Collective Interaction Filtering Approach for Detection of Group in Diverse Crowded Scenes

Pei Voon Wong[1*], Norwati Mustapha[2], Lilly Suriani Affendey[2] and Fatimah Khalid[2]
[1] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman,
31900 Kampar, Perak, Malaysia.
[e-mail: wongpv@gmail.com]
[2] Faculty of Computer Science and Information Technology, University Putra Malaysia,
43400 UPM, Serdang, Selangor, Malaysia.
[e-mail: norwati@upm.edu.my; lilly@upm.edu.my; fatimahk@upm.edu.my]
*Corresponding author: Pei Voon Wong

Received January 17, 2018; revised July 1, 2018; accepted September 6, 2018;
published February 28, 2019

Abstract

Crowd behavior analysis research has revealed a central role in helping people to find safety hazards or crime optimistic forecast. Thus, it is significant in the future video surveillance systems. Recently, the growing demand for safety monitoring has changed the awareness of video surveillance studies from analysis of individuals behavior to group behavior. Group detection is the process before crowd behavior analysis, which separates scene of individuals in a crowd into respective groups by understanding their complex relations. Most existing studies on group detection are scene-specific. Crowds with various densities, structures, and occlusion of each other are the challenges for group detection in diverse crowded scenes. Therefore, we propose a group detection approach called Collective Interaction Filtering to discover people motion interaction from trajectories. This approach is able to deduce people interaction with the Expectation-Maximization algorithm. The Collective Interaction Filtering approach accurately identifies groups by clustering trajectories in crowds with various densities, structures and occlusion of each other. It also tackles grouping consistency between frames. Experiments on the CUHK Crowd Dataset demonstrate that approach used in this study achieves better than previous methods which leads to latest results.

Keywords: group detection, clustering, crowded scenes, trajectory, behavior analysis

1. Introduction

There is a rapid installation of closed-circuit television (CCTV) cameras in streets and businesses around the world for safety monitoring. However, manual effort of studying a large number of crowd phenomenon data is difficult. Therefore, automated understanding of crowd behavior using surveillance videos is a pronounced focus in computer vision.

Crowd is difficult to be analyzed, but worthy to be understood as the event is likely to be limited to the results of interaction between members of the same group [1]–[4]. Division of crowded scenes can done into two clusters: structured and unstructured scences according to the movement of the crowd [5]–[7]. In structured crowded scenes, most of the people move consistently and follow common direction, and each dimensional position of the scene devises a key crowd behavior [5]–[7]. Examples of structured crowded scenes include marathon race, processions, and military parade soldiers march. In unstructured crowded scenes, people appear to be random, with different actors moving at different times in diverse directions, and each dimensional position has several crowd behaviors [5]–[7]. Examples of unstructured crowded scenes include exhibitions, people walking on the zebra crossing in the opposite direction, and crowds in airports, railway stations, and shopping centres. The structured and unstructured crowded scenes are the challenges to detect groups of people in crowded scenes. Besides that, various densities and occlusion of each other are also challenges for group detection. Therefore, most existing studies on group detection focuses on scene-specific which resulting in model overfitting, and thus are hardly useful for other scenes.

Motion feature is the input for group detection. While shape-based, color-based, and texture-based are important visual features [7]-[8], motion feature is the fundamental study area in this paper. Trajectory or tracklet is the motion representation for group detection which is calculated based on a moving object tracks. The complete trajectories are hard to gain in crowded scenes due to occlusion. Hence, Zhou et al. [9] suggested tracklet is a short temporal segment of a trajectory. Cluttered background or occlusion affects the terminations of tracklets [9]. Trajectory or tracklet motion feature researches achieved relatively better performance for structured and unstructured crowded scenes [1], [2], [4], [10]. However, object tracking can hardly achieve good accuracy in cluttered scenes and as the number of people increase cause the occlusion [1], [2], [4], [10].

The current group detection approaches which applied in identify groups by clustering trajectories suffer from the following problems. First, the existing grouping techniques are ineffective in division of detection features points into groups in order to form a dynamic group detector. These weaknesses in occlusion caused by the amount of points changes [1], [2], [11], [12]. Second, the group detection approaches cannot support crowds with various structures and densities [1], [5], [7], [13]. Third, most previous studies [1], [10], [12], [14] emphasis on the motion correlation of individuals within a local region and limited to grouping consistency between frames.

In addressng the above limitations, we propose a novel group detection approach called Collective Interaction (CI) Filtering. Below are our main contributions:

- The key person which remains consistent between all frames in each cluster over time-varying dynamics in crowded scenes is explored to handle grouping consistency between frames.

- We introduce distance, occurrence, and speed correlations of each person with the key person as a matrix to deduce the people's interaction in a tracklet clusters with Expectation-Maximization (EM) algorithm to handle the occlusion.
- We propose group refinement threshold based on the results gathered through the inferences on human relationships in order to tackle the crowds with various densities and structures.

This paper is prepared as follows. Current studies of group detection issues are discussed from different perspectives in Section 2. Section 3 explains the proposed group detection approach in detail. Section 4 describes information on CUHK Crowd Dataset [1] for group detection. Section 5 provides a detailed discussion on experimental results for group detection. Section 6 has concluded with a summary and future directions.

2. Related Work

Motion pattern segmentation which is used in crowd analysis is able to analyze the motion patterns and achieve insightful interpretation [15], [16]. It creates clustering difficulty in group detection studies [7]. The process involves detection and extraction of motion features, followed by grouping the features into similar categories through some resemblance measures or probabilities [7].

Cheriyadat et al. [17] apply a distance measurement based on longest common subsequence for feature trajectories. The method involves optical flow tracking in tracking low-level features individually, followed by clustering these trajectories into a smooth motion. Spatio-Temporal Driving Force Model [18] which based on trajectory information of multiple objects tracking is used to model segmentation into offensive and defensive pattern groups. The proposed model simplifies the segmentation algorithm from diverse period instants and detects the main person with the most likely match. Ge et al. [19] uses bottom-up hierarchical clustering of trajectories based on pairwise objects' distance and speed to detect small groups. Cosar et al. [20] apply group tracking algorithm to detect group. The defination of group is where two or more individuals whose distance near each other with the same direction and coherent movement. All the above motion pattern segmentation methods are categorised as similarity model-based clustering. The advantage of similarity model-based clustering is its speed to measure of the similarity between the tracks. However, the cluster parameters need to be fine-tuned for changes of camera viewpoint. The feature point tracking may be affected by noise. Similarity model-based clustering is generally suitable for low-density of structured and unstructured crowded scenes.

The following similarity model-based clustering methods are suitable to be applied in low, medium, and high density of diverse crowded scenes. Zhou et al. [14] propose the Coherent Filtering (CF) algorithm, which is a clustering technique to detect coherent motion from the time series data. It is based on characterization of the interaction between the individual local spatio-temporal called the Coherent Neighbour Invariance. KLT [21] feature point tracker to extract tracklets and then grouped to form trajectories. CF detection cannot be shared across the group dynamic modelling neighbourhood in measuring coherent movement [1]. It is thus sensitive to tracking failure [1]. In addition, CF is first detected based on a coherent motion between successive frames, and then the group is associated through the whole video clip, so its error is accumulated [1]. Trojanová et al. [12] use *k*-nearest neighbor graph to characterise the topological point of trajectories collected in crowd. The data-driven graph segmentation and clustering can be applied to groups with different scales and arbitrary shapes in

overlapping motion patterns [12]. The results of Trojanová et al. also show groupings consistency from frame to frame.

Probability model-based clustering can be implemented for crowd motion pattern segmentation [7] and overcomes the limitation of similarity model-based clustering. Pellegrini et al. [22] employ Conditional Random Field to predict the trajectories and estimate group connections as potential variables in a short prior. The Random Field Topic model which is based on the motions of objects has been used in semantic area analysis in crowded scenes [9], [23]. Zhou et al. [9], [23] model presume that fragments of trajectories or tracklets are detected in crowded scenes without learning semantic region from complete trajectories or from optical flows. Lu et al. [24] suggest an enhanced floor field cellular automation model based on leader-follower rule to study pedestrian group behaviors. The proposed model is valuable for crowd evacuation. All of the above-discussed researches concentrating on scene-specific which resulting in model overfitting. Hence, these methods are difficult to be applied for other scenes.

Methods suggested by Shao et al. [1] and Solera et al. [2] tackle the limitation of identify groups by clustering trajectories in extremely crowded scenes. These methods also can be used to diverse crowded scenes with dynamic viewpoints and scales. Shao et al. propose Collective Transition (CT) prior algorithm for group detection in diverse crowded scenes. The key idea is to discover group members which fit well within the video clip. Moreover, it relies on the temporal relations between local and related speed of a group member. The methods similar to [1] was suggested by Solera et al. [2] in discovering group of people in diverse crowded scenes. Solera et al. propose a new Correlation Clustering algorithm based on people trajectories for social group detection in diverse crowded scenes. The similarity between crowd members is learned by Structural Support Vector Machine based on their physical and social identity [2]. Grounded on the similar principle, Solera et al. [25] propose a tracking groups technique in a multicamera situation. They apply the Correlation Clustering [2] based on long and consistent trajectories to detect group. The clustering technique is built within a planned learning framework [26] to learn a likeness measure appropriate for group articles. However, [2] and [25] methods focus on sparse trajectories rather than dense trajectories. Therefore, [2] and [25] methods based on sparse trajectories are inappropriate for analyzing scenes of a crowd with dense trajectories.

3. Group Detection

Trajectory or tracklet is the motion feature that used for group detection in diverse crowded scenes [1], [27]–[29]. Trajectories are extracted by KLT [21] feature point tracker from each video clip. KLT [21] feature point tracker is suitable to be used to detect and track objects that are in stable shape, and in enviroment with brightness constancy. KLT [21] feature point tracker is often applied in short-term tracking. There is loss of features points due to occlusion, when the feature point tracker algorithm developes over time. Besides that, moving into shadows causes change of some features points appearance over time. The goal of this method is to extract tracklets and then group them to form trajectories. The features point $I(x, y, t)$ in first frame. x and y are the position coordinates features point, and t is the time stamp for the position information. It moves by distance (dx, dy) in the next frame taken after dt time as Eq. (1).

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{1}$$

Group detection input is trajectories and the output is the groups identified by clustering the trajectories. CI is a proposed clustering method to discover people motion interaction from trajectories in order to deduce people interaction with the EM algorithm. It is fit for group detection in low, medium, and high density of diverse crowded scenes. The CI is based on the algorithm used in Shao et al. [1] with few amendments. The first difference is formation of a new equation to determine the key person in every cluster that is generated by the CF [14], in order to handle grouping consistency between frames. The second difference is in the definition of the EM algorithm which is based on weight computed from distance connectivity, *DC*, occurrence connectivity, *OC*, and speed connectivity, *SC*, to deduce the people's interaction in every cluster that is generated by the CF in order to overcome the crowd scene occlusion. Third difference is a new equation for group refinement threshold to tackle the crowds with various structures and densities. As shown in [30], the *DC*, *OC*, and *SC* are not based on key person in every cluster. The detail steps of the CI are explained as follows:

3.1 Create coherent filtering clusters

CF [14] is used to find out a set of initial tracklet clusters, *{C}*. These clusters are not accurately in line with the view of the group, but they can still be used as a basis for seeking the final group tracklet, *{G}*. The first limitation of the CF is sensitive in tracking errors because it detects coherent motions just by considering neighbour interactions, not the whole dynamics of the whole group. The second limitation of CF is that it only requires coherent motion detection over a short duration (4 or 5 successive frames in a norm) [14] which associates with the group. Therefore, its group detection errors will be accumulated. CF works well in crowded scenes but not in situations when the group merges or splits. The third limitation is that CF fails to capture the subtle differences in the movement of points.

3.2 Determine the key person in every cluster

Duplication scheme begins by randomly selecting a cluster, C_i and find the key person, K_i consistency from the start frame, f_1 to the end frame, f_n using the novel Eq. (2).

$$K_i = \sum_{f=1}^{n}\left[G_i.max\sum_{k=1}^{r}\{z_{f,k}\}\right] \tag{2}$$

G_i is the each group that surrounds a set of tracklets, $\sum_{k=1}^{m}\{z_{f,k}\}$ in every cluster detected by a tracker. r is the total tracklets' length for each group generated by CF. CT as proposed by Shao et al. [1] find that the key person in every cluster based on the long duration and low variance of trajectories. Therefore, the key person with long duration has more chance to get together with individuals on his or her pathway and generates a group surrounding him or her. Besides that, for low variance is unusual speed rarely occur within a group such as running and sudden stops. An individual with high speed or velocity variance cannot last as a member of a group. However, CT only emphasis on the motion correlation of individuals within a local region and limited to grouping consistency between frames. The proposed Eq. (2) solves the limitations of CT by considering key person consistency from the start frame, f_1 to the end frame, f_n. For example (Refer to **Fig. 1** for CT found the key person in the detected cluster based on the long duration of trajectories), index 1 to 7 of trajectories belong to the first cluster in frame 1. The key person is index 6 because the highest of total tracklets' length, 41. The index 3 to 11 of trajectories belong to the cluster 1 in frame 2. The key person is index 10 because the highest of total tracklets' length, 45. The further steps of CT are based on the motion correlation

between key person and members in cluster within a local region and then cause grouping inconsistency between frames.

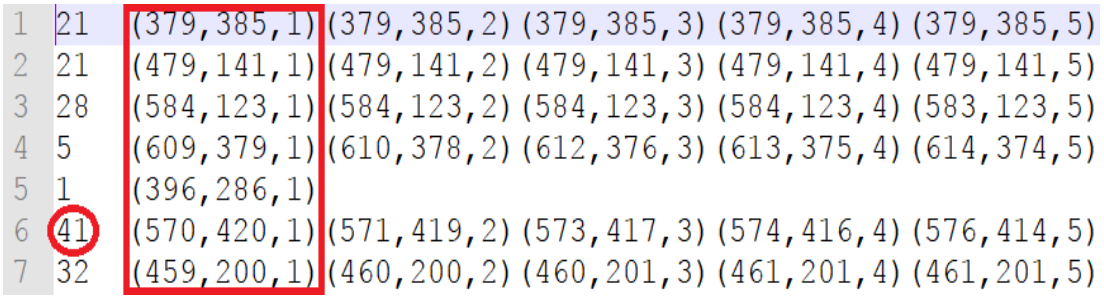


Fig. 1. Found key person based on the long duration of trajectories

3.3 Compute the degree of connectivity of each person with the key person

CT discovers the seeding tracklets based on the high speed correlation with key person in every cluster. Each feature in current frame and search for matching feature within neighbourhood in next frame. Difference in positions is called displacement. Speed is displacement over the frame difference. The threshold for seeding tracklets not suitable for the crowds with various structures and densities. It affects the accuracy of learning CT with EM. EM based on weight computed from *DC*, *OC*, and *SC* are the proposed method to overcome the limitation of the CT.

For each person, p and the key person, K_i in cluster, C_i, $W(p, K_i)$, a weight which is a degree of connectivity between people with the key person, K_i in cluster, C_i from the start frame, f_1 to the end frame, f_n is computed. *DC*, *OC*, and *SC* are the new measurements proposed for approximating the degree of connectivity for each person with key person, K_i in cluster, C_i.

DC measures the degree of distance each person, p with the key person, K_i in cluster, C_i from the start frame, f_1 to the end frame, f_n using the novel Eq. (3).

$$DC_{p,K_i} = \frac{\sum_{f=1}^{n} \sqrt{(D_{pK_i})^2}}{\sum_{f=1}^{n} F_{pK_i}} \tag{3}$$

where F_{pK_i} is the number of frames containing person, p and the key person, K_i. D_{pK_i} is the distance of person, p and the key person, K_i.

OC measures the occurrence of each people, p with the key person, K_i in cluster, C_i from the start frame, f_1 to the end frame, f_n using the novel Eq. (4).

$$OC_{p,K_i} = \frac{\sum_{f=1}^{n} N_{pK_i}}{MAX(N_{p,}N_{K_i,})} \tag{4}$$

where N_{pK_i} is the number of frame containing person, p and the key person, K_i. $N_{p,}$ and $N_{K_i,}$ are the number of frames containing only person, p and the key person, K_i.

SC measures the speed correlation of each person, p with the key person, K_i in every cluster, C_i from the start frame, f_1 to the end frame, f_n using the novel Eq. (5).

$$SC_{p,K_i} = \frac{\sum_{f=1}^{n} S_p \, x \, \sum_{f=1}^{n} S_{K_i}}{\sum_{f=1}^{n} S_{pK_i}} \tag{5}$$

where S_{pK_i} is the same speed of person, p and the key person, K_i. S_p and S_{K_i} are the speed of person, p and the key person, K_i.

DC, *OC*, and *SC* are the three main features to compute the degree of connectivity for each person, p with the key person, K_i in every cluster C_i, as shown by Eq. (6).

$$W(p, K_i) = \frac{2 \, x \, DC_{pK_i} \, x \, OC_{pK_i} \, x \, SC_{pK_i}}{DC_{pK_i} + OC_{pK_i} + SC_{pK_i}} \tag{6}$$

Adjacency matrix is used to deduce the people's interaction in a tracklet clusters with the EM algorithm and group refinement.

3.4 Deduce the people's interaction in a tracklet clusters

Deduce the interaction, R_{pK_i} for each person, p with the key person, K_i in cluster, C_i with the EM based on weights store in an adjacency matrix to handle the occlusion. R_{pK_i} will be used to seek the final group tracklet, $\{G\}$. EM [31], [32] is a frequentative method to obtain maximum probability estimates.

3.5 Group refinement

Fix value of threshold was applied in CT group refinement [1]. The fix threshold cannot tackle the crowds with various structures and densities. Therefore, a new equation for group refinement threshold is proposed to solve the above problems. The group refinement threshold, α is defined as

$$\alpha = \frac{1}{\delta} \, X \, \frac{\sum_{m=1}^{y} W_{m \in S}}{y} \tag{7}$$

All of the $R_{pK_i} < \alpha$, tracklets need to retain to create the final tracklet groups, $\{G\}$. y is the total members of a group in a tracklet clusters, C_i. Substandard tracklets will need to go through the frequentative process repeatedly in order to reflect different groups. S is the weight of connectivity each person, p with the key person, K_i in cluster, C_i from the start frame, f_1 to the end frame, f_n.

The algorithm of group detection (CI) is shown in **Table 1**.

Table 1. CI Algorithm

CI Algorithm

Input:

- Tracklets, T in a video clip.

Output:

- Tracklet groups, *{G}*

Method:

(1) **for each** (f) € T **do** // Tracklet for each frame in a video clip.
initial tracklet clusters *{C}*= CF(f); // Create coherent filtering clusters
end for

(2) **for each** (i) € *{C}* **do** // for each tracklet cluster.
Key person, K € { C_i }; // Determine the key person in every cluster using Eq. (**2**)
end for

(3) **for each** (p, K_i) € *{C}* **do** // for pairwise members in each cluster
$W(p, K_i)$ = WeightFormula($DC(p, K_i)$, $OC(p, K_i)$, $SC(p, K_i)$);
//computing the distance connectivity, $DC(p, K_i)$ using Eq.(**3**)
//computing the occurrence connectivity, $OC(p, K_i)$ using Eq.(**4**)
//computing the speed connectivity, $SC(p, K_i)$ using Eq.(**5**)
//computing the $W(p, K_i)$ weight using Eq.(**6**)
end for

(4) **for each** (W_{pK_i}) € *{C}* **do** // for each tracklet cluster.
R_{pK_i} = EM(W_{pK_i});
// EM - Deduce the people's interaction in a tracklet clusters with weights store in an adjacency matrix
end for

(5) **for each** (p, K_i) € *{C}* **do** // for pairwise members in each cluster
if $R_{pK_i} < \alpha$ **then**
remove (p) € *{C}*; // Group refinement using Eq.(**7**)
end if
end for

return final tracklet groups, *{G}*

4. Crowd Dataset

We used the CUHK Crowd Dataset [1] to compare our results with CF [14] and CI [1] methods. It comprises 474 video clips from 215 different scenes of crowded people. 419 video clips were collected from Getty Image and Pond5. The 419 video clips were existing crowd datasets from Ali and Shah [33], Zhou et al. [10], and Rodriguez et al.'s [34] researches. 55 video clips were captured by Shao et al. [1]. The dataset covers crowd videos with diverse angle scales, densities, and diverse crowded scenes where the pedestrians move to occlude each other or blocked by non-human items. Video resolutions are different, varying from 480 × 360 to 1920 × 1080.

The video clips can be categorized into structured and unstructured crowded scenes. The pedestrians from 215 diverse crowded scenes are divided into 9 categories such as military parade, crowd protest, streets, shopping malls, ports, stations, crossing a zebra crossing, marathon, and escalators. Reference to [35], the low, medium, and high crowd density ranges are defined as shown in **Table 2**. Low density is 1 to 30 people in crowd boundary. 31-100 people in crowd boundary are classified into medium range density. More than 100 people are classified into high range density.

Table 2. Crowd density ranges

Classified Range	**Low**	**Medium**	**High**
Crowd boundary (people)	1-30	31-100	>100

Fig. 2 illustrates the detailed information of dataset. 185 of video clips are categorized into unstructured crowded scenes and 289 video clips are categorized into structured crowded scenes. The breakdown consisted of 122 video clips under the high density, 246 video clips under the medium density, and 106 video clips under the low density. Of the 185 unstructured crowded scenes video clips listed, medium density has the highest video clips (118). The rest are 21 video clips under the high density and 46 video clips under the low density. The crossing a zebra crossing category has the highest video clips (84) from the unstructured crowded scenes. The ports category only has 3 video clips from the unstructured crowded scenes. Of the 289 structured crowded scenes video clips listed, medium density has the highest video clips (128). The rest is 101 video clips under the high density and 60 video clips under the low density. The escalators category has the highest video clips (142) from the structured crowded scenes. The streets, ports, stations categories only have 6 video clips from the structured crowded scenes.

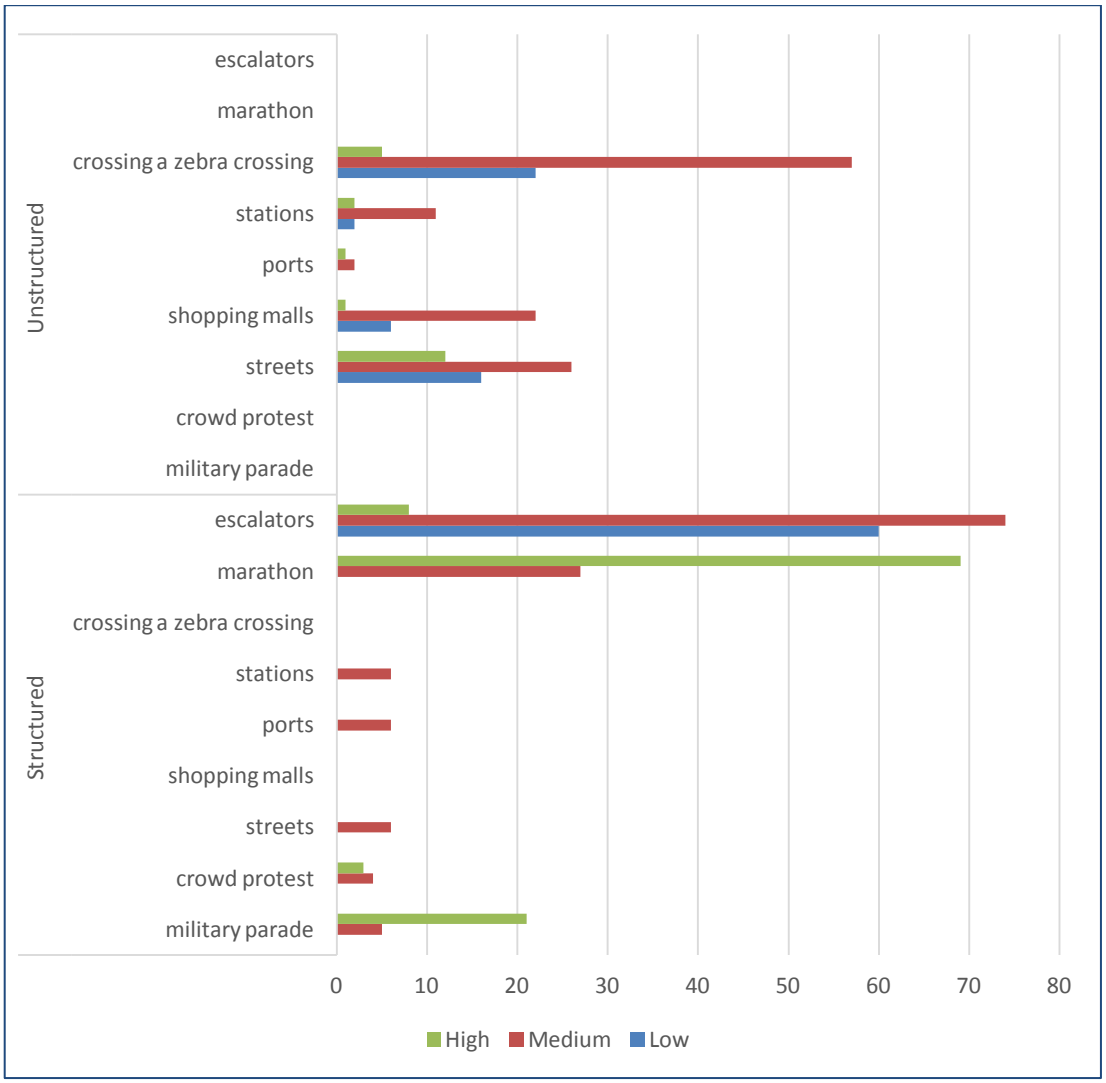


Fig. 2. Detail information of dataset

5. Experimental Results and Analysis

The proposed group detection approach, CI was evaluated by performing extensive experiments on CUHK Crowd Dataset [1] (refer Section 4). The CUHK Crowd Dataset provides 300 video clips tracklets manually descanted into groups. Members of the same group are in the same direction, and the formation of the standard collective motion. Tracklets that do not belong to any group are descanted as outliers. Normalized Mutual Information (NMI) and Rand Index (RI) [1] are applied to obtain the accuracy of the proposed group detection approach. In this research, we only compare our group detection approach with the techniques that are closest to the current approach, respectively explained in [14] and [1]. These methods are known to tackle crowds with various densities, structures, and occlusion in diverse crowded scenes. The proposed group detection approach is implemented with MATLAB. The results are measured on an Intel(R) Pentium(R) CPU G4400 @ 3.30 GHz

processor with 4.0 GB RAM.

The quantitative comparison is presented in **Fig. 3**. The technique used in this paper accomplishes better than the CF [14] and CT [1] by achieving 0.55 for NMI and 0.83 for RI. CI increases 30.9% for NMI and 13.7% for RI compared to the CF. CI increases 14.6% for NMI and 6.4% for RI compared to the CT.

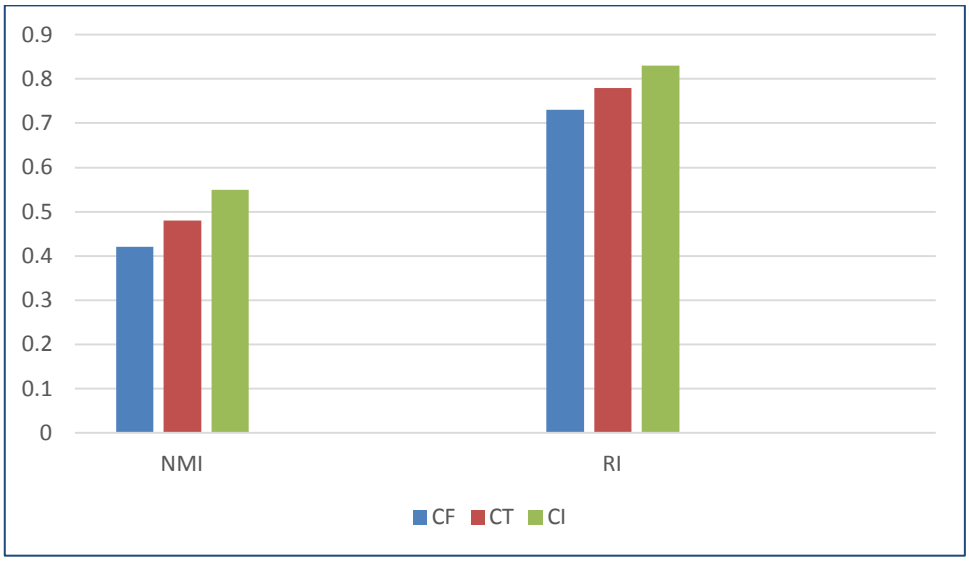


Fig. 3. The quantitative comparison of group detection methods

Fig. 4 shows the examples of ground truth, CF, CT, and CI group detection results for indoor crowded scenes while **Fig. 5** shows the group detection results for outdoor crowded scenes. Groups are distinguished with colour and white circle indicates outliers. The first row in **Fig. 4** shows people on an escalator. CF and CT indicate people with the same direction into different groups because of their limitations to detect global consistency. The second row in Figure **Fig. 4** shows people in a shopping mall. CF and CT indicate the last group member as outlier for the green marked group because they fail to capture the subtle differences in the movement of points. In the third row in **Fig. 4**, CF combines two groups mobilising in diverse directions into one. Subsequent from CF's mistakes made in single frames and the errors are accumulated. Conversely, CT divides a group mobilising in the same pathway into smaller groups. The last row in **Fig. 4** shows people in a station. CF and CT indicate people with the same direction into different groups. The same error also happened in all rows from **Fig. 5**. CF's error is due to it detects coherent motions just by considering neighbour interactions, and not the whole dynamics of the whole group. CT is undependable to find an appropriate threshold for the crowds with various structures and densities.

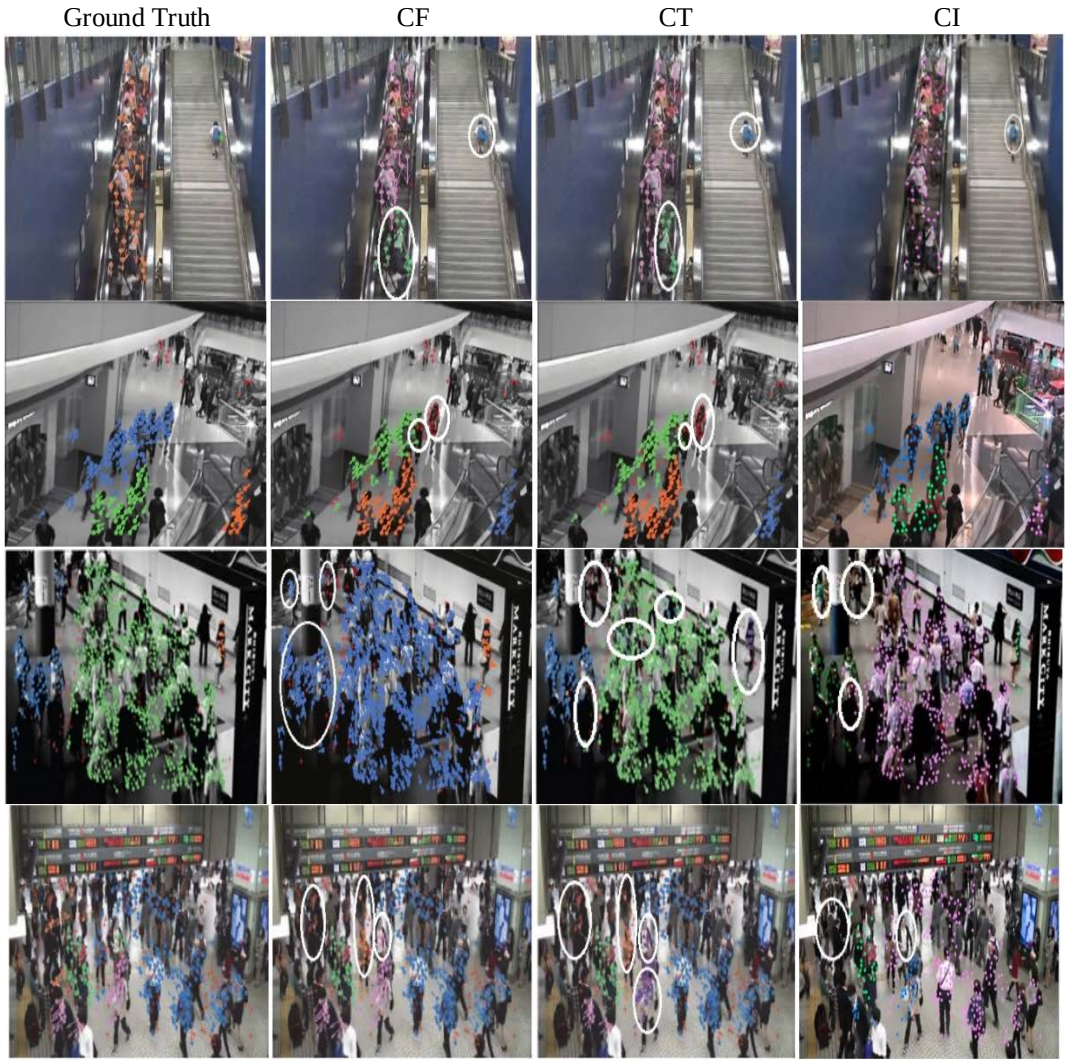
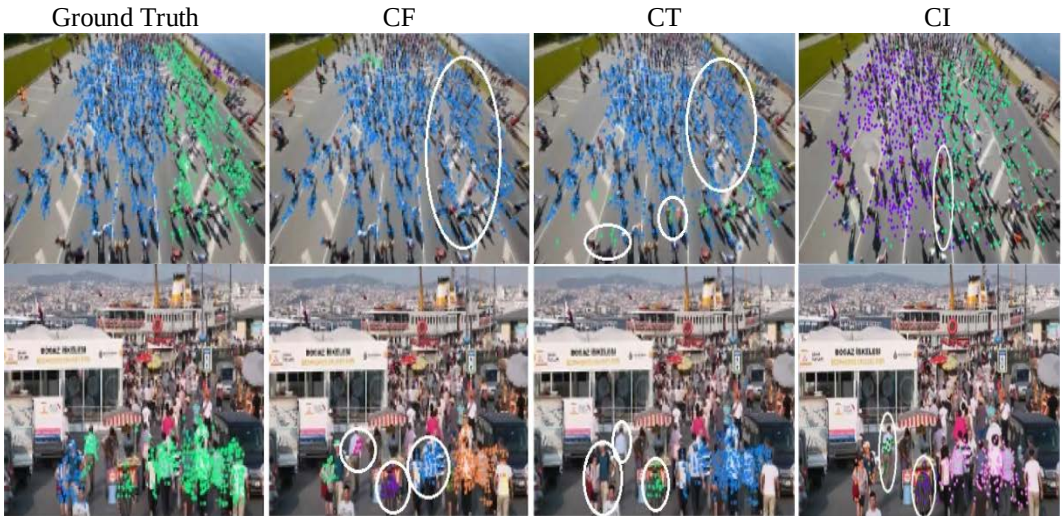


Fig. 4. Examples of ground truth, CF, CT, and CI group detection results for indoor crowded scenes



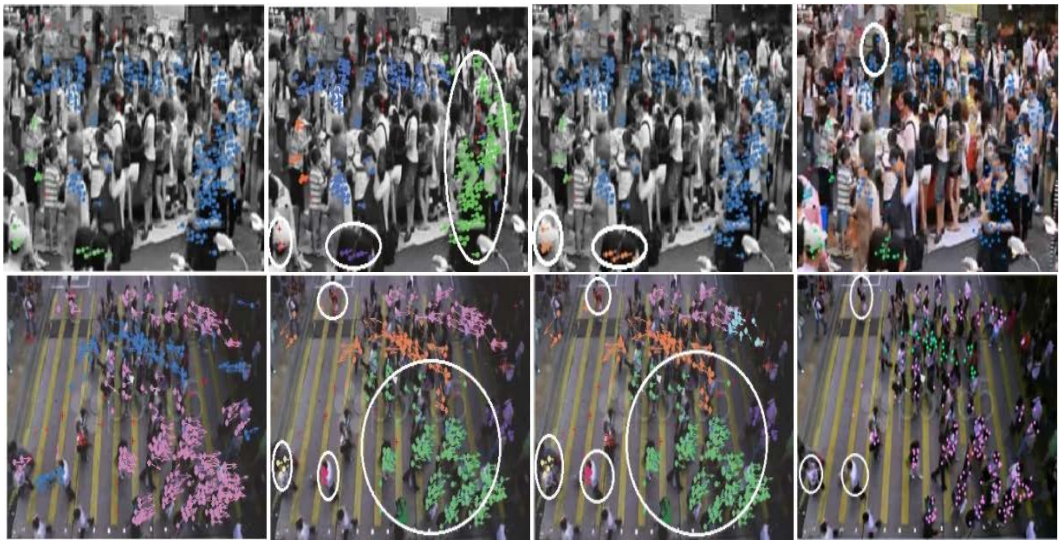


Fig. 5. Examples of ground truth, CF, CT, and CI group detection results for outdoor crowded scenes

The quantitative comparison of CF, CT, and CI methods in low, medium, and high density crowded scenes is shown in **Fig. 6**. The CF, CT, and CI methods are suitable in low, medium, and high density crowded scenes. It is clearly seen that CI method is higher than others in low, medium, and high density crowded scenes. CI method is able to accurately identify groups by clustering trajectories in crowds with various densities, structures and occlusions. It also able to tackles grouping consistency between frames.

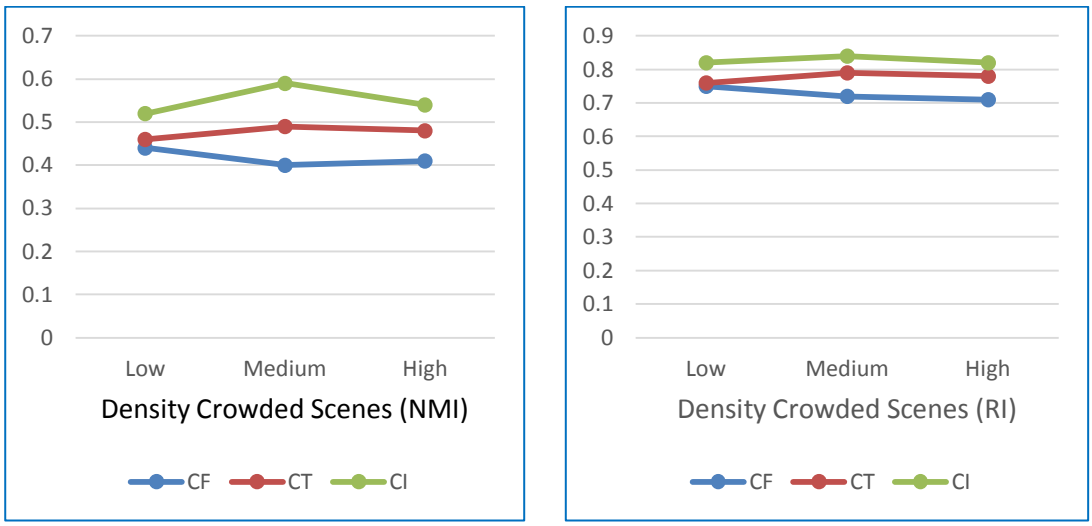


Fig. 6. Quantitative comparison of CF, CT, CI methods in different crowd densities

Fig. 7 and **Fig. 8** illustrate the quantitative comparison of CF, CT, and CI methods in diverse crowded scenes in NMI and RI. The average accuracy of CI method for military parade, streets, shopping malls, ports, stations, crossing a zebra crossing, marathon, and escalators scene categories achieves more than and equal to 0.52 for NMI and 0.80 for RI higher than others. The pedestrians in the escalators scenes category achieves 0.69 for NMI and 0.90 for RI, which is the highest average accuracy for CI method. Mostly people's motions in escalators are

collective and coherent without stationary motions, which improve the detection results. However, the crowd in the protest scene category achieves the worst average accuracy of CI method results, 0.31 for NMI and 0.68 for RI. CF and CT also achieve the worst average accuracy than other categories. The reason is CF, CT, and CI methods are not able to discover trajectories extracted from non-human moving objects such as banners and flags that causes tracking noise.

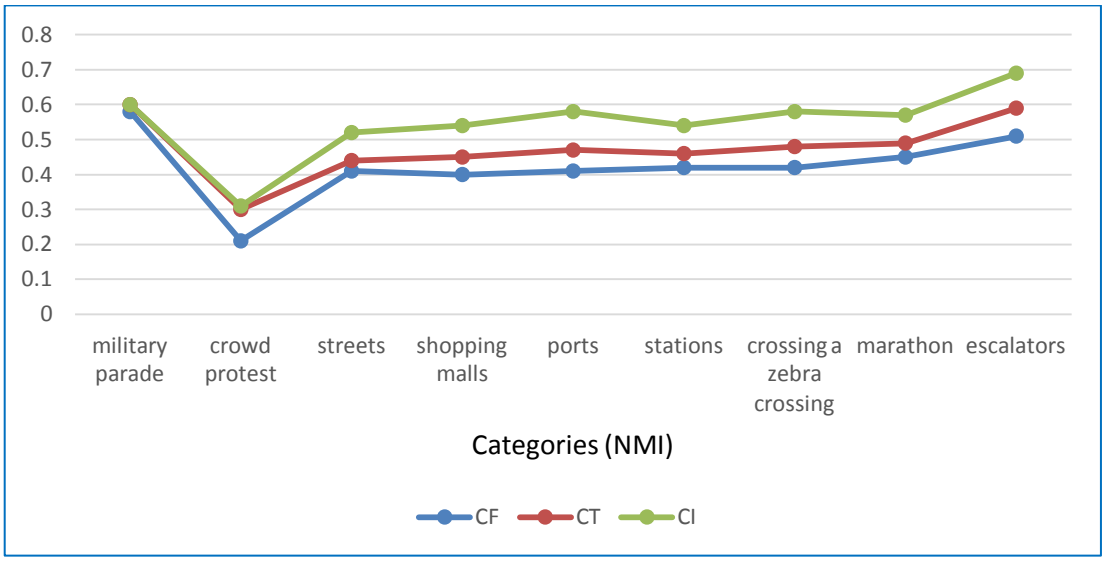


Fig. 7. Quantitative comparison of CF, CT, CI methods in diverse crowded scenes in NMI

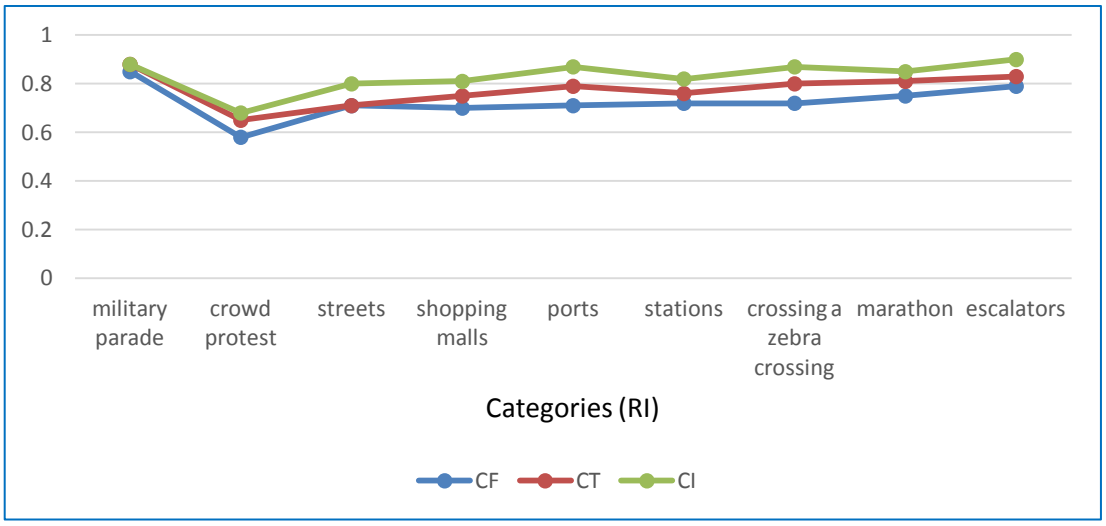


Fig. 8. Quantitative comparison of CF, CT, CI methods in diverse crowded scenes in RI

6. Conclusion

In this study, we recommend a group detection approach called Collective Interaction Filtering. Collective Interaction Filtering is able to determine the key person which remains consistent between all frames in each cluster over time-varying dynamics in crowded scenes as well as to handle grouping consistency between frames. Besides that, the proposed approach form an

inference about human interactions using Expectation-Maximization grounded on distance, occurrence, and speed correlations of each person with the key person to handle the occlusion. Finally, a group enhancement threshold constructed on the results gathered through the inferences on human interactions is applied to tackle the crowds with various structures and densities. The proposed approach shows significant improvements in accuracy of the CUHK Crowd Dataset compared to Coherent Filtering and Collective Transition prior methods.

We would like to explore a technique that able to handle stationary motions and moving objects in our forthcoming work.

Acknowledgements

This work is supported by the Fundamental Research Grant Scheme (FRGS) Malaysia (Project No.: 08-01-17-1918FR).

References

[1] J. Shao, C. C. Loy, and X. Wang, "Learning Scene-Independent Group Descriptors for Crowd Understanding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1290–1303, Jun. 2017. Article (CrossRef Link)

[2] F. Solera, S. Calderara, and R. Cucchiara, "Socially Constrained Structural Learning for Groups Detection in Crowd," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 995–1008, 2016. Article (CrossRef Link)

[3] V. Murino, M. Cristani, S. Shah, and S. Savarese, *Chapter 1 - The Group and Crowd Analysis Interdisciplinary Challenge*, 1st ed. Elsevier Inc., 2017. Article (CrossRef Link)

[4] X. Wang and C. Loy, "Deep Learning for Scene-Independent Crowd Analysis," in *Group and Crowd Behavior for Computer Vision*, 1st ed., Elsevier, pp. 209–252, 2017. Article (CrossRef Link)

[5] J. C. S. J. Junior, S. R. Musse, and C. R. Jung, "Crowd Analysis Using Computer Vision Techniques," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 66–77, Sep. 2010. Article (CrossRef Link)

[6] M. Thida, Y. L. Yong, P. Climent-pérez, H. Eng, and P. Remagnino, "A Literature Review on Video Analytics of Crowded Scenes," in *Intelligent Multimedia Surveillance*, Springer Berlin Heidelberg, pp. 17–36, 2013. Article (CrossRef Link)

[7] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded Scene Analysis: A Survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015. Article (CrossRef Link)

[8] Z. Li and J. Tang, "Weakly Supervised Deep Metric Learning for Community-Contributed Image Retrieval," *IEEE Trans. Multimed.*, vol. 17, no. 11, pp. 1989–1999, Nov. 2015. Article (CrossRef Link)

[9] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," *Cvpr 2011*, pp. 3441–3448, 2011. Article (CrossRef Link)

[10] B. Zhou, X. Tang, H. Zhang, and X. Wang, "Measuring crowd collectiveness," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1586–1599, 2014. Article (CrossRef Link)

[11] R. Liang, Y. Zhu, and H. Wang, "Counting crowd flow based on feature points," *Neurocomputing*, vol. 133, pp. 377–384, 2014. Article (CrossRef Link)

[12] J. Trojanová, K. Křehnáč, and F. Brémond, "Data-Driven Motion Pattern Segmentation in a Crowded Environments," *Florida dental journal*, vol. 9913, no. 4, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, pp. 760–774, 2016. Article (CrossRef Link)

[13] V. J. Kok, M. K. Lim, and C. S. Chan, "Crowd behavior analysis: A review where physics meets biology," *Neurocomputing*, vol. 177, pp. 342–362, Feb. 2016. Article (CrossRef Link)

[14] B. Zhou, X. Tang, and X. Wang, "Coherent filtering: Detecting coherent motions from crowd clutters," *LNCS*, vol. 7573 LNCS, no. PART 2, pp. 857–871, 2012. Article (CrossRef Link)

[15] W. Hu, T. Tan, L. Wang, and S. Maybank, "A Survey on Visual Surveillance of Object Motion and Behaviors," *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.*, vol. 34, no. 3, pp. 334–352, 2004. Article (CrossRef Link)

[16] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–1464, 2006. Article (CrossRef Link)

[17] A. M. Cheriyadat and R. J. Radke, "Detecting dominant motions in dense crowds," *IEEE J. Sel. Top. Signal Process.*, vol. 2, no. 4, pp. 568–581, 2008. Article (CrossRef Link)

[18] R. Li and R. Chellappa, "Group motion segmentation using a Spatio-Temporal Driving Force Model," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2038–2045, 2010. Article (CrossRef Link)

[19] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-Based Analysis of Small Groups in Pedestrian Crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012. Article (CrossRef Link)

[20] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Bremond, "Toward Abnormal Trajectory and Event Detection in Video Surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017. Article (CrossRef Link)

[21] C. Tomasi, "Detection and Tracking of Point Features Technical Report CMU-CS-91-132," *Image Rochester NY*, vol. 91, no. April, pp. 1–22, 1991. Article (CrossRef Link)

[22] S. Pellegrini, A. Ess, and L. Van Gool, "Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings," *LNCS*, vol. 6311 LNCS, no. PART 1, pp. 452–465, 2010. Article (CrossRef Link)

[23] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a Mixture model of Dynamic pedestrian-Agents," in *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2871–2878, 2012. Article (CrossRef Link)

[24] L. Lu, C.-Y. Chan, J. Wang, and W. Wang, "A study of pedestrian group behaviors in crowd evacuation based on an extended floor field cellular automaton model," *Transp. Res. Part C Emerg. Technol.*, vol. 81, pp. 317–329, Aug. 2017. Article (CrossRef Link)

[25] F. Solera, S. Calderara, E. Ristani, C. Tomasi, and R. Cucchiara, "Tracking Social Groups Within and Across Cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 441–453, Mar. 2017. Article (CrossRef Link)

[26] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proc. of Twenty-first international conference on Machine learning - ICML '04*, p. 104, 2014. Article (CrossRef Link)

[27] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4657–4666, 2015. Article (CrossRef Link)

[28] J. Shao, C. C. Loy, K. Kang, and X. Wang, "Slicing Convolutional Neural Network for Crowd Video Understanding," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5620–5628, 2016. Article (CrossRef Link)

[29] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 833–841, 7–12 June, 2015. Article (CrossRef Link)

[30] P. V. Wong, N. Mustapha, L. S. Affendey, and F. Khalid, "A new clustering approach for group detection in scene-independent dense crowds," in *Proc. of 2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, pp. 414–417, 2016. Article (CrossRef Link)

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977. Article (CrossRef Link)

[32] F. Nielsen, "K-MLE: A fast algorithm for learning statistical mixture models," in *Proc. of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 869–872, June 2012. Article (CrossRef Link)

[33] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–6, 2007. Article (CrossRef Link)

[34] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *Proc. of 2011 International Conference on Computer Vision*, pp. 1235–1242, 2011. Article (CrossRef Link)

[35] M. Jiang, J. Huang, X. Wang, J. Tang, and C. Wu, "An Approach for Crowd Density and Crowd Size Estimation," *J. Softw.*, vol. 9, no. 3, pp. 757–762, Mar. 2014. Article (CrossRef Link)

Pei Voon Wong received the B.S. degree in Faculty of Computer Science and Information Technology from University of Malaya in 2004. She received the M.S. in Faculty of Computer Science and Information Technology from University Putra Malaysia in 2012. She is currently a part time Ph.D. candidate student in the Faculty of Computer Science and Information Technology from University Putra Malaysia. She is currently working as a lecturer in the Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman. Her research interests include data mining, social networks, and crowd behavior analysis.

Norwati Mustapha received her BSc degree in Computer Scince from University Putra Malaysia (1991) and MSc degree in Information Systems from University of Leeds (1995). She also obtained her PhD in Artificial Intelligence from University Putra Malaysia (2005). Norwati is an active researcher in the area of data mining, web mining, social networks and intelligent computing. Now, she is working as Associate Professor at University Putra Malaysia.

Lilly Suriani Affendey is an Associate Professor in the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She received her Bachelor of Computer Science degree in 1991 from the Universiti Pertanian Malaysia and in 1994 received her MSc in computing degree from the University of Bradford, UK. In 2007, she received her PhD degree from Universiti Putra Malayia. Her current research interest is in multimedia databases, video content-based retrieval, data science and big data analytics.

Fatimah Khalid is an Associate Professor in the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She hold a Bachelor in Computer Science (University of Technology, Malaysia, UTM), a Masters in Information Technology (National University of Malaysia, UKM) and a Doctorate in Computer Science (UKM). She has been lectured in Computer Science subjects for the past seventeen years, in UPM and currently she attach with Tabuk University for one and the half years. Before entering UPM, She worked as a lecturer with Sal College and as a System Analyst with UKM. Her research interest areas include image processing and computer vision.