IEICE TRANS. INF. & SYST., VOL.E100–D, NO.7 JULY 2017

LETTER

Small Group Detection in Crowds using Interaction Information*

Kai TAN[†a)], ***Nonmember***, **Linfeng XU**[†], ***Member***, **Yinan LIU**[†], ***and*** **Bing LUO**[†], ***Nonmembers***

SUMMARY Small group detection is still a challenging problem in crowds. Traditional methods use the trajectory information to measure pairwise similarity which is sensitive to the variations of group density and interactive behaviors. In this paper, we propose two types of information by simultaneously incorporating trajectory and interaction information, to detect small groups in crowds. The trajectory information is used to describe the spatial proximity and motion information between trajectories. The interaction information is designed to capture the interactive behaviors from video sequence. To achieve this goal, two classifiers are exploited to discover interpersonal relations. The assumption is that interactive behaviors often occur in group members while there are no interactions between individuals in different groups. The pairwise similarity is enhanced by combining the two types of information. Finally, an efficient clustering approach is used to achieve small group detection. Experiments show that the significant improvement is gained by exploiting the interaction information and the proposed method outperforms the state-of-the-art methods.

key words: *small group detection, interaction, trajectory clustering*

1. Introduction

Detecting small groups in crowds plays an important role in computer vision field, since it provides a fundamental support for high-level semantic analysis and contains a wide range of practical applications, such as crowd scene classification, crowd activity recognition and so on [1]. Small group is first defined as a collection of pedestrians who have interdependent relations to a certain extent [2]. Generally speaking, small group detection in crowds is formulated as a segmentation [3], [4] or a trajectory clustering problem where the persons travelling with each other are grouped together. An instance is illustrated in Fig. 1. The persons with the same color bounding boxes mean that they are clustered into the same group.

One of the key issues for trajectory clustering is how to measure the pairwise similarity between trajectories. In general, most of trajectory-clustering approaches are designed based on trajectory information [1], [5]. In [5], they propose a series of measures called 'Coherent Neighbor Invariance' to describe the local spatiotemporal relations of moving points in coherent motion. Shao *et al* [1] assume that the motion of points in the same group have a finite number of collective transition priors. Ge *et al* [2] firstly propose to discover small groups and propose a hierarchical clustering algorithm.

Fig. 1 Persons bounded with the same color mean that they are walking together and clustered into the same group.

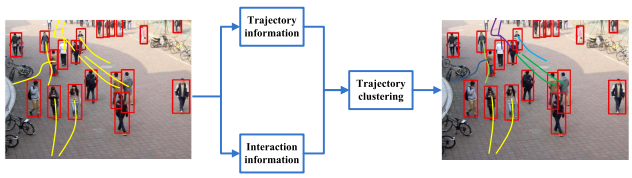


Fig. 2 The main technical pipeline for small group detection. Given a video sequence, trajectory information and interaction information are combined to compute edge weight. Then an agglomerative hierarchical clustering method is used to get trajectory groups. Trajectories with the same color mean that they belong to the same group.

In recent years, the trajectory based methods have made great progress in small group detection. However, due to the sensitivity to the variations of group density and interaction behaviors, the trajectory information is still far from enough to accurately measure the pairwise similarity and new domain information is needed [6]. Fortunately, social scientists have found that the structure of interaction or relationship is highly related to small groups [7]. This inspires us that we should not only exploit the trajectory information but also take into consideration of interaction information conveyed from video content.

In this letter, we incorporate the interaction information into small group detection. Thus, we propose two types of information, i.e., trajectory and interaction information, to cluster trajectories. The trajectory information is used to describe the spatial and motion information. The interaction information consisting of conversation and hand-holding classifiers is exploited to capture the interactive behaviors from video sequence. The goal of the interaction

Manuscript received September 18, 2016.
Manuscript revised March 27, 2017.
Manuscript publicized April 17, 2017.

[†]The authors are with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China.

*This work was supported in part by National Natural Science Foundation of China, under grant number 61601102.

a) E-mail: kaitanuestc@gmail.com

DOI: 10.1587/transinf.2016EDL8192

information is to enhance the pairwise similarity of group members. The pipeline of our method is illustrated in Fig. 2.

This letter is organized as follows. Section 2 introduces the proposed method. Experimental results are provided in Sect. 3. Finally, this letter is concluded in Sect. 4.

2. Small Group Detection

Given a video sequence, N trajectories $L = [l_1, l_2, \ldots, l_N]$ are generated. For the j-th trajectory, it is represented as a set of tuples $l_j = [b_j, p_j, vl_j, t_j]$, where b^i_j is a 4D vector containing top-left and bottom-right coordinates of the j-th bounding box in the i-th frame, as shown in Fig. 2. p^i_j and vl^i_j denote its position and velocity respectively in the i-th frame. t_j is an indicator to record its starting time and ending time. The goal of this paper is to discover small groups $Group = [g_1, g_2, \ldots, g_K]$, where K is the number of groups and $g_k = [l_1, l_2]$ denotes the k-th group consisting of the trajectories l_1 and l_2 if l_1 and l_2 belong to the same group.

2.1 Weights Calculation

2.1.1 Trajectory Information

In the social science literature, Mcphail *et al* [7] conduct a series of experiments to find out which people are travelling with each other. Their experiments indicate that group member satisfies the following properties: people are close enough with each other and not separated by the others;people have the same speed within 0.5 feet per second and motion direction within 3 degrees [2]. According to the aforementioned properties, we know that any two people are considered as group members if they have very small spatial distance, the same speed and direction at the same time. Based on this underlying consideration, a trajectory information weight $s_l(k, j)$ is defined to compute the pairwise similarity between trajectories:

$$\begin{aligned} s_l(k,j) &= exp(\frac{-d_l(k,j)^2}{2\sigma^2}) \\ d_l(k,j) &= d_s(k,j) + \alpha \cdot d_v(k,j) \end{aligned} \tag{1}$$

where $d_s(k,j) = \frac{1}{\Gamma}\sum_i \|\frac{p^i_k - p^i_j}{\sigma_s}\|$ and $d_v(k,j) = \frac{1}{\Gamma}\sum_i \|\frac{v^i_k - v^i_j}{\sigma_v}\|$. $d_l(k,j)$ is a combination of spatial proximity and velocity difference. α is a weight factor that controls the contribution of each term. We set $\alpha = 1$ which means that each term is equally important. Γ denotes the overlap time between any two trajectories. $d_s(k,j)$ and $d_v(k,j)$ compute the average spatial distance and the average velocity difference among Γ scaled by a normalization factor σ_s (or σ_v), respectively.

2.1.2 Interaction Information

The video sequence itself contains rich semantic information. Exploiting these semantic information from video can help detect small groups much better. In particular, the interaction between pedestrians is a very important kind of semantic information. It is highly discriminative and can help observers identify group members in crowds more accurately.

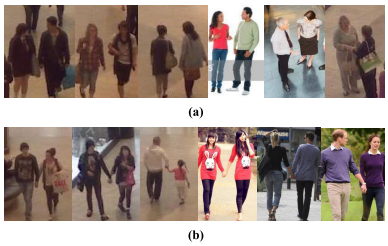


Fig. 3 Some conversation samples (a) and hand-holding samples (b).

To import these interaction information, we define two classifiers that can capture interactive behaviors between pedestrians. The first classifier which we call conversation classifier, recognizes the conversation activity as shown in Fig. 3 (a). The second classifier, we name it hand-holding classifier, assesses the likelihood of containing the behavior that two persons hold the hands. Some samples of these two behaviors are shown in Fig. 3. Both of them are measured to evaluate whether there is a connection between any two persons. In general, persons who converse with each other or hold the hands should be grouped together.

Given two nodes (trajectories) (v_k, v_j), the interaction weight for the edge $e = (v_k, v_j)$ is computed via the following three steps:

1. Calculate class score $c^i_{kj,m}$ for the window W^i_{kj} in the frame I_i by the m-th classifier. Given trajectories (l_k, l_j), we have their proposals in the i-th frame $b^i_k = [x_{1,k}, y_{1,k}, x_{2,k}, y_{2,k}]$ and $b^i_j = [x_{1,j}, y_{1,j}, x_{2,j}, y_{2,j}]$. The window W^i_{kj} is the minimum bounding box containing these proposals b^i_k and b^i_j. Then, two interaction classifiers are performed on the window W^i_{kj} to get the score $c^i_{kj,1}$ and $c^i_{kj,2}$.

2. Compute confidence score $s_1(k,j)$ and $s_2(k,j)$ for the interaction term. As inspired by the method [8], [9], we define that two persons have interaction if they are classified as either conversation or hand-holding class many times along the overlap time Γ. Let $s_m(k,j)$ denote the m-th kind of interaction confidence score. We have

$$s_m(k,j) = \frac{1}{\Gamma}\sum_{i=1}^{\Gamma} \delta(c^i_{kj,m} - \theta) \tag{2}$$

where $\delta(x)$ is a binary function that assigns 1 to x if x is greater than 0, and 0 otherwise. θ is a threshold which is set as 0.5. Thus, the confidence score $s_m(k,j)$ is represented by the times that the window W^i_{kj} is classified as m-th class scaled by the overlap time Γ. This favors that the more times two nodes are classified as m-th class the more likelihood they have this kind of interaction.

3. Calculate the interaction weight $s_h(k,j)$ by a linear combination [10]:

$$s_h(k,j) = \alpha_h \cdot s_1(k,j) + (1 - \alpha_h) \cdot s_2(k,j) \tag{3}$$

where α_h is a weight factor to control the impact of each cue. We set $\alpha_h = 0.5$ which means that each term is equally important. Finally, the edge weight w_{kj} is defined as:

$$w_{kj} = s_l(k, j) \cdot (1 + s_h(k, j)) \quad (4)$$

In this formulation, the trajectory information term can be considered as a base, which means that pedestrians who have spatial proximity and similar velocity are likely to be grouped together; the interaction term can be regarded as an enhancement. It will increase the possibility of being grouped together if pedestrians have high confidence of interaction with each other.

2.2 Clustering Trajectories

The small group detection is formulated as a clustering problem. We use an agglomerative hierarchical clustering method called single-linkage clustering (SL) [11] to get connected sub-graphs. Initially, this approach regards each node as a separate cluster and then gradually merges clusters with the maximum weight until the maximum weight is less than the threshold T. The weight between two clusters G_1 and G_2 is determined by a single element pair $g(G_1, G_2) = \min\{w_{kj} \mid , k \in G_1, j \in G_2\}$.

3. Experiments

3.1 Database

We construct 10 video sequences with different densities ranging from approximately 15 to 40 pedestrians. The detail of video sequences are summarized in Table 1. To obtain the ground truth, the whole body of each person is annotated with a bounding box. To reduce the manual workload, we use Faster R-CNN [12] to generate detections and manually refine the inaccurate ones to get the final ground truth. These bounding boxes then consist of ground truth trajectories by using the tracking method [†]. To get reliable trajectories, we manually delete the wrong trajectories. Consequently, for each person in the video, its trajectory with a unique ID is generated. We follow the procedure and experimental setting as [2]. Briefly speaking, nine subjects watch the videos with IDs covered on all pedestrians. All subjects are asked to identify small groups and give same label to group members with IDs to represent which group they belong to. For ambiguous IDs, they are assigned to the most common labels.

3.2 Evaluation Metrics

Since small group detection is formulated as a trajectory clustering problem, we use Normalized Mutual information (NMI)[††] to evaluate clustering results. NMI is one of the most common measurement in clustering, information retrieval, feature selection and so on. The other measurement we exploit in the experiment is Rand Index (RI)[†††]. The higher RI is, the closer clusters get to ground-truth.

Table 1 The detail of all video sequences in our dataset.

	resolution	frame rate	frames	pedestrian range
Video	640×480	30	150	$15 \sim 40$

[†]https://cvl.gist.ac.kr/project/cmot.html

3.3 Implementation Details

To train the two classifiers, we firstly collect a large number of crowd pictures from the mall dataset[††††] and Web. We get the conversation(/hand-holding) samples from the collected pictures by three steps. Step1, get the bounding box for each pedestrian by following the procedure in Sect. 3.1. Step2, crop the minimum windows containing two bounding boxes as initial samples. Step3, manually select conversation(/hand-holding) samples from the initial samples obtained from step2. Finally, the collected dataset has 6900 images consisting of 2600 conversation samples, 2700 hand-holding samples and 1600 non-interaction samples. To train the conversation (/hand-holding) classifier, we have about 1800 conversation (/hand-holding) samples and about 3000 negative samples consisting of hand-holding and non-interaction samples. We fine-tune the Alex net on our training set with the pre-trained model on the ImageNet dataset.

3.4 Performance

We compare our method with 3 state-of-the-art approaches: CF [5], CT [1] and HC [2]. Since CF and CT are used to detect group in every frame, their performances are obtained by averaging the results in all frames. To perform a fair comparison, we instead use our provided trajectories for CF and CT; the parameters in HC τ_s and τ_v are consistent to ours σ_s and σ_v. τ_s and τ_v are two thresholds which measure spatial distance and velocity difference, respectively. σ_s and σ_v are empirically set as 80 and 2. The threshold T is set as 0.75.

The quantitative evaluation of our method compared with 3 state-of-the-art methods is given in Table 2. From Table 2, we can see that our approach outperforms other

Table 2 Grouping results of all methods. SL indicates that we instead use the clustering algorithm SL in all methods. *ori* denotes the method itself.

		CF [5]	CT [1]	HC [2]	Ours
NMI	ori	0.7565	0.7768	0.9032	0.9317
	SL	0.7830	0.7832	0.8892	0.9317
RI	ori	0.8241	0.8264	0.9542	0.9672
	SL	0.8831	0.8831	0.9455	0.9672

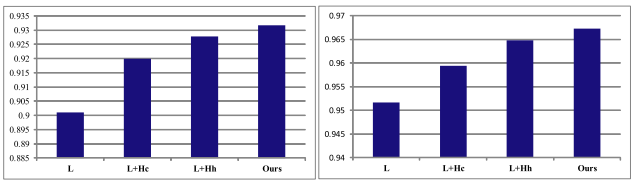
[††]https://arxiv.org/abs/cond-mat/0505245

[†††]http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356

[††††]http://personal.ie.cuhk.edu.hk/~ccloy/downloads_mall_dataset.html

Table 3 The average running time(sec) of different clustering algorithm.

	CF	CT	HC	Ours
Time	0.2240	1.3325	0.0934	0.0646



(a) NMI **(b) RI**

Fig. 4 Analysis of each information in our model. L denotes that only the trajectory information are used. H_c and H_h denotes the conversation and hand-holding information, respectively. Ours is our method that two types of information are used.

Table 4 Classification accuracies.

	Classification accuracies (%)	
Classifier	Training	Testing
ClassC	79.43	76.84
ClassH	81.94	78.49

methods. In particular, our method achieve about 3% improvement of NMI on HC and about 14% improvement on CF and CT. The result is rational, since our method not only uses the trajectory information such as spatial and motion information but also takes into consideration of interaction information. To perform a more fair comparison, we instead use the clustering algorithm (SL) in all methods. The results are shown in Table 2. The performance of CF and CT are still lower than our method, though using SL improves their performance. This indicates the efficiency of the proposed trajectory and interaction information. On the other hand, we replace SL in HC in the experiments which results in performance reduction. This shows that the clustering algorithm (SL) we used is not the best. However, our method still outperforms the other methods. This further verifies the efficiency of the proposed interaction information. In addition, we compare the average running time of different clustering algorithms as shown in Table 3. We run Matlab code on the computer with a 2.8GHz processor. We can see that our method takes less time than others.

We evaluate the contribution of the interaction information proposed in our method. The results are shown in Fig. 4. We can see that two classifiers in interaction term make their contribution to our method respectively. The best performance can be achieved when all information are combined together.

To demonstrate the performance of the two interaction classifiers, we use Cross Validation method to verify the classifiers' performance. The collected data is randomly divided into two nonoverlapping subsets, where 50% images are training samples and the others are testing samples. The randomly splitting and training is performed 50 times. Table 4 shows the classification accuracy. ClassC denotes the conversation classifier and ClassH means the hand-holding classifier. As can be seen, the accuracy of two classifiers attains about 78%. Intuitively, the performance is not very well. Since the samples collected from mall dataset are low resolution and body posture is not obvious enough, it is very difficult to classify the samples collected from mall dataset.

4. Conclusion

In this letter, we propose two types of interaction information to detect small pedestrian groups. To our best knowledge, such idea is first applied in small groups detection. The experiments demonstrate that our method outperforms the state-of-the-art methods and the significant improvement is gained by exploiting the interaction information.

References

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

[2] W. Ge, R.T. Collins, and R.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, May 2012.

[3] H. Li, K.N. Ngan, and Q. Liu, "Faceseg: Automatic face segmentation for real-time video," Trans. Multi., vol.11, no.1, pp.77–88, Jan. 2009.

[4] H. Li and K.N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," Journal of Visual Communication Image Representation, vol.19, no.5, pp.320–333, 2008.

[5] B. Zhou, X. Tang, and X. Wang, "Coherent filtering: Detecting coherent motions from crowd clutters," European Conference on Computer Vision, vol.7573, pp.857–871, 2012.

[6] Q. Wu, H. Li, F. Meng, K.N. Ngan, and S. Zhu, "No reference image quality assessment metric via multi-domain structural information and piecewise regression," Journal of Visual Communication and Image Representation, vol.32, pp.205–216, 2015.

[7] C. McPHAIL and R.T. Wohlstein, "Using film to analyze pedestrian behavior," Sociological Methods and Research, vol.10, no.3, pp.347–375, 1982.

[8] H. Li, F. Meng, and K.N. Ngan, "Co-salient object detection from multiple images," IEEE Transactions on Multimedia, vol.15, no.8, pp.1896–1909, Dec. 2013.

[9] Q. Wu, H. Li, F. Meng, K.N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," IEEE Trans. Circuits Syst. Video Techn., vol.26, no.3, pp.425–440, 2016.

[10] H. Li and K.N. Ngan, "A co-saliency model of image pairs," IEEE Transactions on Image Processing, vol.20, no.12, pp.3365–3375, Dec. 2011.

[11] R. Sibson, "Slink: An optimally efficient algorithm for the single-link cluster method," The Computer Journal, vol.16, no.1, pp.30–34, 1973.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., p.1, 2016.