Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

A Multiview-Based Parameter Free Framework for Group Detection

Xuelong Li, Mulin Chen, Feiping Nie, Qi Wang*
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
xuelong_li@opt.ac.cn, chenmulin001@gmail.com, feipingnie@gmail.com, crabwq@gmail.com

Abstract

Group detection is fundamentally important for analyzing crowd behaviors, and has attracted plenty of attention in artificial intelligence. However, existing works mostly have limitations due to the insufficient utilization of crowd properties and the arbitrary processing of individuals. In this paper, we propose the Multiview-based Parameter Free (MPF) approach to detect groups in crowd scenes. The main contributions made in this study are threefold: (1) a new structural context descriptor is designed to characterize the structural property of individuals in crowd motions; (2) an self-weighted multiview clustering method is proposed to cluster feature points by incorporating their motion 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. Extensive experiments on various real world datasets demonstrate the effectiveness of the proposed approach, and show its superiority against state-of-the-art group detection techniques.

Introduction

When people walk in a crowd space, they tend to sense each other and group together. Within each group, the pedestrians exhibit consistent behaviors and share similar properties. Since groups are the primary components of a crowd and convey plenty of information about crowd phenomenon, the detection of groups has motivated a surge of interest in the context of artificial intelligence. It also involves a wide range of practical applications, such as event recognition (Mehran, Oyama, and Shah 2009; Yuan, Fang, and Wang 2015), crowd counting (Rabaud and Belongie 2006) and semantic scene segmentation (Lin et al. 2016). Though many efforts have been conducted in the past years, the achieved performance is still unsatisfying.

A major difficulty in group detection comes from the inadequate utilization of features. Due to the serve 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 pedestrians' 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 movement of pedestrians accurately all the time. It's necessary to develop a stable feature to perceive the pedestrians' motion patterns.

Another limitation shared by existing works is the arbitrary clustering procedure. After obtaining the feature points' adjacent graph, previous works cluster those points by thresholding the graph (Lin et al. 2016; Zhou, Tang, and Wang 2012; Zhou et al. 2014; Shao, Loy, and Wang 2014; Wu, Ye, and Zhao 2015). This strategies is popular because it doesn't need the prior about the desired cluster number, and it's helpful in some occasions. However, since crowd densities vary across scenes, it's not practical to find a threshold that suitable for all crowds. In addition, these arbitrary approaches neglects 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. However, existing works are limited to decide the group number automatically based on the graph structure.

To alleviate the impact of above issues, a Multiview-based Parameter Free (MPF) framework is proposed in this study. Our main contributions can be summarized as follows.

1. A structural context descriptor is designed to express the structure 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 motion and context correlations of points. Unlike existing approaches, the proposed method doesn't involve any hyperparameter, which makes it applicable for various clustering tasks.
3. A novel framework for group detection is proposed, which has salient properties: (1) the incorporation of features on multiple views; (2) the automatic decision of group number without involving any arbitrary threshold; (3) the capability of dealing with crowd scenes with varying densities.

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

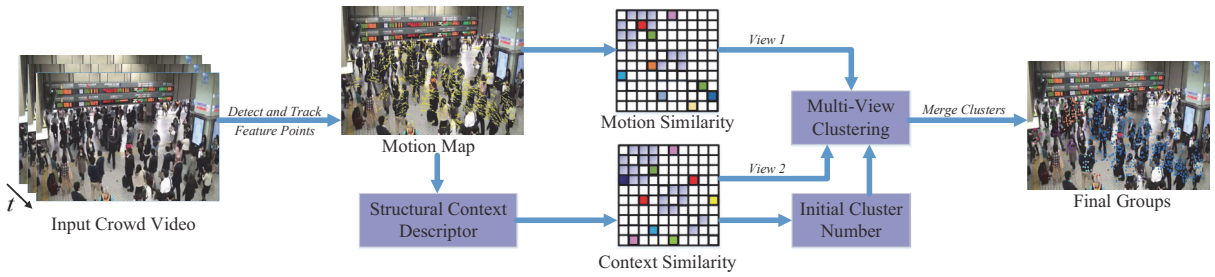


Figure 1: The pipeline of the proposed framework. First, a motion graph is built according to the feature points' motion similarities. Then, a structural context Descriptor is proposed to describe the structures of points. Third, the graphs are integrated by a novel self-weighted multiview clustering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

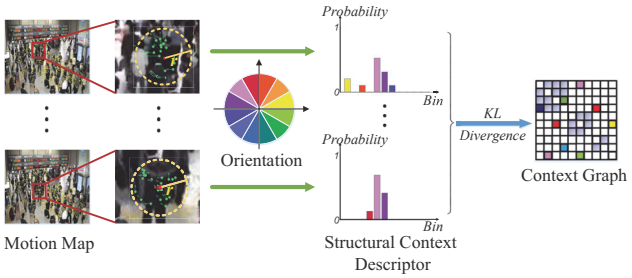


Figure 2: Details about the construction of context graph.

Related Work

In this section, we briefly review the recent studies of group detection and multiview clustering, and discuss their limitations.

Group detection has attracted a wealth of attentions in artificial intelligence. Ali and Shah (2007) divided flow into different dynamics by introducing a Lagrangian Coherent Structure. Lin et al. (2016) detected coherent motions by transfer the optical flow into a thermal energy field. Ge, Collins, and Ruback (2012) introduced a hierarchical clustering strategy to detect groups. Zhou, Tang, and Wang (2012) developed a coherent filtering method to detect groups in crowd scenes. Shao, Loy, and Wang (2014) proposed a coherent prior-based approach to refine the groups obtained by Zhou, Tang, and Wang (2012). Zhou et al. (2014) detected groups by manifold learning. Wu, Ye, and Zhao (2015) designed a multi-stage merging method to detect groups. These methods investigate the motion coherency of pixels or feature points to detect groups. However, the velocities of pixels and points are too microcosmic to reflect the real movements of pedestrians. In addition, all the above approaches involve arbitrary thresholds, so they are not practical for various crowd occasions.

Multiview clustering is also a well-studied topic in machine learning (Guan et al. 2015). In order to capture the relation between different perspectives, many approaches build a graph for each view and integrate them to get a unified graph. Kumar, Rai, and Daum (2011) clustered multiview data by extending the co-regularization method into the spectral clustering scheme. Cai et al. (2011) proposed a multi-modal spectral clustering method to utilize image features from different views. Xia et al. (2014) learned a low-rank transition probability for each view, and input them into a Markov chain to achieve clustering task. There are also some other methods (Xia et al. 2010; Wahid, Gao, and Andreae 2015; Li et al. 2015), which combined the graphs with a conventional weight learning strategy. A shortcoming shared by the above approaches is that they all appealed to a hyper parameter (denoted as γ in following parts for convenience), which restricted their applicabilities to deal with various kinds of data.

Multiview-based Parameter Free Approach

In this section, a Multiview-based Parameter Free (MPF) framework is presented. First, due to the difficulty of extracting pedestrians in crowd scenes, feature points are taken as study objects. And a motion graph is built based on the points' orientations. Then a novel structural context descriptor is developed to represent the structure of each point and a context graph is constructed. After that, a self-weighted multiview clustering method is proposed to cluster points into subgroups by integrating the motion and context graphs. At last, a tightness-based cluster merging strategy is introduced to combine the coherent subgroups into final groups.

Adaptive Motion Description

In order to identify the underlying patterns inside a crowd motion, it's fundamental to compare the pedestrians' motion dynamics. Due to the serious occlusion and noise in crowd scenes, we alternatively take feature points as study objects. To extract feature points, a generalized Kandae-Lucas-Tomasi (gKLT) tracker (Zhou et al. 2014) is employed, which jointly combines the detecting and tracking stages with efficient computation. Then the points' motion similarity is investigated. According to (Li, Chen, and Wang 2016), points always keep connection only with their neighbors and the similarity for points without neighbor relationship should be 0, so it's necessary to find the neighbors of each point.

Some existing works find a point's neigbors by kNN method, which involves a parameter k. However, for crowd motions with varying density, it's not infeasible to find an

optimal k that applies to all situations. Here we propose an adaptive way to find the neighbors of a point. Considering a frame with $\mathbb{N}$ points, the spatial position of a point i $(i = 1, ..., \mathbb{N})$ 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 a spatial distance between point i and j is denoted as

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

Suppose there exists a variable r, and point i and j are considered as neighbors if $d(i,j) < r$. Then the motion graph W_m can be defined as

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

where the max function is used to prevent the similarity from being negative.

It's manifest that the quantity r is crucial for the computation. As a rule of thumb, r is adopted as the $\mathbb{N}$-th smallest element in all pairs of the distance d. Throughout experiments, we find this setting is reasonable. Specifically, when $\mathbb{N}$ is fixed, a higher crowd density corresponds to a smaller r, which complies the fact that the neighbors should resides within a small radius if the crowd motion is with a high density. In addition, existing tracking methods (Wang, Fang, and Yuan 2014; Fang, Wang, and Yuan 2014) are limited to deal with the large variation between consecutive frames. Thus, for videos with a low frame rate, there may be only a small amount of detected feature points although the crowd density is high, then the incorporating of $\mathbb{N}$ will prevent r to be too large.

Structural Context Description

In the above step, a motion graph is built for feature points in crowd scenes, however, the local motion of those points are too microcosmic to reflect the behavior of pedestrians because of motion deviation. So it's necessary to formulate a descriptor to represent point from the macroscopic view. As mentioned before, a point in crowds relate to its neighbors, so its structure can be consequently profiled by its correlations with neighbors. For this purpose, a novel Structural Context (SC) descriptor is developed 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 introduced in the above stage. Then, we divide the orientation space into 12 bins, as shown in Figure 2. 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 the motion of a point may be disrupted due to the limitation of tracking method, in this case its neighbors' motions 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 (Kullback 1968). Thus, the context graph is capable of describing the similarity of points' structures.

Self-weighted Multiview Clustering

Group detection can be considered as the clustering of points. During the above two stages, both the motion and context graphs of feature points are constructed. Here, the graphs are integrated to cluster the points. We first briefly review the Constrained Laplacian Rank (CLR) method (Nie et al. 2016), which conducts clustering task based on a single-view graph. Suppose there are n samples to be classified into c clusters, the objective of CLR is

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

where $S \in \mathbb{R}^{n \times n}$ is a target graph with exactly c connected components, and $||\cdot||_F$ is the Frobenius Norm (Peng et al. 2015). $W \in \mathbb{R}^{n \times n}$ is the input graph, which indicates the similarity of samples. And $L_S \in \mathbb{R}^{n \times n}$ is the laplacian matrix of S, whose to be $n - c$ to guarantee that the c connected components in C correspond 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 (Nie et al. 2011; Cai et al. 2011; Xia et al. 2014), it doesn't need any post-processing.

To integrate the data captured from different aspects, we extend CLR to the multiview clustering scheme. Denote n and n_v as the number of samples and views respectively, and 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 (3), we aim to find a S that approximate each of the views, so the optimization problem is

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

where scalar variable $w^{(v)}$ is the weight of the graph $G^{(v)}$. Without prior knowledge, an intuitive thought is assigning the equal weight to each graph, just as (Kumar, Rai, and Daum 2011). 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, a self-conducted weight learning algorithm is proposed to solve problem (6).

When S is fixed, the problem seems complicated to solve because it can't be directly decoupled into rows. So we transform problem (6) 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 $w = [w^{(1)}, w^{(2)}, ..., w^{(n_v)}]^T \in \mathbb{R}^{n \times 1}$, Eq.(6) naturally becomes a vector form problem

$$\min_{\sum_v w^{(v)}=1, w^{(v)} \geqslant 0} ||A - Bw||_2^2, \tag{7}$$

which is much easier to solve. Spreading the terms in Eq. (7), the problems becomes

$$\min_{\sum_v w^{(v)}=1, w^{(v)} \geqslant 0} \frac{1}{2} w^T B^T B w - w^T B^T A. \tag{8}$$

The above function is a standard quadratic programming (QP) problem, which can be readily solved by an efficient iterative algorithm (Huang, Nie, and Huang 2015) or other existing convex optimization packages.

When $\{w^{(1)}, w^{(2)}, ..., w^{(n_v)}\}$ is fixed, the above problem becomes Eq.(5) and the details of optimization can be referred to (Nie et al. 2016).

Thus, given an initial w, the closed form solution of problem (6) can be computed by updating S and w alternately until convergence. Different from existing multi-view clustering algorithms (Kumar, Rai, and Daum 2011; Cai et al. 2011; Xia et al. 2014), the proposed method is totally self-weighted, and doesn't resort to any hyperparameter. This property is promising because we don't need to tune those additional parameters when handling various crowds.

In our 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 considered to be the number of strongly connected components in the context graph, which can be efficiently computed by the Depth First Search method (Tarjan 1972). Then, graphs on both the motion and context views are utilized to learn the target graph S by solving problem (6). Since S assigns a cluster index to each point, the clustering procedure is accomplished immediately when the optimal solution of problem (6) 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 multi strongly connected ones, leading to an overestimation. Thus, it's necessary to merge obtained subgroups that actually belonging to the same group.

Tightness-based Merging

To combine the coherent subgroups acquired by the previous stage, a tightness-based cluster merging strategy is put forward. Denoting the optimal weight of the motion and context graph as w_m and w_c respectively, which are obtained by the previous clustering procedure. Then a integrated graph is presented as

$$G = w_m G_m + w_c G_c. \tag{9}$$

(a) Subgroups (b) Final Groups

Figure 3: Comparison of groups before and after merging. Scatters with different colors indicate different detected groups, and arrows indicate motion orientations.

The graph G has the ability to approximate both the motion and context graph of points. The reason that we don't use the learned target graph S is that because of 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.

Inspired by the phenomenon that pedestrians in one group tend to keep connection with each other, a tightness measure is introduced to capture the intra-correlations of subgroups. Similar to (Shao, Loy, and Wang 2014), we assume there exists an anchor point within each subgroup, which has the capability to reflect the motion dynamic of the subgroup it belongs to. Then the tightness of a subgroup is considered to be the consistency between the anchor point and others.

Given the weight graph G, we start by determining the anchor points. 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

$$\phi_i = \sum_{j \in sub_\alpha} G(i, j). \tag{10}$$

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,

$$q = \max_{i \in sub_\alpha} (\phi_i + \delta_i), \tag{11}$$

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,

$$T(sub_\alpha) = \phi_q. \tag{12}$$

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

$$T(sub_\alpha + sub_\beta) \geqslant max[T(sub_\alpha), T(sub_\beta)]. \tag{13}$$

By merging consistent subgroups iteratively, the final groups are obtained. Since the sequence of merging will affect the result, we just combine those pairs with the highest value of $T(sub_\alpha + sub_\beta)$ in each iteration.

	CF	CT	CDC	MCC	MPF-m	MPF-c	MPF	r=15	r=25	r=35
ACC	0.7034	0.7523	0.6743	0.6810	0.7263	0.7003	**0.8032**	0.7133	0.7816	0.7056
F-score	0.6714	0.7369	0.6714	0.6657	0.7327	0.6931	**0.7949**	0.7248	0.7793	0.7198

Table 1: Quantitative comparison on group detection. Best results are in bold face.

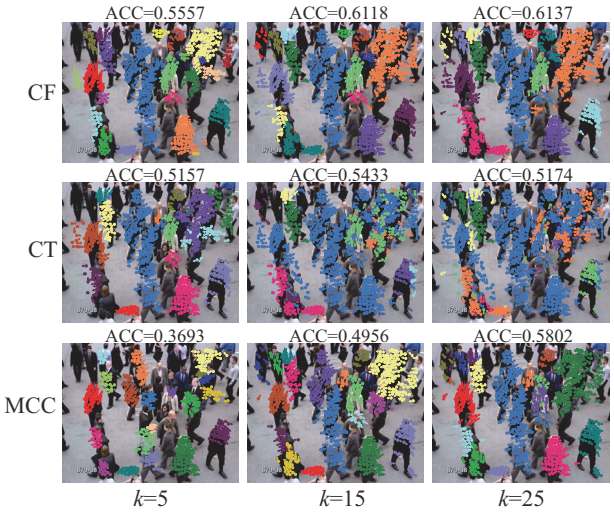


Figure 4: Performance of CF, CT and MCC with varying k. Scatters with different colors indicate different detected groups, and arrows indicate motion orientations.

Through the merging operation, local coherent motions are automatically amalgamated into several global motions, as shown in Figure 3. The proposed merging procedure stops when no subgroups qualified to be combined, so it provides a more principled termination criterion than setting a threshold manually (Zhou, Tang, and Wang 2012; Zhou et al. 2014; Shao, Loy, and Wang 2014; Wu, Ye, and Zhao 2015).

Experiments

In this section, the proposed framework is evaluated from two aspects. Two widely used metrics, the accuracy (ACC) (Nie, Wang, and Huang 2014) and F-score (Xia et al. 2014) are used as measurements to evaluate all the methods quantitatively. Throughout all the experiments, we make the competitors use their respective optimal parameters.

Group Detection Results

In this work, the CUHK Crowd Dataset (Shao, Loy, and Wang 2014) is used to verify the proposed framework's performance on group detection. Four state-of-the-art group detection methods are chosen for comparison.

Dataset: CUHK Crowd Dataset consists of 474 crowd videos, whose frame rates vary 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 conduct group detection on every video and average the obtained ACC and F-score as experimental results.

Competitors: To validate the effectiveness of the proposed framework, four state-of-the-art methods Coherent Filtering (CF) (Zhou, Tang, and Wang 2012), Collective Transition (CT) (Shao, Loy, and Wang 2014), Measuring Crowd Collectiveness (MCC) (Zhou et al. 2014) and Coherent Density Clustering (CDC) (Wu, Ye, and Zhao 2015), are taken for comparison.

Performance: The quantitative comparison of different group detection methods is visualized in Table 1. It can be seen in Table 1 that the proposed MPF method achieves the highest averaged ACC and F-score, which means MPF outperforms other methods. CF and CT detect groups by extracting the invariant neighbors of each point. MCC detect collective motions by thresholding the collectiveness of points. CDC employs a density-based approach to cluster points. All of them utilize only the motion feature, and neglects the structural properties of points. So they tend to be affected by tracking failures and motion deviation. The proposed MPF jointly incorporates the motion and context features with a multiview clustering method, so it has the capability to accurately perceive the movement of pedestrians.

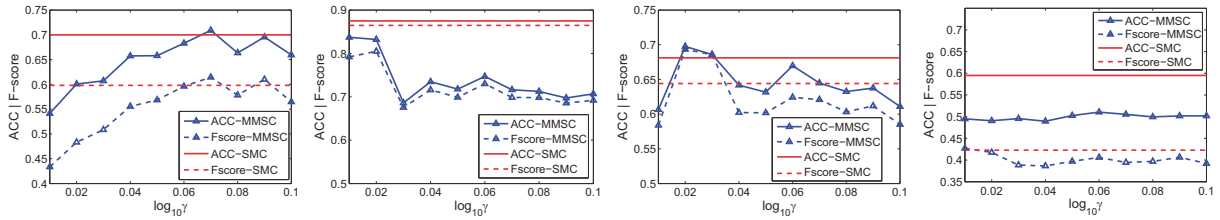
MPF has the salient property that no parameter or threshold is needed. To better illustrate the importance of this property, we compare the result of CF, CT and MCC with varying parameter. The above three methods are chosen because they all involve a kNN processing. Figure 4 shows the clustering result of CF, CT and MCC on a video clip with k is set as 5, 15 and 25. The corresponding ACC is also visualized. Experimental results show that the performance of the three methods is sensitive to the value of k. For crowd motions with various densities, it's not practical to chose an appropriate k that satisfies all occasions. Though CDC doesn't need the kNN procedure, it has multiple additional thresholds to be tuned, so it's not applicable as well. The proposed MPF doesn't have this problem because it's totally parameter free.

We also show the performance of utilizing motion view and context view separately, denoted as MPF-m and MPF-c. As exhibited in Table 1, MPF-m achieves better result than MPF-c. It doesn't mean that context feature fails on all videos. Through experiments, we have found that motion feature performs well when the videos are captured from an overlooking perspective, where pedestrians are small and their velocities can be approximated by those of feature points. However, for videos with serious motion deviation, context feature captures pedestrians' movements better and shows satisfying result. Besides, Table 1 shows MPF is better than MPF-m and MPF-c, so we conclude that the proposed Structural Context descriptor (SC) assists the motion aspect, and the combination of them is reasonable.

In addition, MPF utilizes an adaptive way to decide the relationship threshold r. To demonstrate the validity, we fix

	MSRC-v1		Digits		Caltech101-7		Caltech101-20	
	ACC	F-score	ACC	F-score	ACC	F-score	ACC	F-score
Co-reg	0.7000	0.5905	0.3705	0.3252	0.4252	0.4450	0.4818	0.3921
RMSC	0.6714	0.5937	0.7740	0.6898	0.5855	0.5566	0.5122	**0.4621**
MMSC	**0.7100**	**0.6144**	0.8375	0.7920	**0.6976**	**0.6934**	0.5104	0.4059
SMC	0.7000	0.5981	**0.8750**	**0.8646**	0.6811	0.6411	**0.5951**	0.4227

Table 2: Clustering results on four datasets. Best results are in bold face, and the second-best results are underlined.



(a) MSRC-v1 (b) Digits (c) Caltech101-7 (d) Caltech101-20

Figure 5: Performance comparison of MVSC and MPF on four datasets. We can see that MVSC is sensitive to the value of γ, while MPF sustains good performance on different datasets.

r to be 15, 25 and 35, and show the corresponding performance in Table 1. With r equals to 15 and 35, the performance drops dramatically. This is because a small value of r will make a group divided into parts, while a large r will bring some noise. The performance is relatively well when r is 25, but it's not so good as CPF because a fixed r can't be suitable for crowd videos with various densities and frame rates. Thus, the adaptive decision of parameter r does improve the overall performance of the MPF.

Effectiveness of SMC

In this part, experiments are conducted on various datasets to demonstrate the effectiveness of the proposed Self-weighted Multiview Clustering (SMC) method.

Datasets: The proposed SMC is evaluated on four standard multiview datasets, MSRC-v1 (Winn and Jojic 2005), Digits (van Breukelen et al. 1998), Caltech101-7 and Caltech101-20 (Li, Fergus, and Perona 2007). MSRC-v1 dataset contains 210 images from 7 classes, and the features are extracted from 5 views. Handwritten numerals dataset consists 2000 data points from 10 digit classes, and 5 features are published for clustering. Caltech101-7 and Caltech-20 are composed of images with 6 kinds of features, belonging to 7 and 20 classes respectively.

Competitors: The proposed SMC is compared with Co-regularized spectral clustering (Co-reg) (Kumar, Rai, and Daum 2011), Robust Multiview Spectral Clustering (RMSC) (Xia et al. 2014) and Multi-Modal Spectral Clustering (MMSC) (Cai et al. 2011). 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.

Performance: Table 2 exhibits the averaged ACC and F-score of Co-reg, MMSC and the proposed SMC. It can be seen that SMC achieves the top two performance on all datasets. 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. Except the highest F-score on Caltech101-20, the performance of RMSC is unsatisfactory because it tends to be seriously influenced by those weak views. MMSC obtains competitive results, however, it's not so practical as SMC because it relies on a haperparameter γ. For a better interpretation, we compare the performance of SMC and MMSC on different datasets, and MMSC is set with different values of $\log_{10}\gamma$ (varying from 0.1 to 2 with a 0.2 spacing), as shown in Figure 5.

In Figure 5(a)(c), we note that MMSC enjoys satisfying results at the optimal γ, but its performance drops dramatically with the change of γ. So the value of γ influences the performance of MMSC. However, it can be seen that the optimal values on the four datasets are different. As a result, it's not practical to chose a γ that is suitable for different applications. The proposed SMC performs well under all circumstances because it doesn't rely on any parameter.

Conclusions

In this paper, a context-aware parameter-free (MPF) framework is proposed to detect groups in crowd motions. The Structural Context descriptor is designed to capture the structure property of feature points and the Self-weighted Multiview Clustering method is developed to fuse the information from motion and context views. A tightness-based cluster merging strategy is introduced to discover the global consistency in crowds. Experiments on various datasets show that our method outperforms the state-of-the-art approaches. One of our future works is tackling the detecting and tracking problems, which will tremendously improving the achieved performance. The other is to design more effective features to profile crowds.

References

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

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

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

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

Guan, Z.; Zhang, L.; Peng, J.; and Fan, J. 2015. Multi-view concept learning for data representation. *IEEE Trans. Knowl. Data Eng.* 27(11):3016–3028.

Huang, J.; Nie, F.; and Huang, H. 2015. A new simplex sparse learning model to measure data similarity for clustering. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 3569–3575.

Kullback, S. 1968. On the convergence of discrimination information (corresp.). *IEEE Trans. Information Theory* 14(5):765–766.

Kumar, A.; Rai, P.; and Daum, H. 2011. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems*, 1413–1421.

Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2750–2756.

Li, X.; Chen, M.; and Wang, Q. 2016. Measuring collectiveness via refined topological similarity. *TOMCCAP* 12(2):34.

Li, F.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* 106(1):59–70.

Lin, W.; Mi, Y.; Wang, W.; Wu, J.; Wang, J.; and Mei, T. 2016. A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. *IEEE Trans. Image Processing* 25(4):1674–1687.

Mehran, R.; Oyama, A.; and Shah, M. 2009. Abnormal crowd behavior detection using social force model. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 935–942.

Nie, F.; Zeng, Z.; Tsang, I. W.; Xu, D.; and Zhang, C. 2011. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Trans. Neural Networks* 22(11):1796–1808.

Nie, F.; Wang, X.; Jordan, M. I.; and Huang, H. 2016. The constrained laplacian rank algorithm for graph-based clustering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 1969–1976.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *International Conference on Knowledge Discovery and Data Mining*, 977–986.

Peng, X.; Lu, C.; Yi, Z.; and Tang, H. 2015. Connections between nuclear norm and frobenius norm based representation. *CoRR* abs/1502.07423.

Rabaud, V., and Belongie, S. J. 2006. Counting crowded moving objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, 705–711.

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

Tarjan, R. E. 1972. Depth-first search and linear graph algorithms. *SIAM J. Comput.* 1(2):146–160.

van Breukelen, M.; Duin, R. P. W.; Tax, D. M. J.; and den Hartog, J. E. 1998. Handwritten digit recognition by combined classifiers. *Kybernetika* 34(4):381–386.

Wahid, A.; Gao, X.; and Andreae, P. 2015. Multi-objective multi-view clustering ensemble based on evolutionary approach. In *IEEE Congress on Evolutionary Computation*, 1696–1703.

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

Winn, J. M., and Jojic, N. 2005. LOCUS: learning object classes with unsupervised segmentation. In *10th IEEE International Conference on Computer Vision*, 756–763.

Wu, Y.; Ye, Y.; and Zhao, C. 2015. 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*, 361–370.

Xia, T.; Tao, D.; Mei, T.; and Zhang, Y. 2010. Multiview spectral embedding. *IEEE Trans. Systems, Man, and Cybernetics, Part B* 40(6):1438–1446.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2149–2155.

Yuan, Y.; Fang, J.; and Wang, Q. 2015. Online anomaly detection in crowd scenes via structure analysis. *IEEE Trans. Cybernetics* 45(3):562–575.

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

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