Journal of Computers Vol. 31 No. 1, 2020, pp. 34-45
doi:10.3966/199115992020023101003

A Crowd Video Retrieval Framework Using Generic Descriptors

Pei Voon Wong[1*], Norwati Mustapha[2], Lilly Suriani Affendey[2], Fatimah Khalid[2]

[1] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia
wongpv@gmail.com

[2] Faculty of Computer Science and Information Technology, University Putra Malaysia, UPM, Serdang, Selangor, Malaysia
{norwati, lilly, fatimahk}@upm.edu.my

Received 20 April 2018; Revised 2 September 2018; Accepted 5 October 2018

Abstract. In the era of data mining and analytics, retrieval of crowd video with desired motion pattern segmentation plays a significant role in surveillance video management. The retrieval of crowd video with desired motion pattern segmentation poses challenges in finding generic descriptors to describe crowd patterns and similarity matching. This paper presents a novel crowd video retrieval framework using generic descriptors to overcome the above challenges. The anticipated structure comprises of four core components, namely motion feature extraction, group detection, learning generic descriptors, and crowd video retrieval. Results obtained indicate that the proposed framework can improve performance of crowd video retrieval compared with the existing crowd motions on CUHK Crowd Dataset.

Keywords: crowd motion, crowd scenes, crowd video retrieval, generic descriptor, motion pattern segmentation

1 Introduction

Currently, multiple surveillance cameras are installed in public places such as shopping malls, offices, banks, parking lots, airports, train stations, and retail stores to improve safety by real-time monitoring of personnel activities [1-5]. This information is also captured and recorded for future analysis. It is important to efficiently recover the preferred videos from the enormous surveillance video dataset for video management [1-5].

Oftentimes, people use keywords to search videos from websites [1-2, 5-6]. In our study, crowd motion patterns differ significantly in local dynamic features and global spatial structures, causing it hard to precisely label them with some keywords. As the users come from different backgrounds, we may have different descriptions of the same type of motion in this study. Therefore, often it is inconvenient for users to enter keywords to recover crowd videos. In recent years, researchers have shift their research interest to propose generic descriptors to describe crowd patterns. Then, automatically to retrieve the preferred videos based on measuring the likeness between video queries and crowd patterns contained in crowd videos [3-10].

Crowd videos consists of multiple pedestrians walking down the road, or causal, like people participating in a marathon or protest. Crowd videos are difficult to segment the motion pattern because of people move to occlude each other or blocked by non-human items [3-13]. Besides that, crowds with low, medium and high densities in structured and unstructured crowd scenes also challenge to segment the motion pattern [3-13]. These characteristics cause the difficulty to find generic descriptors to describe crowd patterns [3-13], which commonly make the difficulty of measuring the likeness between video queries and crowd patterns enclosed in crowd videos [3-5, 7]. Group descriptors [3] and bilinear curl and divergence (CD) descriptors [4] are the recent works to tackle the above challenges. However, these

* Corresponding Author

descriptors cannot perform well for the motion dynamics caused by pedestrians walk in different directions with extremely diverse behaviors; such as pedestrians in streets or shopping malls.

In this paper, a novel crowd video retrieval (CVR) framework using generic descriptors is proposed. Motion feature extraction, group detection, learning generic descriptors, and crowd video retrieval are the main components for CVR framework. Our framework is able to mine motion pattern from scenes of crowds with various densities, structures, and occlusion with Expectation-Maximization (EM) method. This is followed by group detection using clustering trajectories. Our framework is also capable to find out collectiveness, uniformity, stability, and conflict generic descriptors from the groups frequent motion pattern for retrieve crowd videos. The Euclidean distance and Chi-Square (X^2) distance are used to quantify the likeness matching between the query video and the remaining video clips.

The rest of this paper is prepared as follows. The study of crowd motion representations are discussed in Section 2. In Section 3, we explain a comprehensive explanation of the proposed CVR framework using generic descriptors. Section 4 explains a detailed information on The Chinese University of Hong Kong (CUHK) Crowd Dataset [3] for CVR. Section 5 provides a detailed discussion on experimental results for CVR. Section 6 has concluded with a summary and future direction.

2 Related Work

The motion features are the most useful and representative part in video frames [1, 3-13]. Hence, different crowd motion representations have been selected for crowd behavior analysis and further used in CVR [3-5, 7]. Recently, some researchers proposed generic descriptors, such as curl, divergence, collectiveness, uniformity, stability, and conflict, from the computer vision point of view to describe crowd with various densities, structures, occlusion and different scenes [3-10].

Liu et al. [11] presented a framework to extract motion features based on multiple exemplar agent-based motion models (AMMs) from the crowd trajectories. The proposed Extended Kalman Smoother and KL-divergence were applied in correlation measure. To describe crowd motions, individual feature and holistic feature were learnt. AMMs cannot be applied automatically in any given crowd video without manually estimating the perception transformation.

Stationary-time estimation algorithm was suggested by Yi et al. [12] for stationary crowd analysis. The 3D stationary-time map was formulated as an L_0 optimization issue to discover stationary crowd analysis. Cosar et al. [13] recommended crowd motion representation based on snapped trajectories which essentially used to detect irregular group behaviors in crowd scenes. However, the limitation of the descriptors proposed by Yi et al. and Cosar et al. is only focus on individual attributes and neglect collective attributes about the crowd.

Shao et al. [3] devised a set of scene-independent group descriptors, namely collectiveness, conflict, uniformity and stability from clustering trajectories for crowd behavior analysis. The group descriptors was applied to different scenes group states analysis and crowd video classification. The group descriptors were further applied in CVR. Similar to the Shao et al. work, Zhang et al. [7] discovered intra and inter-group level attributes for CVR. This was followed by deep attribute-embedding graph ranking to measure the video difference between the rankings on group attributes. Both researches are inadequate to handle motion dynamics caused by pedestrians walk in different directions with extremely diverse behaviors.

Wu et al. [8] proposed curl and divergence of motion trajectories (CDT) descriptors to describe crowd motion patterns. They applied the CDT descriptors to detect counterclockwise arch, clockwise arch, lane, fountainhead, and bottleneck crowd behaviors. Based on the same idea, Wu et al. [5] proposed motion structure coding algorithm, which encoded motion vector fields using curl and divergence and produced (CDT^P) for CVR based on hand-drawn sketch. The CDT^P descriptors were used to find out same types of crowd behaviors in [5]. Then, they applied Ranking SVM for similarity matching between sketch queries and crowd videos. However, the CDT [8] and CDT^P [5] descriptors were restricted to structured crowd scenes and were not able to be applied to unstructured crowd scenes [4]. To tackle the limitation of CDT [8] and CDT^P [5] descriptors, Wu et al. [4] proposed bilinear CD descriptors for crowd behavior analysis in structured and unstructured different crowd scenes. The CD descriptors were further applied in CVR. The limitation of CD descriptors is same as Shao et al. [3] and Zhang et al. [7], which is incapable to handle motion dynamics.

However, the current tracking approaches are difficult to capture motion interaction accurately among people with various densities, structures, occlusion and different scenes. These approaches affect the effective of generic description in CVR accurately [3-13]. Table 1 reports the summary of the above review.

Table 1. Summary of crowd motion representations

Author, Year	Crowd Motion Representation	Scenes	Limitation
Liu et al., 2016	AMMs	Structured, Unstructured	Manually estimating the perception transformation of crowd videos
Zhang et al., 2016	Group descriptors	Structured, Unstructured	Inadequate to handle motion dynamics
Cosar et al., 2017	Snapped trajectories	Structured, Unstructured	Focus on individual attributes and neglect collective attributes about the crowd
Shao et al., 2017	Group descriptors	Structured, Unstructured	Inadequate to handle motion dynamics
Wu et al., 2017	CDT descriptors	Structured	Restricted to structured crowd scenes
Wu et al., 2017	CDT^P descriptors	Structured	Restricted to structured crowd scenes
Wu et al., 2017	Bilinear CD descriptors	Structured, Unstructured	Inadequate to handle motion dynamics
Yi et al., 2017	Stationary descriptors	Unstructured	Focus on individual attributes and neglect collective attributes about the crowd

3 The Proposed Crowd Video Retrieval Framework

A CVR framework using generic descriptors has been proposed to find generic descriptors to describe crowd patterns and retrieve the preferred videos based on measuring the likeness between video queries and crowd patterns contained in crowd videos. The block diagram of the proposed CVR framework using generic descriptors for crowds with various densities, structures, occlusion and different scenes as illustrated in Fig. 1.

3.1 Motion Feature Extraction

Motion feature extraction is performed as a first component in proposed CVR framework using generic descriptors. Kanade-Lucas-Tomasi (KLT) [14] feature point tracker is used to detect and track moving objects, and then tracklets are grouped to form trajectories.

3.2 Group Detection

Second component is to detect groups by clustering trajectories. The Coherent Filtering (CF) [15] is used to find out a set of initial tracklet clusters. Then, determine the key person in every cluster and compute the degree of connectivity of each person with the key person [16]. Finally, infer the people's relationship in a tracklet clusters by EM and group refinement.

3.3 Learning Generic Descriptors

The third component is to employ the group motion pattern mining and prediction approach to find out collectiveness, uniformity, stability, and conflict generic descriptors based on interaction among people in groups as shown in Fig. 2. Uniformity generic descriptor is members of intra group in the whole scene who are evenly distributed in space with fixed group size and same direction. The conflict characterizes members in different group are approaching each other in opposite or different directions in different group size and inconsistent distance between members in intra group. The collectiveness descriptor indicates members' intra group move in same direction without fixed group size and distance between

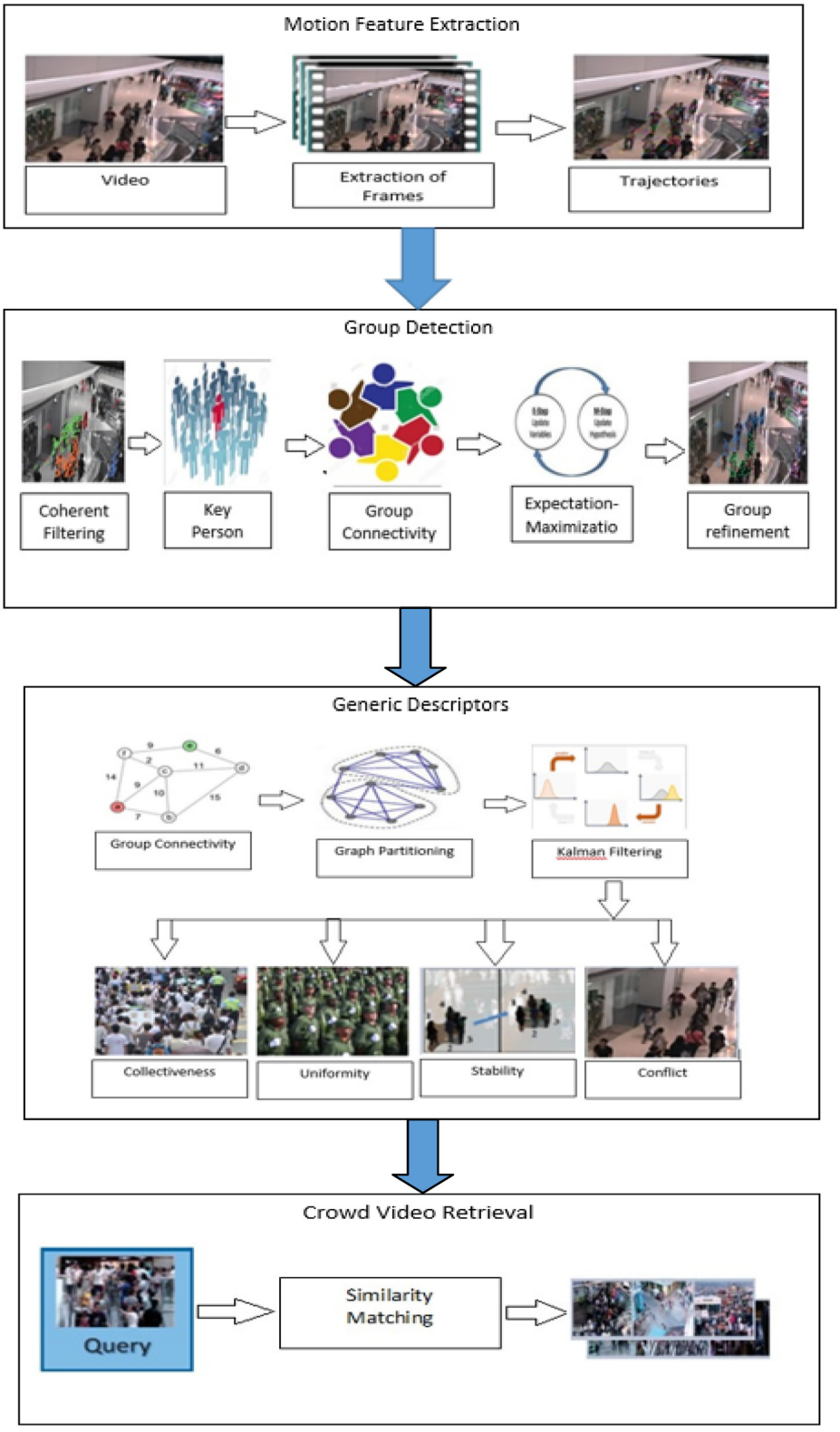

Fig. 1. Block diagram of the proposed crowd video retrieval framework using generic descriptors

members. The stability characterizes members' intra group move in same direction and fixed group size but members are not evenly distributed in space. The first step is to calculate group size, evenly space, and speed direction of connectivity between members in group and create an adjacency matrix. The second step is to create undirected graph to mine the group interaction pattern by using graph partitioning algorithm. Finally, Kalman filtering [17] is used to predict generic descriptors. The generic descriptors are represented by graph-based descriptors.

Collectiveness

Uniformity

Stability

Conflict

Fig. 2. Example for generic descriptors

3.4 Crowd Video Retrieval

The fourth component is employment of the graph-based descriptors in CVR. The two main tasks for CVR are query and retrieval. The Euclidean distance and Chi-Square (X^2) distance are applied to measure similarity matching.

The important attributes for CVR as shown in Table 2. *ID*, and *Name* are the labels assigned to the crowd video and their related name. F_{cm} indicates the covariance matrix feature of generic descriptors for the detected groups.

Table 2. Crowd video retrieval attributes

Label	Meaning
ID	Unique identification of the crowd video
Name	Name of the crowd video
F_{cm}	Covariance matrix feature

The query-by-example is used in query task for the proposed CVR framework using generic descriptors. Users can submit the query video and then relevant results are retrieved by similarity matching of the generic descriptors.

In retrieval task, the Euclidean distance and X^2 distance are applied to measure similarity matching between the query video and the remained video clips for generic descriptors. The relevant results are those videos with the smallest distances of their generic descriptors of the given query.

The Euclidean distance [18] is used to calculate the distance between two generic descriptors from different crowd videos. p and q are the two generic descriptors in two dimensional n video frames, while the distance (d) from p to q, or from q to p is given by the Pythagorean formula as:

$$d(p,q) = d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}, \\ = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}, \tag{1}$$

X^2 distance, a nonlinear matric is extensively utilized to compare histograms such as texture, object and shape classification, and image retrieval [19]. It is defined as:

$$X^2_{p,q} = \frac{1}{n}\sum_{i=1}^{n}\frac{[p(i)-q_i)]^2}{p(i)+q(i)}, \quad \textbf{(2)}$$

p and q represent the probability distributions of two generic descriptors p and q for frame, $i = 1, 2,\ldots, n$.

The similarity distance for generic descriptor between the different crowd videos is calculated based on the average of the Euclidean distance and X^2 distance. It is defined as:

$$Sim(p,q) = \frac{1}{2}[d(p,q) + X^2(p,q)] \quad \textbf{(3)}$$

Finally, the results are ranked according their top smallest of similarity values. Table 3 summarizes all the necessary steps for the proposed CVR framework using generic descriptors.

Table 3. Description of the symbols in the algorithm

Label	Meaning
F_v	The set of all videos frames
F_{cm}	The set of n covariance matrix feature about generic descriptors for all crowd video frames
F_Q^{cm}	The set of m covariance matrix feature about generic descriptors of the given query
F_{rel}	The set of M relevant video frames

for all generic descriptors in F_v
  begin
    $d(F_Q^{cm}, F_{cm})$ //Using Equation (1)
    $x^2(F_Q^{cm}, F_{cm})$ //Using Equation (2)
    $sim(F_Q^{cm}, F_{cm}) = \frac{1}{2}[d(F_Q^{cm}, F_{cm}) + X^2(F_Q^{cm}, F_{cm})]$ //Using Equation (3)
  end
    $F_{rel} = \max j(\min(sim(F_Q^{cm}, F_{cm}))), j=1,2,\ldots,M;$
    //Select top M relevant retrieved frames
return F_{rel}.

4 Dataset

In this paper, the experiments are carried out using the CUHK Crowd Dataset [3]. It comprises of 474 video clips from 215 different scenes. 419 video clips were collected from Getty Image and Pond5. These 419 video clips are the existing crowd datasets from Ali and Shah [20], Rodriguez et al. [21], and Zhou et al. [22] researches. While 55 video clips are captured from Shao et al. [3]. The dataset covers crowd videos with different scenes, diverse densities and angle scales where the pedestrian move to occlude each other or blocked by non-human items.

The crowd videos are manually assigned into 8 classes provided by Shao et al. [3] as shown in Table 4 and Fig. 3. These 8 classes are crowd behaviors that usually happen in crowd video and used for crowd management and traffic control.

Table 4. List of crowd behavior classes

Class	Class Name	Total videos
1	Highly mixed pedestrian walking	15
2	Crowd walking following a main stream and well organized	153
3	Crowd walking following a main stream but poorly organized	72
4	Crowd merge	9
5	Crowd split	13
6	Crowd crossing in opposite directions	70
7	Intervened escalator traffic	21
8	Smooth escalator traffic	121

Fig. 3. Example for 8 classes crowd behavior

In class 1, pedestrians are seen walking in different directions with extremely diverse behaviors; such as pedestrians in streets or shopping malls. Most people follow the main stream with stable relative

positions in class 2 and there are little overtaking events. Military parade, and marathon race are categories in class 2. In class 3, most people follow the main stream, though not in a well-organized manner; such as crowd protest, marathon race, pedestrians in streets, or exiting regions examples ports and stations. Class 4 describes crowd merge and class 5 displays crowd split to avoid overcrowded spaces and collisions all exacerbate the dangers. Both classes are commonly happen in pedestrians in market or in shopping malls. Class 6 mostly is shown in crossing a zebra crossing. Classes 7 and 8 are vital to keep escalator traffic smooth to avoid blocked exits and collisions.

5 Experimental Results and Analysis

In this segment, a set of experiment is presented to validate the proposed CVR framework using generic descriptors and related crowd features. The evaluation is performed by using CUHK Crowd Dataset [3]. Each time a query video is picked from the CUHK Crowd Dataset [3], while the remained 473 videos form the set for retrieval.

Since the user is usually only interested in the top results returned by the search engine, the videos related to the query should be as high as possible [23]. Therefore, CVR performance is measured by the average precision (*AP*) in the top k retrieved samples (*AP@k*) [3-4]. *AP@k* is a variant of *AP* where only the top *k* ranked videos are considered [23]. *AP@K* is average precision value from the rank positions where a relevant video was retrieved.

$$\text{Precision } (P) = \frac{\text{Number of relevant video retrieved}}{\text{Total number of video retrieved}}, \tag{4}$$

$$AP = \frac{\sum_{k=1}^{n} p(k)\ x\ rel(k)}{\text{Number of relevant video}} \tag{5}$$

where *P(k)* is precision at rank *k* and *rel(k)* is an indicator function in which equal to 1 if the video at rank *k* is a relevant video, zero otherwise. For example, a query that has three relevant crowd videos which are retrieved at ranks 1, 3, and 9. The actual precision attained when each relevant crowd videos is retrieved is 1, 0.67, and 0.33. The *AP@k* over all relevant crowd videos for this query is 0.67.

The influence of video frames using the graph-based descriptors in average top 10 precision evaluation are shown in Fig. 4. The generic descriptors obtained from the proposed CVR framework are represented by graph-based descriptors. The graph-based descriptors presents an increasing trend when increase video frames from 10 to 30 frames. It achieves 49% in average top 10 precision. The graph-based descriptors do not improve accuracy of CVR significantly with continuous increase of the video frames from 40 to 80 frames. Even though longer frames generate more temporal information in which benefit in motion modelling, but the result achieved by 80 frames is 52% in average top 10 precision. Result shows only a subtle improvement from 30 frames onwards. Hence, satisfactory CVR results can be obtained based on fewer frames because the motion patterns in each segmented clip remains similar within a crowd video. Therefore, we only takes the 30 frames from each video clip for executing the graph-based descriptors.

Fig. 5 illustrates the quantitative comparison of the graph-based descriptors with group descriptors [3], and bilinear CD descriptors [4] for CVR average precision at top *k* (*AP@k*) over all the crowd behavior classes. Fig. 5 represents that the graph-based descriptors perform better than previous works in average top 10 precision to average top 100 precision. It obtained 49% in average top 10 precision and increases 16.7% if compare with group descriptors [3] and increases 8.9% if compare with bilinear CD descriptors [4]. The results progressively become similar between them with the increasing value of *k* if compare with group descriptors [3] and bilinear CD descriptors [4] in average top 100 precision. It only increases 7.3% if compare with group descriptors [3] and increases 4.9% if compare with bilinear CD descriptors [4].

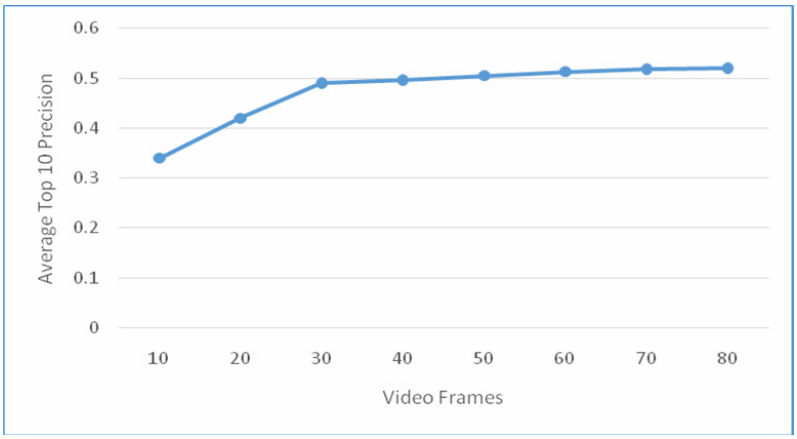

Fig. 4. Influence of video frames to the graph-based descriptors in average top 10 precision

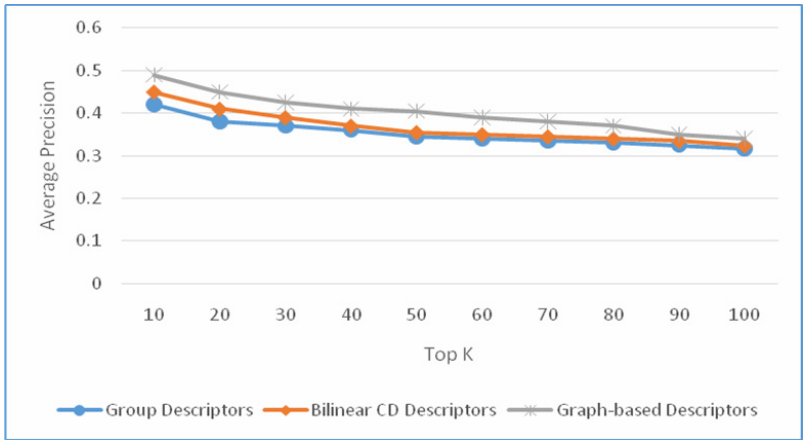

Fig. 5. The quantitative average precision at top *k* comparison of the graph-based descriptors with previous works

Fig. 6 illustrates the accomplishment of the average precision at top 100 of each crowd behavior classes. The graph-based descriptors show slight improvement on some classes if compared with group descriptors [3] and bilinear CD descriptors [4] in average top 100 precision. It performs better on the 1, 3, and 5 crowd behavior classes. Fig. 7 shows the example of query videos in 1, 3 and 5 crowd behavior classes and their retrieval results. The first query video represents the crowd behavior of class 1, in which people in shopping malls and its top five similar crowd motions. Graph-based descriptors achieve 3 of the top 5 retrieval results are relevant if compared with group descriptors [3] and bilinear CD descriptors [4]. The irrelevant retrieval results belong to the class 3 or 5. Graph-based descriptors in class 1 describe more accurately crowd motions in unstructured crowd scenes with pedestrians walk in dissimilar directions with extremely diverse behaviors. The second query video describes the crowd behavior in class 3 in which most people follow the main stream not in a well-organized manner when exiting from station. However, this second query video resembles the crowd split as in class 5. All the irrelevant retrieval results in second query video are considered belong to the class 5. Hence, results show that group descriptors [3] and bilinear CD descriptors [4] do not work well in crowd split videos of class 5 due confusion of the query video is very similar to class 3. Group descriptors [3] and bilinear CD descriptors [4] pose the same reason as mentioned above in third query video. The third query video describes the crowd behavior in class 5. All the irrelevant retrieval results in third query video for group descriptors [3] and bilinear CD descriptors [4] are belongs to the class 3. Graph-based descriptors in this study improve the correct track of each person involved in the occlusion after they split up from crowd merge in class 5.

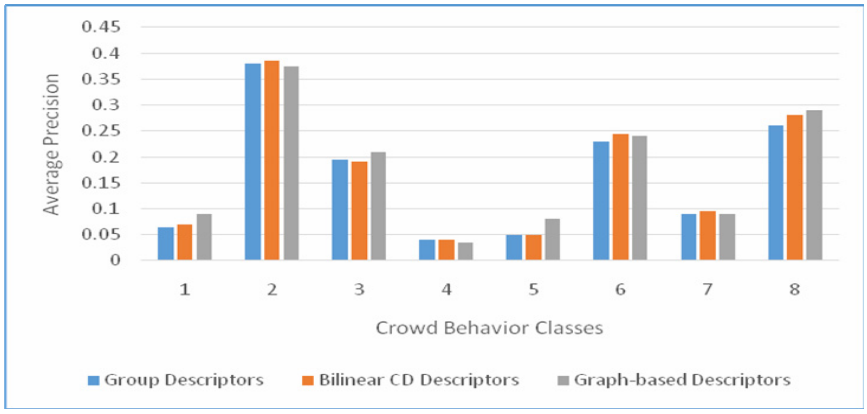

Fig. 6. The quantitative average precision at top 100 of each crowd behavior classes

Fig. 7. The example of query videos in 1, 3 and 5 crowd behavior classes and their retrieval results

Overall, the performance improvement reveals the effectiveness of the learning generic descriptors for CVR in different crowd scenes, and crowds with various densities and structure.

6 Conclusion

In this study, an effective crowd video retrieval framework that employs generic descriptors has been presented. The whole framework includes four main components, namely motion feature extraction, group detection, learning generic descriptors, and crowd video retrieval. Kanade–Lucas–Tomasi feature point tracker is used to detect and track moving objects, and then tracklets are grouped to form trajectories. Second component is detect groups by clustering trajectories. The third component is to employs group motion pattern mining and prediction approach to find out collectiveness, uniformity, stability, and conflict generic descriptors to describe crowd behavior for different crowd scenes. The last component is the Euclidean distance and Chi-Square distance are applied to quantify the similarity matching between the query video and the remained video clips. Finally, the most relevant videos are retrieved. The experimental results have shown significant improvement in accuracy. The proposed crowd video retrieval framework using generic descriptors overcomes the challenges of crowd video retrieval in crowds with various densities, structures, occlusion and different scenes.

This study only achieves a subtle improvement when used longer frames in motion modelling for crowd video retrieval. Therefore, we will exploit deep learning techniques to find out effective descriptors for crowd understanding in future work. We will also explore metric learning algorithms to improve the retrieval ranking results.

Acknowledgements

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

References

[1] F. Chamasemani, L.S. Affendey, N. Mustapha, F. Khalid, A framework for automatic video surveillance indexing and retrieval, Res. J. Appl. Sci. Eng. Technol. 10(11)(2015) 1316-1321.

[2] F. Chamasemani, L.S. Affendey, N. Mustapha, F. Khalid, Speeded up surveillance video indexing and retrieval using abstraction, in: Proc. 2017 IEEE International Conference on Signal and Image Processing Applications, 2017.

[3] J. Shao, C.C. Loy, X. Wang, Learning scene-independent group descriptors for crowd understanding, IEEE Trans. Circuits Syst. Video Technol. 27(6)(2017) 1290-1303.

[4] S. Wu, H. Su, H. Yang, S. Zheng, Y. Fan, Q. Zhou, Bilinear dynamics for crowd video analysis, J. Vis. Commun. Image Represent. 48(2017) 461-470.

[5] S. Wu, H. Yang, S. Zheng, H. Su, Q. Zhou, X. Lu, Motion sketch based crowd video retrieval, Multimed. Tools Appl. 76(19)(2017) 20167-20195.

[6] J. Dong, X. Li, C.G.M. Snoek, Predicting visual features from text for image and video caption retrieval, IEEE Trans. Multimed. 9210(2018) 1-12.

[7] Y. Zhang, L. Qin, S. Zhao, R. Ji, X. Lu, H. Yao, Q. Huang, Crowd video retrieval via deep attribute-embedding graph ranking, in: Proc. 2016 IEEE International Conference on Multimedia and Expo, 2016.

[8] S. Wu, H. Yang, S. Zheng, H. Su, Y. Fan, M.-H. Yang, Crowd behavior analysis via curl and divergence of motion trajectories, Int. J. Comput. Vis. 123(3)(2017) 499-519.

[9] X. Wang, C. Loy, Deep learning for scene-independent crowd analysis, in: V. Murino, M. Cristani, S. Shah, S. Savarese (Eds.), Group and Crowd Behavior for Computer Vision, Elsevier, 2017, pp. 209-252.

[10] J. Shao, K. Kang, C.C. Loy, X. Wang, Deeply learned attributes for crowded scene understanding, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[11] W. Liu, R.W.H. Lau, D. Manocha, Robust individual and holistic features for crowd scene classification, Pattern Reconit. 58(2016) 110-120.

[12] S. Yi, X. Wang, C. Lu, J. Jia, H. Li, L_0 regularized stationary-time estimation for crowd analysis, IEEE Trans. Pattern Anal. Mach. Intell. 39(5)(2017) 981-994.

[13] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, F. Bremond, Toward abnormal trajectory and event detection in video surveillance, IEEE Trans. Circuits Syst. Video Technol. 27(3)(2017) 683-695.

[14] C. Tomasi, Detection and tracking of point features technical report CMU-CS-91-132, Image Rochester NY 91(4)(1991) 1-22.

[15] B. Zhou, X. Tang, X. Wang, Coherent filtering: detecting coherent motions from crowd clutters, LNCS. 7573(2)(2012) 857-871.

[16] P.V. Wong, N. Mustapha, L.S. Affendey, F. Khalid, A new clustering approach for group detection in scene-independent dense crowds, in: Proc. 2016 3rd International Conference on Computer and Information Sciences, 2016.

[17] P. Bagherpour, S.A. Cheraghi, M. Bin Mohd Mokji, Upper body tracking using KLT and Kalman filter, Procedia Comput. Sci. 13(2012) 185-191.

[18] E. Deza, M.M. Deza, Encyclopedia of Distances, Springer Berlin Heidelberg, 2009.

[19] O. Pele, M. Werman, The Quadratic-Chi histogram distance family, LNCS. 6312(2)(2010) 749-762.

[20] S. Ali, M. Shah, A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: Proc. 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.

[21] M. Rodriguez, J. Sivic, I. Laptev, J.Y. Audibert, Data-driven crowd analysis in videos, in: Proc. 2011 IEEE International Conference on Computer Vision, 2011.

[22] B. Zhou, X. Tang, H. Zhang, X. Wang, Measuring crowd collectiveness, IEEE Trans. Pattern Anal. Mach. Intell. 36(8)(2014) 1586-1599.

[23] Z. Cheng, J. Shen, H. Miao, The effects of multiple query evidences on social image retrieval, Multimed. Syst. 22(4)(2016) 509-523.