



Salient Points Driven Pedestrian Group Retrieval

Xiao-Han Chen^{1,2,4} and Jian-Huang Lai^{1,3,4}(✉)

¹ School of Data and Computer Science, Sun Yat-sen University,
Guangzhou 510006, People's Republic of China
chenxh45@mail2.sysu.edu.cn, stsljh@mail.sysu.edu.cn

² Faculty of Mathematics and Computer Science, Guangdong Ocean University,
Zhanjiang 524088, People's Republic of China

³ School of Information Science and Technology, XinHua College,
Sun Yat-sen University, Guangzhou 510006, People's Republic of China

⁴ Guangdong Key Laboratory of Information Security Technology,
Sun Yat-sen University, Guangzhou 510006, People's Republic of China

Abstract. Groups are the primary constituent units of crowd and the study on groups can help us better understand the collective phenomena in public area. In this paper, collection of stable individuals with some social relationship in public area, called group, is selected as the research object, and a novel task of pedestrian group retrieval is introduced. Different from the individual person matching, groups often show high aggregation due to their inherent characteristics, individuals in the group are more occluded. Therefore, the performance of individual person based detection and matching will be affected. At the same time, group matching also needs to handle difficulties like variations in the shape and ordering of people within the group. We then design a salient points driven framework for pedestrian group retrieval across non-overlapping cameras. The work focuses on the problems of overall appearance characteristics extraction of a deformable pedestrian collection and matching of groups at varying scales. Experiments on Pedestrian-Groups dataset demonstrate the effectiveness of our proposed framework for Pedestrian Group retrieval.

Keywords: Salient points · Pedestrian group · Group retrieval · Group entire descriptor

1 Introduction

With increasing need for public safety, crowd management in public area with the computer vision technology is essential to improve the management ability of security departments. In the past decade, crowd analysis has been a active field and widely studied and applied in crowd event detection [2, 24], crowd counting [21, 25] and segmentation [1, 15]. Nevertheless, as one of the major constituent units of crowd, the group is also the important research object in public safety

management field. It contains the information that facilitates the understanding about collective phenomena, which raises great interest from the researchers on the study of group, such as group detection [10, 12, 19] or group activity recognition [11, 16, 22]. However, the understanding of group remains challenging, especially the cross-camera group retrieval, which is still few studies.

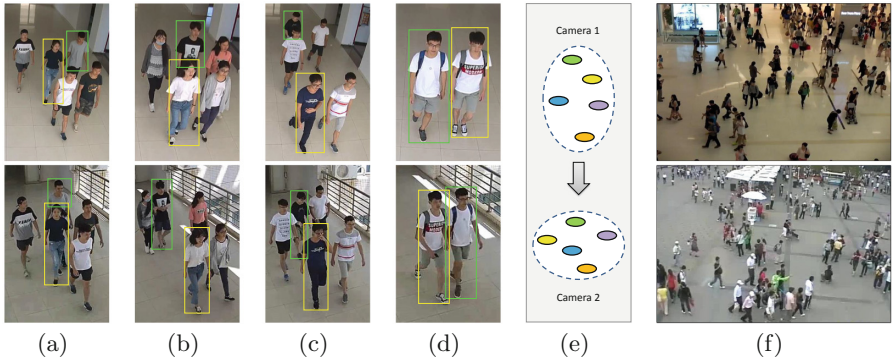


Fig. 1. The challenges in group retrieval. (a)–(d) The two rows of pictures are collected from different cameras, and those in the one column are of the same group obtained from two cameras. Changes could be seen in relative location, distance, and way of occlusion of individuals (the rectangles of the same color in every column indicate the same person in different camera scene). (e) Group may modify their configuration on movement. (f) Pedestrian Groups in real public area (Color figure online)

The major goal of group retrieval is to search and match the same pedestrian collective from different non-overlapping cameras images. As defined in [19], pedestrian group is a cluster of members who tend to move together for a sustained period of time. They usually have some kind of social relationships, e.g. friends or family members. In here, we mainly focus on small groups composed of several pedestrians. The study on group retrieval is of practical value for security management, such as anomaly crowd source tracing or group movement route detection. Therefore, group retrieval has become one of the urgent issues for the security departments. This paper conducts an exploration into this less studied issue and proposes a salient points driven framework for pedestrian group retrieval.

Compared with individual re-identification, group re-identification confronts the following challenges. The first challenge is that the relative position and distance of individuals may change with the movement, and the group structure will also change accordingly. This makes it impossible to measure the similarity through the global matching method. As in Fig. 1(a) and (d), we can see that there are significant changes in the relative locations of the individuals in the group. In addition, the individuals in Fig. 1(b) and (c) are also becoming farther or closer from each other while in motion, and the shape of the group also

changes accordingly. The second challenge concerns the inconsistent scales in the matching of different group due to the difference in individual quantity. The third challenge results from the highly aggregated individuals in the group, this causes them to highly occlude each other. The individual-based matching method turns out to be increasingly hard to conduct in such case. Furthermore, the occlusion varies in form and is largely random. For instance, it can be observed from Fig. 1(b) that the part of occlusion changes for some individuals.

In this study, we first find the key points of the salient feature in the group image by using the key point detection algorithm. Combining with perspective transformation, the sampling of group appearance characteristics is conducted with these salient key points as the center, and a series of non-uniform image blocks are obtained to form a group appearance representation collection. We then extract appearance features vectors such as color, texture and structure for each block in the collection, and carry out the clustering analysis for these features with normalization process, the obtained cluster centers are called group entire descriptor (GED). Finally, for two groups to be matched, we compute the optimal GED matching distance, and use it as a metric for groups identification.

The main contributions of this work are summarized as follows: (1) a local sampling of group characteristics method is proposed to generate group appearance representation collection in dealing with the deformation caused by group movement; (2) a salient points driven framework for group retrieval is designed to solve the issue concerning group matching in different camera; and (3) we also introduce a new cross cameras group dataset *Pedestrian-Groups* for evaluating our proposed methods; it contains various situations such as occlusion each other, group shape and relative position or distance changes.

2 Related Works

Group detection and group activity recognition have attracted a great deal of attention of many researchers [10–12, 16, 22]. A context-aware parameter-free (MPF) framework is proposed to detect groups [12]. In this framework, feature points are detected and clustered by using the motion information. Approach in [10] proposes an instant group motion refining framework based on group motion. Our proposed method also detects feature points at first, but we use the motion clues for denoise processing. In the methods of group activity recognition [11, 16, 22], the representation of group activities often based on individual actions or pairwise interactions. In contrast to these approaches, we use a sub-individual way to construct a group entire descriptor.

About the retrieval problems, single person re-identification has been well studied in surveillance video analysis. In general, these approaches can be categorized into two classes: hand-crafted algorithms [5, 13] and deep learning methods which benefit from the use of large data [20, 23]. One of the crucial tasks of person re-identification is to extract the description of color and textural information of a person image. For this purpose, a person image is usually divided according to a certain principle. In [13], an image is equally divided into six

horizontal stripes, and an effective feature representation called Local Maximal Occurrence (LOMO) is proposed. A method adopts perceptual principles to find two horizontal axes of asymmetry to isolate three main body regions [5]. In [23], for learning different body parts representations, part bounding boxes are obtained by an unsupervised person part generation procedure. In our framework, we fetch a series of image blocks according to the detected feature points.

Perspective normalization is an important step in crowded scene analysis. Abnormal crowd behavior detection [2,3], person detection and tracking [18] and crowd counting [25] have considered the effects of perspective in their proposed methods. To obtain perspective map, the approach in [25] randomly selects several adult pedestrians, and then labels them from head to toe. Similarly, we conduct perspective estimation by selecting and artificially labeling pedestrians in the group image to acquire their heights.

3 Extraction of Group Appearance Representation

When we are identifying groups, we generally first pay our attention to some discriminating local parts that can distinguish it from others instead of directly giving an overall matching. Then, those characteristics will be found from some individuals of one group. The more significant regions that can be matched at another group, the higher possibility that they are the same group. In this process, the entirety of group is decomposed, and some blocks with distinguishable features are selectively re-collected as the basis for final matching. Inspired by such intuition, a series of salient points in group image are first extracted, then local sampling processing is performed to divide a complete group into some units of smaller granularity in order to provide useful clues for the subsequent group matching. In the following, how to develop local appearance sampling on group as driven by those salient points, and acquire the distinguishable group appearance representation collection are going to be discussed.

3.1 Finding the Salient Points Associated to Group

Groups usually have no regular shape, the grid cell division on the images may easily introduce some unrelated information. Therefore, interest point detection is combined with noise reduction by motion information to find out the significance points on the group image. In this paper, Harris detector is employed to extract the salient points [6], while those points usually exist on the locations with critical information. In Fig. 2(a), we provide the detection result of the significance points in the scenario, from which it could be observed that most of the key points are located on the individuals and their locations contain abundant characteristics of the group images. Apart from that, it is also found that some points extracted in such a case may fall into the areas outside the moving groups just like those on the door of background in Fig. 2(a), and those noise points must be eliminated. In view of those, the noise points are removed on using the moving information of optical flow with two continue frames, and then

the target salient points of the group are obtained. As shown in Fig. 2(b), these salient points have a higher correlation with the group.

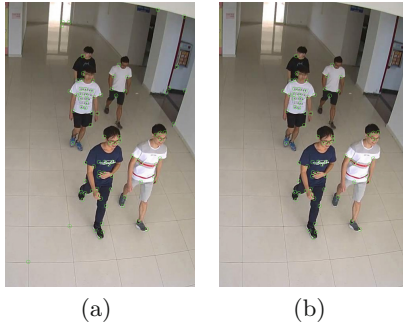


Fig. 2. The detection result of salient points associated to group (marked with green circles). (a) The points are extracted with detection operator; some noise points resided outside the group area. (b) The points in the group area after being treated with a noise removal process. (Color figure online)

3.2 Local Sampling of Group Based Salient Points

After obtaining a series of salient points of group, local appearance sampling is performed as per those points. With points as the centers, some square image blocks with certain width are extracted to represent the characteristics of the group. However, it is noticed that under the effect of perspective distortion, the group pixel size in the scenario changes with the distance from the cameras. This causes the lack of uniformity in image scale. If the sampling process is carried out with the same size in the scenario, the scale of the body parts contained in the image blocks would be inconsistent. To cope with those problems, reasonable perspective transformation should be conducted towards the size of the image blocks. As in [25], to ensure the extracted image blocks could fit the change in individual size as much as possible, the size of sampling window is determined by its location coordinates and scale factor. Assume \mathbf{P} is a point in the set of salient points of a group at location (x_P, y_P) , its sampling width w with the image block is defined as

$$w = (\eta_1 + \eta_2 \cdot y_P) / \rho, \quad (1)$$

where η_1, η_2 are scale factors, and ρ is the height coefficient, value of ρ means the width of image block w is $1/\rho$ of that of a pedestrian image height, ρ is set to 7. The scale factors η_1, η_2 can be calculated by estimating the camera perspective, and each camera is to be estimated for one time. It is assumed that the height of an individual varies linearly from near to far in camera scene. Several images taken by one camera are randomly selected and artificially labeled to acquire the pixel height from head to toe of K individuals, the height of the i th individual

is supposed to be h_i , and the coordinates of its central point is denoted (x_i, y_i) . After that, the scale factors η_1 and η_2 can be calculated by

$$\min_{\eta_1, \eta_2} \sum_{i=0}^K (\eta_1 + \eta_2 \cdot y_i - h_i)^2. \quad (2)$$

Figure 3(a) presents the marking results of one image using bounding boxes. The height of the yellow box is the pixel value of the individual height, whereas the green cross in the middle of each box is the position of central point.

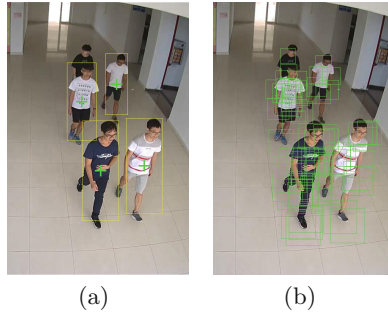


Fig. 3. The local appearance sampling of group. (a) The height of the yellow box is the pixel value of the resulting individual height and green cross in the middle corresponding its location coordinates. (b) The sampling box size is adjusted adaptively according to its location after a perspective transformation (marked in green). (Color figure online)

According to the definition of Eq. (1), image blocks with non-uniform size are next extracted from every point in the salient points set acquired in Sect. 3.1 according to its location. The bounding boxes in Fig. 3(b) illustrate the local appearance sampling results of a group, and the sampling box size is adjusted adaptively according to the specific location. It can be seen from the figure that the size of image blocks is inconsistent due to the varying distances, while the scale of body part in image blocks is close to each other. Although the size of the sampling block is approximate, this kind of perspective transformation can effectively normalize each image block in scale. This is helpful for the subsequent group matching. In Fig. 4, the comparison of the extracted group appearance image blocks is further presented. To facilitate the comparison, we select some image blocks containing person's head; the first row in Fig. 4 is the result extracted in the non-uniform block sampling method as described in this paper, while the second row is the result of the extraction in a uniform width. As shown in Fig. 4, the scale of person's head in the images of first row appears to have a better consistency.

In addition, as the salient points of significance obtained from Sect. 3.1 are usually too densely. When the Euclidean distance between the central points of

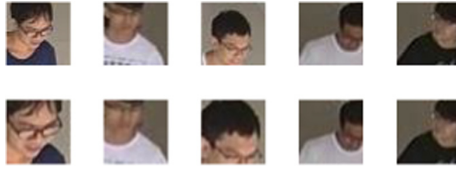


Fig. 4. The comparison of the group local appearance image blocks extraction results with non-uniform and uniform in row 1 and row 2, respectively; the scale of person's head in the images of first row appears to have a better consistency.

the two image blocks is lower than τ , $\tau = 10$, one of them is randomly reserved in order to reduce the redundancy of image blocks and computation as well. Finally, all of these extracted image blocks are resized to the same size (40×40 pixels), and used to construct a group appearance representation collection.

3.3 Extraction of Appearance Feature Vectors

When the image blocks in the group appearance representation collection obtained in Sect. 3.2 are reviewed (Fig. 5), it can be easily observed that those blocks have well reserved the significant appearance characteristics of their group. By extracting the non-uniform appearance blocks of the group, the deformed collective composed of the pedestrian individuals is disassembled into some units of sub-individual granularity; they retain the local characteristics of the group so as to overcome the impact of deformation.

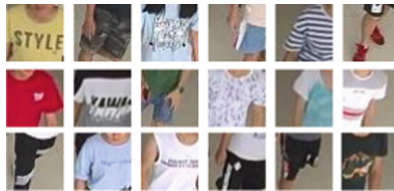


Fig. 5. The group appearance representing image blocks. They can preserve the significant appearance characteristics of group well.

Then, the appearance feature vectors of all the image blocks of each group will be extracted. We use a sliding window of size of 10×10 with an overlapping step of 5 pixels to the 40×40 image block and extract such characteristic information as color, texture. Color is an important characteristic for depicting the appearance of group. The Retinex algorithm is firstly used to enhance each image block [7, 8], so that the image lightness and color under different cameras could be more consistent; we then extract a 512×49 dimensions HSV histogram vector. Next, the scale invariant local ternary pattern (SILTP) descriptor is applied

to obtain the illumination invariant texture feature [14]. The SILTP histograms of two scales (radius: 3 and 5) are extracted separately to get texture feature vectors with the dimension of $81 \times 2 \times 49$. Finally, a 576-dimensional histogram of oriented gradients (HOG) vectors is extracted to gain the structural characteristics of the images [4]. Those results constitute all the feature vectors of the group appearance representation collection.

4 Group Matching

The acquisition of group appearance representation collection enables a complete group to be represented with a set of image units of smaller granularity. However, due to the difference in scenario or group, the number of appearance blocks obtained by each group image is different. Therefore, the appearance feature vectors extracted from the image blocks cannot be directly applied to the matching of groups. We solved the problem in two steps. Firstly, clustering analysis is carried out with the feature vectors of every group by a fixed class number, and the second step is to view the clustering centers resulting from previous step as the group entire descriptor for the final matching.

4.1 Construction of Group Entire Descriptor

Assume \mathbf{R} , \mathbf{S} , \mathbf{T} are the HSV, HOG and SILTP features matrixes of image blocks in all group appearance representation collections, respectively. Each row in those matrixes is a feature vector of one image block. We then denote the vectors \mathbf{m}_{hsv} , \mathbf{m}_{hog} , \mathbf{m}_{siltp} as the mean and \mathbf{s}_{hsv} , \mathbf{s}_{hog} , \mathbf{s}_{siltp} as the standard deviation of each column of matrixes.

For the i th group C_i , there are L image blocks in its group appearance representation collection, and its color, texture and structure features sets can be given by $HSV_i = \{hsv_i^1, \dots, hsv_i^l, \dots, hsv_i^L\}$, $HOG_i = \{hog_i^1, \dots, hog_i^l, \dots, hog_i^L\}$ and $SILTP_i = \{siltp_i^1, \dots, siltp_i^l, \dots, siltp_i^L\}$, where hsv_i^l , hog_i^l and $siltp_i^l$ indicate the HSV, HOG and SILTP feature vectors of the l th image block, respectively. We next conduct the normalization process by Eqs. (3)–(5) on feature vectors hsv_i^l , hog_i^l , $siltp_i^l$, and obtain the results as r_i^l , s_i^l , t_i^l , respectively.

$$\mathbf{r}_i^l = (\mathbf{hsv}_i^l - \mathbf{m}_{hsv}) \oslash \mathbf{s}_{hsv}, \quad (3)$$

$$\mathbf{s}_i^l = (\mathbf{hog}_i^l - \mathbf{m}_{hog}) \oslash \mathbf{s}_{hog}, \quad (4)$$

$$\mathbf{t}_i^l = (\mathbf{siltp}_i^l - \mathbf{m}_{siltp}) \oslash \mathbf{s}_{siltp}, \quad (5)$$

where the operator \oslash is the Hadamard division, which denotes element-wise division. A new feature fusion vector \mathbf{hhs}_i^l is then obtained by vector concatenation, $\mathbf{hhs}_i^l = (\mathbf{r}_i^l, \mathbf{s}_i^l, \mathbf{t}_i^l)$, and new features set of all image blocks of C_i is denoted as $HHS_i = \{\mathbf{hhs}_i^1, \dots, \mathbf{hhs}_i^l, \dots, \mathbf{hhs}_i^L\}$ accordingly. The k-means clustering analysis with N classes is performed on HHS_i , clustering centers of group C_i are then achieved and denoted as $GED_i = \{\mathbf{ged}_i^1, \dots, \mathbf{ged}_i^n, \dots, \mathbf{ged}_i^N\}$, where \mathbf{ged}_i^n is

the n th clustering center vector. Through this processing, the set of N vectors, namely GED , is derived for group C_i .

As clustering to the same number of classes, the difference in the number of image block resulting from the scale difference of groups has been transformed to have the same matching dimensions.

4.2 Optimal Group Entire Descriptors Matching

For a pair of groups to be matched, the i th group C_i and the j th group C_j , it is necessary to compute the final general distance between them. Suppose GED_i and GED_j indicate the GED of C_i and C_j , respectively. The bipartite graph matching method is adopted to find the optimal GED matching distance between C_i and C_j at a minimum cost. The distance between each element of GED_i and GED_j is firstly computed by the Bhattacharyya distance to generate a cost matrix D_{ij} ; then a bipartite graph $G = (V, E)$ (in which $V = GED_i \cup GED_j$) for GED_i and GED_j is constructed, and it is assumed that they are both disjoint sets. In Graph G , every edge $e \in E$ means the cost (distance) between two vertices from GED_i and GED_j , respectively, namely certain value in D_{ij} . Therefore, the optimal matching of bipartite graph G can be determined through solving of Hungarian algorithm [9], and the total cost corresponding to the optimal matching can also be figured out and denoted as $D(GED_i, GED_j)$, that is GED distance between groups C_i and C_j . It means the criteria for measuring the similarity of groups.

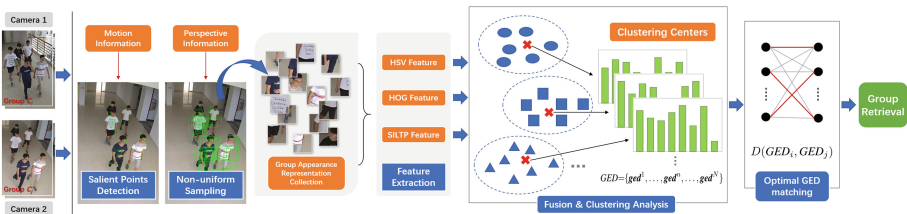


Fig. 6. Overview of the proposed framework for pedestrian group retrieval.

5 Experiments

To evaluate our proposed method, the experiments are conducted on Pedestrian-Groups dataset. In experiments, the clustering parameter N is set to 11. And the framework of our proposed approach is shown in Fig. 6.

5.1 Pedestrian-Groups Dataset

Groups retrieval is one of the new and less studied issue in crowd video analysis field. In order to verify the method proposed in this paper, a new dataset, the

Pedestrian-Groups dataset is constructed. In this dataset, there are 120 images of 30 groups, and two parts of data are collected from each group under two non-overlapping cameras. For the purpose of obtaining the motion information, there are two continuous image frames of one pedestrian group under each camera. In each group, the number of pedestrian individuals ranged from 2 to 5. The dataset is inclusive of such situations as the changing shape of groups, the inter-occlusion and the change in relative location or distance between different individuals.

5.2 Experimental Results

The method proposed in this paper is evaluated on Pedestrian-Groups dataset. The results are shown by the Cumulative Matching Characteristics (CMC) curve. The CMC curve can be associated with the ratio of the probe group found in the top r matches in the group image gallery. In Fig. 7, we can see that the proposed method can effectively identify group across camera, and the matching correctness of rank 1 identification rate is 70.0%. Furthermore, we use single person re-identification based match method as a comparison baseline, which first extracts individuals in group by pedestrian detection method [17] and leverages the person re-identification approach [13] to compute the optimal match cost as a group pair distance. As shown in Fig. 7, person re-identification based match method has lower results than ours for the reason of the highly aggregation feature of group.

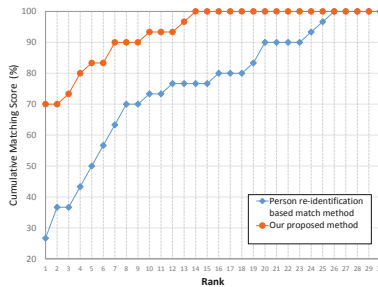


Fig. 7. The CMC curves on the Pedestrian-Groups dataset by comparing the proposed method to the person re-identification based match method.

The effect of perspective transformation-based sampling method proposed in this paper is analyzed by comparing the experimental results from the sampling window of non-uniform and uniform in the group appearance image blocks extraction. As shown in Fig. 8(a), the rank 1 identification rate of non-uniform block extraction method is 70.0% and of the uniform method is 43.3%, it increases by 26.7%, which indicates the perspective transformation processing is conducive to the improvement of the identification effect. For the number of clusters, we conduct the experiments on varying the value of N . Figure 8(b)

shows how the number of clusters affects the rank 1 identification rate. It shows that the performance obtains higher precision when N is more than 8. We fix $N = 11$ in our other experiments.

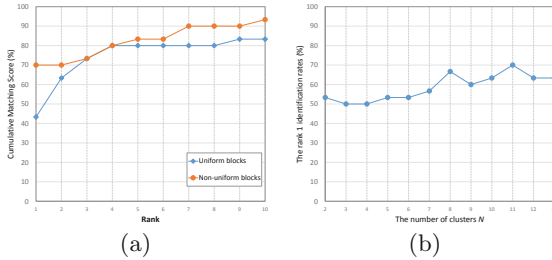


Fig. 8. (a) Results of the uniform and the non-uniform sampling window. (b) Comparing the results of rank 1 identification rates at different number of clusters N .

Table 1. Comparison of different feature representation methods (%); concatenation means that all of the features are directly concatenated without normalization

Features	Rank 1	Rank 10	Rank 20
HSV	66.67	100.00	100.00
HOG	20.00	60.00	83.33
SILTP	40.00	80.00	96.67
Concatenation	40.00	83.33	93.33
Normalization	70.00	93.33	100.00

A comparative evaluation of using the different feature representation methods is carried out. Table 1 presents the evaluation results based on features of HSV, HOG, and SILTP. Although the rank 1 identification rate achieves 66.7% by using just HSV features, combining HOG and SILTP features can increase to 70%, it shows that the combined feature can improve the performance of group identification in rank 1. At last, the results also indicate that treating the direct concatenating of feature without conducting the normalization process has failed to generate a significant effect on matching.

6 Conclusion

In this paper, local sampling method is adopted to decompose the deformative group and a group retrieval framework driven by salient points is proposed. Experimental evaluations are conducted on our Pedestrian-Groups dataset. The experimental results suggest our method can effectively identify the group across cameras. As a preliminary exploration in group matching, we will extend the

research on terms of group characteristic representation and modeling of group matching. Meanwhile, group detection or segmentation may be combined to apply the group retrieval to some more complicated practical scenarios in our future work.

Acknowledgments. This work was supported by National Key Research and Development Program of China (2016YFB1001003), the NSFC (61573387).

References

1. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–6. IEEE (2007)
2. de Almeida, I.R., Cassol, V.J., Badler, N.I., Musse, S.R., Jung, C.R.: Detection of global and local motion changes in human crowds. *IEEE Trans. Circ. Syst. Video Technol.* **27**(3), 603–612 (2017)
3. Chen, X.H., Lai, J.H.: Detecting abnormal crowd behaviors based on the Div-Curl characteristics of flow fields. *Pattern Recogn.* **88**, 342–355 (2019)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
5. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367. IEEE (2010)
6. Harris, C., Stephens, M.: A combined corner and edge detector. In: *Alvey Vision Conference*, vol. 15, pp. 10–5244. Citeseer (1988)
7. Jobson, D.J., Rahman, Z.U., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **6**(7), 965–976 (1997)
8. Jobson, D.J., Rahman, Z.U., Woodell, G.A.: Properties and performance of a center/surround retinex. *IEEE Trans. Image Process.* **6**(3), 451–462 (1997)
9. Kuhn, H.W.: The Hungarian method for the assignment problem. *Naval Res. Logist. Q.* **2**(1–2), 83–97 (1955)
10. Li, N., Zhang, Y., Luo, W., Guo, N.: Instant coherent group motion filtering by group motion representations. *Neurocomputing* **266**, 304–314 (2017)
11. Li, X., Choo Chuah, M.: SBGAR: semantics based group activity recognition. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2876–2885 (2017)
12. Li, X., Chen, M., Nie, F., Wang, Q.: A multiview-based parameter free framework for group detection. In: *AAAI*, pp. 4147–4153 (2017)
13. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206 (2015)
14. Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., Li, S.Z.: Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In: *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1301–1306. IEEE (2010)

15. Lin, W., Mi, Y., Wang, W., Wu, J., Wang, J., Mei, T.: A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. *IEEE Trans. Image Process.* **25**(4), 1674–1687 (2016)
16. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: stagNet: an attentive semantic RNN for group activity recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018*. LNCS, vol. 11214, pp. 104–120. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_7
17. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. *arXiv* (2018)
18. Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.Y.: Density-aware person detection and tracking in crowds. In: *2011 IEEE International Conference on Computer Vision (ICCV)*, pp. 2423–2430. IEEE (2011)
19. Shao, J., Loy, C.C., Wang, X.: Learning scene-independent group descriptors for crowd understanding. *IEEE Trans. Circ. Syst. Video Technol.* **27**(6), 1290–1303 (2017)
20. Shen, Y., Li, H., Xiao, T., Yi, S., Chen, D., Wang, X.: Deep group-shuffling random walk for person re-identification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2265–2274 (2018)
21. Shi, Z., et al.: Crowd counting with deep negative correlation learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5382–5390 (2018)
22. Shu, T., Todorovic, S., Zhu, S.C.: Cern: confidence-energy recurrent network for group activity recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2 (2017)
23. Yao, H., Zhang, S., Hong, R., Zhang, Y., Xu, C., Tian, Q.: Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.* **28**, 2860–2871 (2019)
24. Yuan, Y., Feng, Y., Lu, X.: Structured dictionary learning for abnormal event detection in crowded scenes. *Pattern Recogn.* **73**, 99–110 (2018)
25. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 833–841 (2015)