Convergence analysis of SMC_L1 and SMC_L2 on MSRC-v1 and Digits datasets.



(A) SMC_L1 on MSRC-v1 (B) SMC_L1 on Digits (C) SMC_L2 on MSRC-v1 (D) SMC_L2 on Digits

Fig. 12. ACC curves of SMC_L1 and SMC_L2 on MSRC-v1 and Digits datasets.

effectiveness of the proposed Self-weighted Multiview Clustering (SMC) method. The L1-norm and L2-norm versions of SMC are denoted as SMC_L1 and SMC_L1 respectively.

Block Diagonal Synthetic Dataset

Datasets: The synthetic dataset consists of two 100×100 matrices, as shown in the first two columns of Fig. 9. View 1 denotes the affinities of the points in the first cluster (1 to 50), and View 2 records the affinities in the second cluster (51 to 100). Each cluster corresponds to a block, and the affinity data within each block is randomly generated in the range of 0 and 1. And the noise data outside the blocks is generated in the range of 0 and σ. In the first, second and third rows of Fig. 9, the noise percent σ is set as 0.6, 0.7 and 0.8 respectively. An ideal multiview clustering method should integrate the two matrices and produce two clusters.

Performance: The clustering results of SMC_L1 and SMC_L2 are shown in the third and fourth columns in Fig. 9. Under different percent of noises, both SMC_L1 and SMC_L2 accurately partition the points into the correct clusters. SMC_L1 and SMC_L2 learn the optimal graph by exploiting the correlation of input graphs adaptively, so both of them are able to recover points' relationship from multiple views, and are robust to noise.

Real-world Datasets

Datasets: Four standard real-world datasets are employed to evaluate the proposed SMC, including MSRC-v1 [47], Digits [48], Caltech101-7 and Caltech101-20 [49]. The details of the datasets are as follows.

- MSRC-v1. Following Nie et al. [50], we choose 210 images from 7 classes in the MSRC-v1 dataset, and extract 5 visual features (CMT, HOG, LBP, GIST and CENT).
- Digits. This dataset consists of 2000 handwritten images from 10 classes, and 6 features (FOU, FAC, KAR, PIX MOR and ZER) are released for clustering.
- Caltech101-7. The dataset is composed of 1474 images with 6 kinds of features (Gabor, WM, LBP, SIFT, GIST and CENT), belonging to 7 classes.
- Caltech101-20. The dataset contains 2386 images from 20 classes, and the features are the same as those in Caltech101-7.

Competitors: The proposed SMC_L1 and SMC_L2 are compared with five state-of-the-art multiview clustering methods, including Co-regularized spectral clustering (Co-reg) [29], Robust Multiview Spectral Clustering (RMSC) [32], Multi-Modal Spectral Clustering (MMSC) [30], Auto-weighted Multiple Graph Learning (AMGL) [51], and Iterative Views Agreement (IVA) [52]. Since the results of competitors may be influenced by the post-processing, the experiments are repeated for 30 times, and the averaged result is reported. For our methods and MMSC, the multiview graphs are constructed with an efficient method [22], where the neighborhood size is set as 10. For other competitors, the graphs are constructed with the suggested approaches.

Performance: Table 5 and 6 exhibit the averaged ACC and F-score of Co-reg, MMSC and the proposed methods. It can be seen that SMC_L1 achieves the best performance on most occasions. The performance of SMC_L2 is also promising. The results of all the competitors are dependent on the initialization of K-means, while ours are stable because no post-processing is needed. Co-reg fails in most cases because it requires prior knowledge to determine the weights of different views, which is not provided in the datasets. The performance of RMSC is unsatisfactory because it tends to be seriously influenced by the weak views. AMGL reformulates the spectral learning model, and searches the new representation of original data adaptively. But its performance is sensitive to the post-processing. MMSC and IVA obtain close results to SMC_L2, however, they are not so practical as our method because they rely on the extra parameters. For a better interpretation, we compare the performance of SMC_L2 and MMSC on different datasets. The hyperparameter γ of MMSC, which controls the distribution of different weights, is set with different values, as shown in Fig. 10.

In Fig. 10, we note that MMSC enjoys satisfying results at the optimal γ on MSRC-v1 and Caltech101-7, but its performance drops dramatically with the change of γ. The value of γ influences the performance of MMSC. But the optimal values on the four datasets are different, and it's unrealistic to chose a γ that is suitable for different applications. The proposed SMC_L2 performs well under all circumstances because it does not rely on any parameter. It is worthwhile to mention that compared to the slight difference on group detection, SMC_L1 outperforms SMC_L2 a lot on multiview clustering. This is because that the multiview clustering datasets contain more data points with higher dimensions, and the view number is larger. So the data graphs contain more outliers. Since the L1-norm is more robust to outliers, SMC_L1 achieves better performance than SMC_L2. To sum up, the L1-version is suitable to handle the data with large noise, and the L2-version is appropriate to process the data with less outliers.

Convergence Study

Here we prove the convergence of the proposed optimization algorithms. For both the SMC_L1 and SMC_L2,

the optimal S, F and w are searched in each iteration. So the objective value decreases monotonically and finally converges to a local optima. The objective values of SMC_L1 and SMC_L2 at each iteration are plotted in Fig. 11. As observed from the figure, the optimization algorithms converge quickly in less than ten iterations.

In addition, the ACC curves are also shown in Fig. 12. The clustering accuracies increase during the iterations, which means that the objective values of SMC_L1 and SMC_L2 are consistent to the accuracies. So we come to a conclusion that our objectives are appropriate for multiview clustering.

7 Conclusion and Future Work

This paper proposes a Multiview-based Parameter Free framework (MPF) for group detection. A novel Structural Context descriptor is put forward to profile the structural properties of feature points. Two versions of the Self-weighted Multiview Clustering method are designed to integrate the points' correlations from both the orientation and context views. A tightness-based merging strategy is developed to combine the coherent local groups reasonably. Extensive experiments on various kinds of datasets demonstrate the effectiveness of the proposed group detection framework and the multiview clustering method.

In the future work, we want to tackle the detection and tracking problems in crowd scenes, which will improve the achieved performance to a great extent. It's also desirable to design more effective features to perceive the crowd behaviors.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, and the National Natural Science Foundation of China under Grant 61773316.

References

[1] 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.

[2] Y. Yuan, J. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 562–575, 2015.

[3] S. Yi, H. Li, and X. Wang, "Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4354–4368, 2016.

[4] S. Yi, X. Wang, C. Lu, and J. Jia, "L0 regularized stationary time estimation for crowd group analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2219–2226.

[5] 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.

[6] X. Liu, "Multi-view 3d human tracking in crowded scenes," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 3553–3559.

[7] R. Mazzon, F. Poiesi, and A. Cavallaro, "Detection and tracking of groups in crowd," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2013, pp. 202–207.

[8] H. Yu, Y. Zhou, J. P. Simmons, C. Przybyla, Y. Lin, X. Fan, Y. Mi, and S. Wang, "Groupwise tracking of crowded similar-appearance targets from low-continuity image sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 952–960.

[9] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 705–711.

[10] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589–597.

[11] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using group tracking and local features," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 218–224.

[12] 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 Transactions on Image Processing*, vol. 25, no. 4, pp. 1674–1687, 2016.

[13] 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.

[14] 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, p. 14431454, 2011.

[15] 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.

[16] 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.

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

[18] 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.

[19] Y. Wu, Y. Ye, and C. Zhao, "Coherent motion detection with collective density clustering," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, 2015, pp. 361–370.

[20] 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.

[21] F. Solera, S. Calderara, and R. Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 38, no. 5, pp. 995–1008, 2016.

[22] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 1969–1976.

[23] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[24] 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.

[25] M. Hu, S. Ali, and M. Shah, "Learning motion patterns in crowded scenes using motion flow field," in *International Conference on Pattern Recognition*, 2008, pp. 1–5.

[26] R. Mehran, B. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *European Conference on Computer Vision*, 2010, pp. 439–452.

[27] M. Chen, Q. Wang, and X. Li, "Anchor-based group detection in crowd scenes," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 1378–1382.

[28] Y. Zhang, L. Qin, R. Ji, S. Zhao, Q. Huang, and J. Luo, "Exploring coherent motion patterns via structured trajectory learning for crowd mood modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 635–648, 2017.

[29] A. Kumar, P. Rai, and H. Daum, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[30] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1977–1984.

[31] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 6, pp. 1438–1446, 2010.

[32] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 2149–2155.
[33] A. Wahid, X. Gao, and P. Andreae, "Multi-objective multi-view clustering ensemble based on evolutionary approach," in *IEEE Congress on Evolutionary Computation*, 2015, pp. 1696–1703.
[34] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *AAAI Conference on Artificial Intelligence*, 2015, pp. 2750–2756.
[35] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *IEEE International Conference on Computer Vision*, 2015, pp. 1582–1590.
[36] M. Li, X. Liu, L. Wang, Y. Dou, J. Yin, and E. Zhu, "Multiple kernel clustering with local kernel alignment maximization," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 1704–1710.
[37] X. Liu, Y. Dou, J. Yin, L. Wang, and E. Zhu, "Multiple kernel k-means clustering with matrix-induced regularization," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 1888–1894.
[38] X. Liu, S. Zhou, Y. Wang, M. Li, Y. Dou, E. Zhu, and J. Yin, "Optimal neighborhood kernel clustering with multiple kernels," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2266–2272.
[39] M. Ballerini, "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study," *Proceedings of the national academy of sciences*, vol. 105, no. 4, pp. 1232–1237, 2008.
[40] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, 2014.
[41] J. Fang, Q. Wang, and Y. Yuan, "Part-based online tracking with geometry constraint and attention selection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 854–864, 2014.
[42] S. Kullback, "On the convergence of discrimination information (corresp.)," *IEEE Transactions on Information Theory*, vol. 14, no. 5, pp. 765–766, 1968.
[43] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Trans. Neural Networks*, vol. 22, no. 11, pp. 1796–1808, 2011.
[44] F. Nie, H. Huang, X. Cai, and C. H. Q. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Advances in Neural Information Processing Systems*, 2010, pp. 1813–1821.
[45] R. E. Tarjan, "Depth-first search and linear graph algorithms," *SIAM Journal Computing*, vol. 1, no. 2, pp. 146–160, 1972.
[46] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 977–986.
[47] J. Winn and N. Jojic, "LOCUS: learning object classes with unsupervised segmentation," in *IEEE International Conference on Computer Vision*, 2005, pp. 756–763.
[48] M. van Breukelen, R. Duin, D. Tax, and J. Hartog, "Handwritten digit recognition by combined classifiers," *Kybernetika*, vol. 34, no. 4, pp. 381–386, 1998.
[49] F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.
[50] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 1881–1887.
[51] ——, "Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 1881–1887.
[52] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan, "Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 2153–2159.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Mulin Chen received the B.E. degree in software engineering and the M.E. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2016 respectively. He is currently pursuing the Ph.D. degree with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His current research intersts include computer vision and machine learning.

Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He is currently a Professor with the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. He has authored over 100 papers in prestigious journals and conferences like the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the International Conference on Machine Learning, the Conference on Neural Information Processing Systems, and the Conference on Knowledge Discovery and Data Mining. His current research interests include machine learning and its applications fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Dr. Nie serves as an Associate Editor or a PC Member for several prestigious journals and conferences in the related fields.

Xuelong Li (M'02-SM'07-F'12) is currently a Full Professor with the School of Computer Science and with the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China.