

# An Effective Crowd Property Analysis System for Video Surveillance Application

Shuying Yang<sup>1(✉)</sup>, Hua Yang<sup>1,2</sup>, Jijia Li<sup>1,2</sup>, and Ji Zhu<sup>1,2</sup>

<sup>1</sup> Institute of Image Communication and Network Engineering,  
Shanghai Jiao Tong University, Shanghai, China  
{labeouf,hyang}@sjtu.edu.cn

<sup>2</sup> Shanghai Key Laboratory of Digital Media Processing and Transmission,  
Shanghai, People's Republic of China

**Abstract.** For public security, an intelligent video surveillance system that can analyze large-scale crowd scenes has become an urgent need. In this paper, we propose a system that integrates multiple crowd properties, including stationary and dynamic features, local and global characteristics, and historic statistics analysis in a unified framework. Specially our system consists of four modules. Crowd density module describes global density level and local density distribution with sparse spatial-temporal local binary pattern. Crowd segmentation module presents both global crowd grouping and local moving directions based on spatial-temporal dynamics. In crowd saliency module, salient regions are detected to alarm abnormal behaviors. At last, in order to analyze the historic features of video streaming, a historical statistics analysis module is introduced. Experiments on different crowd datasets show that our system is robust and feasible, and satisfies the requirements of video surveillance applications.

**Keywords:** Density distribution · Motion consistency · Crowd saliency · Historical statistics analysis

## 1 Introduction

With the rapid growth of population and human activities, video surveillance system for analysis of crowd scenes has attracted much attention in the field of computer vision. Public security or management in a high density crowd becomes a significant challenge, due to excessive number of individuals, extreme clutters and complexity of scenarios.

In the past decades, researchers have made great progress in the computer vision community for intelligent video surveillance. Automatic crowd analysis can be conducted at both microscopic level and macroscopic level. At the microscopic level, we concern about the movements of each individual. Mehran [1] detected abnormal behaviors in the crowd using the SFM and Zhou [2] proposed a dynamic pedestrian-agent model for semantic region analysis. Methods at microscopic level cannot adapt to high density crowds. To analyze crowd

scenes at the macroscopic level, Ali [3] designed a framework which is used to segment high density crowd scenes. In paper [4], a novel method was proposed to anomaly detection with crowd flow model. In order to analyze complex high density crowd scenes, many methods based on dynamics have been designed [5,6]. These methods are not applied in the real complex scenarios with heavy crowded scenes and can only achieve to analyze single kind of features. As for the intelligent surveillance system, it has gone through three stages, the analog video surveillance system, the analog-digital monitoring system, network video surveillance [7]. In [8], a dual-camera system was proposed to accomplish the vision-based recognition but it failed to detect the object when there are multiple moving objects. Carnegie Mellon University cooperated with DARPA and other institutes to develop an automatic video understanding technique for future city and war. University of Reading pursued research in tracking of pedestrian and vehicles and the interaction between them. Swedish Axis network communication company launched a product which includes an AXIS 242S IV video server and a people counting module application. Conventional surveillance systems fail to adopt to high density crowd scenes, regarding both accuracy and computation [9]. They only focus on specific scenes and are not applicable for large-scale crowd scenes. These crowd monitoring system cannot be applied in real scenarios because of the simplicity and the impractical assumptions of the modeling function.

Based on the need of public security, our goal is monitoring safety hazard in crowded scenes, and formulating specific technical routines with quantitative evaluation methods. Ideally we will help improve the public security departments capability of dealing with sudden mass disturbance, as well as surveillance on important events. The movements in crowded scenes are complicated, and thus it is difficult to characterize group behaviours. In order to gain effective and quick perception of abnormal group behaviours in large-scale crowd scenes, this paper proposes an intelligent surveillance system using multichannel videos, by leveraging the advanced technique of machine learning and computer vision. The system can not only estimate dynamic and stationary features, but also describe local and global characteristics in an unsupervised way.

The rest of this paper is organized as follows. In Sect. 2 we present the framework of our system and illustrate our system in detail. Section 3 describes software implementation of our system. In addition, the experiments and conclusions are presented in Sects. 4 and 5.

## 2 The Framework of Crowd Analysis System

In this paper, the novelty of our research is that we combine stationary and dynamic features, local and global characteristics, and historic statistics analysis in a unified framework, and design an intelligent video surveillance system as illustrated in Fig. 1. The proposed system consists of four main modules: crowd density module, crowd segmentation module, crowd saliency module, and historical statistics analysis module.

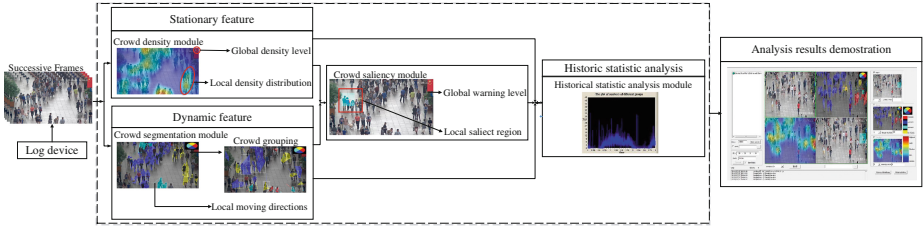


Fig. 1. Framework of the proposed system

These four modules cooperate with each other in the whole system. The crowd density module extracts stationary feature, sparse spatial-temporal local binary pattern (SST-LBP) [10], and provides local density distribution as well as global density level. The crowd segmentation module describes the whole crowd grouping with dynamic features including temporal motion grouping and distribution grouping. Meanwhile, this module also presents local moving direction of each grouping. The saliency module detects salient regions, where abnormal behaviors will happen. Abnormal activities can be detected with the density level results and the semantic segmentation, both locally and globally. The historical statistics analysis module can give out statistical distribution of crowd features, crowd density level and numbers of moving groups with different directions. In following sections, we provide the details on how to implement the four functional modules.

## 2.1 Crowd Density Module

As illustrated in Fig. 1, in our crowd density module stationary SST-LBP feature is extracted to deal with the crowd density, which contains two properties: local density distribution and global density level. We calculate the density level of crowd scenes and present crowd density spatial distribution. The density of crowd scene provides information for warning potential risk in the crowd due to the fact that panic-stricken stampede and overcrowding occur in large-scale crowd. In this module, we analyze distribution of different density levels for large-scale crowded scenes. The diagram of this module is shown in Fig. 2.

In real world crowd, there exist different crowd scenes with different density levels. An improved SST-LBP algorithm is applied to analyze crowd density in this module. Based on the original SST-LBP algorithm [10], which has good

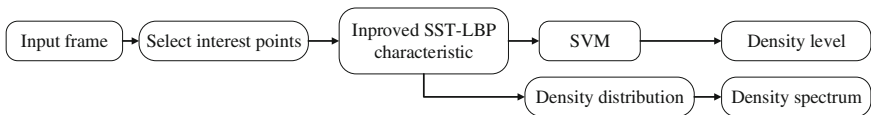
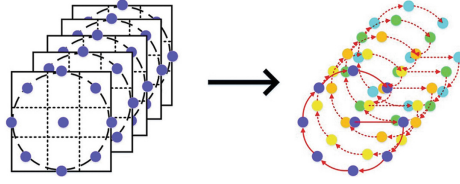


Fig. 2. Pipeline of crowd density module.



**Fig. 3.** Schematic diagram of the calculation of SST-LBP code.

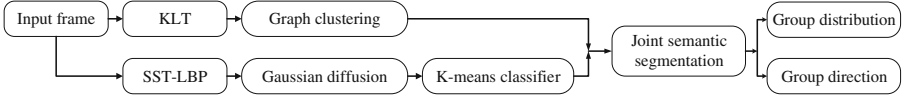
performance on all density levels especially the large-scale crowd, we improve the calculation of spatial-temporal local binary as shown in Fig. 3. This method can present local density distribution. The volume to calculate SST-LBP code is set to  $3 \times 3 \times 5$ . The SST-LBP code is calculated along the helix. Also to improve the performance of density distribution, we sum up the Fourier coefficients of the center pixel and its neighboring pixels as the measure of the density of the center pixel and then apply Gaussian smoothing. The histogram of the spectrum analysis is the feature representation of the SVM model to specify the density of crowd.

Through above analysis, local density distribution can easily be found by labeling different colors and the places where there are more people are labeled with more attractive colors like red. We also obtain the global density level which is helpful for high-density crowd analysis and salient regions detection.

## 2.2 Crowd Segmentation Module

In this crowd segmentation module, crowd semantic segmentation based on spatial-temporal dynamics is used to perform both global and local grouping features. Global feature is addressed by segmenting, whereas local feature is performed using different directions of moving groups.

Individuals in a crowd interact with each other, whose trajectories are affected by both themselves and their neighbors. When two individuals are distant, there are few interactions between them. These two individuals cannot reflect each other. In this module, we propose crowd semantic segmentation based on spatial-temporal dynamics to achieve crowd segmentation as shown in Fig. 4. First, we use KLT algorithm [11] to detect trajectories of the crowd. In order to segment the crowd based on motion consistency, we cluster the trajectories with graph clustering algorithm proposed by [12]. After these steps, we obtain the temporal motion grouping. Then distribution grouping is processed to detect individuals with small distance and similar distribution, which are classified as one group. SST-LBP describes local particle distribution in neighbors. A set of Gaussian kernels with 5 different sizes are designed to organise the local distribution to generate the global distribution. After obtaining the density diffusion maps, k-means is used as a classifier to generate a 5-level grouping, which is called spatial distribution grouping. At last, we combine the spatial and temporal distribution group as semantic descriptions, which contain motion and



**Fig. 4.** Pipeline of crowd segmentation module.

distribution consistency. Group directions are defined as the mean velocity of individuals in one group. In this module, we utilize spatial-temporal distribution grouping to gather individuals with similar moving directions. The moving direction is divided into four groups, direction 3 o'clock, direction 6 o'clock, direction 9 o'clock and direction 12 o'clock. Also, individuals in the same group are labeled with the same color.

In summary, this module shows moving groups with different directions and the numbers of groups, which can reflect the degree of chaos of crowd motion. Masses of moving directions in a crowd scene lead to collision between individuals, which often cause abnormal behaviors. Moving groups with different directions are detected in this module to provide information for analysing interaction between individuals in crowd scenes, which makes contribution to detecting abnormal behaviors in crowd saliency and alarm module.

### 2.3 Crowd Saliency Module

Crowd saliency module detects salient regions and reports abnormal behaviors, such as an individual maneuvering through crowd, two groups meet and unstable regions. Different from traditional definition of saliency, our saliency based on abnormal behaviors detection describes possibility of unstable conditions. This module presents a warning level describing the salient degree. Based on the above two modules' results, we can analyze interactions between different groups and density of the crowd. Salient regions can be detected by combining density level and motion consistency.

A novel framework transforming low-level features into global similarity structure is applied to identify and localize salient regions proposed in [13]. The framework estimates crowd motion field to extract features to represent crowd dynamics, stability map and phase shift map. Stability map presents particle advection demonstrating spatial distribution and phase shift map describes the velocity phase similarity among individuals in a crowd scene. According to the novel framework, we can conclude that spatial distribution and velocity phrase difference can be applied to detect salient regions. Thus, density distribution which reflects spatial distribution can be denoted as stability map in this module. Motion consistency and directions of moving groupings presenting similarity of global motion field are related to phase shift map. Considering the density level and the motion consistency, we can detect salient regions, where there exist abnormal behaviors. When the salient regions are detected, the warning level are defined as in Table 1.

**Table 1.** Explanation of the saliency index.

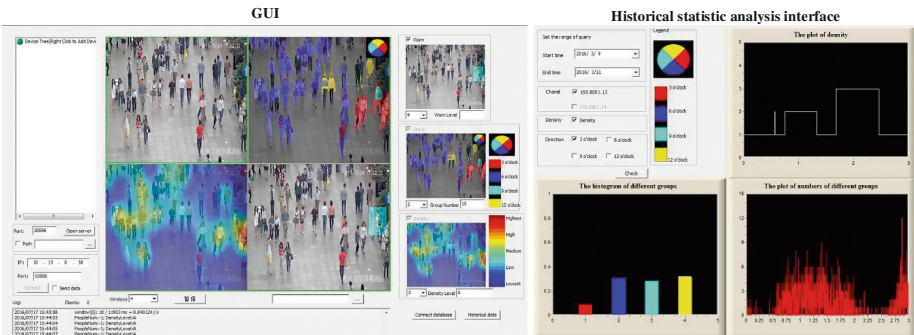
Density level	Group numbers	Direction numbers	Warn level
<C	<5	<=2	1
<C	<5	<=2	2
>=C	<5	>=3	3
>=C	>=5	>=3	4

We propose this module for analysis of group behaviours in crowded scenes and application model, in order to analyze the characterization, categorization and evaluation of abnormal incidents, setting the foundation for alarming in public security video surveillance.

### 2.4 Historical Statistics Analysis Module

In order to observe both the trend of crowd density and the number of groups with different directions and report the level of crowd directly, a historical statistics analysis module is designed. This module presents the historic statistic analysis of our system. The interface of this module is shown in Fig. 5.

We choose NTGraph control to show the results from the query in the database of different periods as users set. There are three schemes showing different information of the selected channel. The first scheme describes the trends of density level. The second one depicts the changing proportion of different groups. The last one shows the changes in the number of groups with different moving directions. In this module, users can set the range of time they want to query. Then they can choose which kind of feature and which channel they inquire. This system shows different query results with regard to different users.



**Fig. 5.** The user interface of surveillance system.

### 3 Software Implementation of System

The user interface of our video surveillance is depicted by Fig. 5, including GUI and historical statistics analysis interface. After logging in device, video stream (as shown in Fig. 1) obtained from Hikvision IP camera located in East Nanjing Road and the Bund is transmitted to an analysis server, which provides analysis information. Its basic functions including playing videos, connecting to server, enrolling devices are implemented in Hikvision SDK. In following subsections, we will introduce software implementation of our system in detail.

#### 3.1 Multichannel Monitoring

Multichannel monitoring plays a vital role in an effective iss. Our system can monitor 1, 4, 9, 16 channels simultaneously. In our system, the users need to firstly log in the system and then choose a channel to monitor. The GUI provides multiple windows to show multichannel video streaming with each window belonging to the same dialog class that can change the size of windows and play video. In addition, the analysis results of three video surveillance modules can all be shown in the windows.

#### 3.2 Multithread and Buffer

Threads are designed to analyze different features of crowd scenes. Correspondingly, we utilize Hikvision SDK as the general framework and multithread is applied in historical statistics analysis module when users query data of different features. To avoid conflictions between writing and reading data, we create an array as a buffer, whose data queues depending on basic first-in-first-out structure. After the frame analysis we record the frame information into the buffer and then read data from it. It should be noted that in each thread, buffer should be locked when the process is reading or writing data.

#### 3.3 Database Design and Statistics Display

Our system analyzes the density level, numbers of groups with different directions, and the crowd saliency level, all of which should be stored for historical statistics analysis module. We choose MySQL database and call MySQL statements in Visual Studio 2013. Moreover, in order to manage data effectively we design two tables.

To deal with the diversity of transmission channels and devices, a MySQL database is applied to record the information of channels, which can be mapped to ‘deviceid’ field in a table shown in Table 2. This table is used to store the information of channel and to avoid repeating storage the same channel. Another table as shown in Table 3 presents all properties of each frame, where the ‘id’ of each ‘cameracode’ has the same meaning as ‘deviceid’ in second table. When users select a channel to play video, the related information will be added to to the first table if the same channel is not stored.

**Table 2.** Device information.

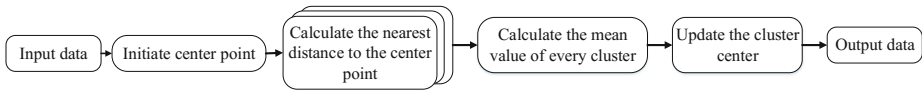
Column	id	Camerascode	Name	Remarks
Meaning	Device id	Channel number	Camera name	Extra information

**Table 3.** Features information.

Column	Meaning	Column	Meaning
Timestamp	Time	imgpath	Image storage address
Deviceid	Device id	dirnum0	3 o'clock direction
Densitylevel	Density level	dirnum1	6 o'clock direction
Groupnum	Total groups	dirnum2	9 o'clock direction
Warnlevel	Salience	dirnum3	12 o'clock direction

### 3.4 Improvement of Speed Based on CUDA

The speed of segmentation and crowd density is mainly determined by k-Means and Gaussian blur process respectively. In this section, we propose parallel algorithms for k-Means Cluster and Gaussian blur and implement them on GPU, which can improve the processing speed of system.

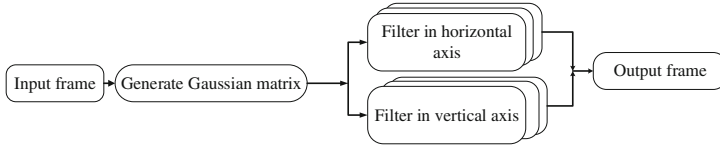
**Fig. 6.** Illustration of k-Means clustering algorithm.

In k-means algorithm the distance between each point to its corresponding cluster center is computed in step 3 as shown in Fig. 6, which is parallelized at patch-level. The kernel function computes the k-nearest points in different threads. In the Gaussian blur process, each point in original images is filtered in horizontal and vertical axis, the process of which includes lots of matrix calculations. CUDA has a good performance on matrix calculation so we can rewrite Gaussian blur as shown in Fig. 7 on CUDA to improve the speed of our system. Compared with traditional k-Means cluster and Gaussian blur, the CUDA implementation of the algorithms achieves a speedup over the CPU implementation.

## 4 Evaluations of System Performance

To evaluate our system performance, we conduct experiments on PETS, UFC, CUHK datasets and the real word datasets. In our system, the specifications of hardware are as follows, Core Intel i7-3770U 3.4 GHz and 8 GB RAM.





**Fig. 7.** Illustration of Gaussian blur algorithm.

**Table 4.** Experimental results of density level estimation.

Scene	Method	Density level accuracy			
		Free	Restricted	Dense	Jammed
PETS	Pixel statistic	0.923	0.87	0.82	0.79
	Texture feature	0.87	0.82	0.86	0.904
	Our system	0.92	0.89	0.87	0.89
Real Scenario	Our system	0.943	0.903	0.861	0.899

As for the crowd density module, we choose PETS dataset and our own sets of video of East Nanjing Road and the Bund. In order to evaluate the density level accuracy of the proposed system, a quantitative comparison is conducted between different features based on effective region feature extraction [14] and our system as shown in Table 4. The results indicate that our system achieves favorable performance on all density levels. The crowd density distribution is shown in Fig. 8, from which we can see that spatial distribution are presented accurately. The regions with more people are labeled with catchier colors.



**Fig. 8.** The density distribution experimental results of the crowd density module.

As for the crowd segmentation module, we introduce Precision, Recall,  $F_1$ -score and segmentation accuracy to analyze the performance of our system. Definition of these metrics are given as below,

(1) Precision = Number of truly detected groups/Number of all detected groups;

(2) Recall = Number of truly detected groups/Number of all ground truth groups;

(3)  $F_1$ -score =  $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$ ;

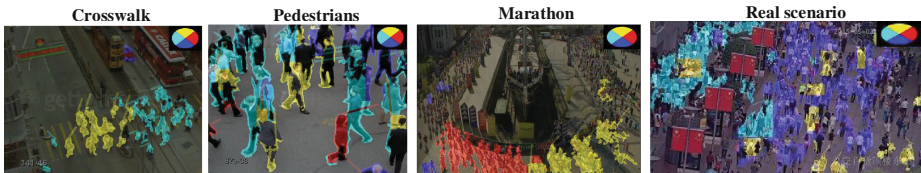
(4) Segmentation Accuracy = Number of truly labeled pixels/Number of all labeled ground truth pixels.

A truly detected group is defined as the group which has more than 50% area overlapping with ground truth group. The performance of this module is evaluated on UCF dataset, CUHK dataset compared with FTLE [3] and CF [15]. Evaluation results are shown in Table 5, from which we can see our system performs better compared with the other two methods on  $F_1$ -score and accuracy of segmentation areas. The qualitative segmentation results of datasets and real scenarios are shown in Fig. 9. We can see that our system segments crowd accurately with continuous and complete patches. Our system also provides moving directions of different groups.

**Table 5.** Results of different methods on the representative crowd datasets. The best results are marked in bold.

Method	Scene	Group detection			Group segmentation
		Precision	Recall	$F_1$ -score	Accuracy
FTLE [3]	Mecca	0.33	0.48	0.39	0.86
	Crosswalk	0.45	0.66	0.54	0.76
	Pedestrians	0.24	0.24	0.24	0.53
CF [15]	Mecca	0.16	0.22	0.19	<b>0.93</b>
	Crosswalk	0.64	0.29	0.40	0.88
	Pedestrians	0.35	0.56	<b>0.43</b>	0.71
Our system	Mecca	0.42	0.50	<b>0.46</b>	0.90
	Crosswalk	0.62	0.54	<b>0.58</b>	<b>0.90</b>
	Pedestrians	0.33	0.63	<b>0.43</b>	<b>0.72</b>

The performance of this module is evaluated on UCF dataset, CUHK dataset compared with FTLE [3] and CF [15]. Evaluation results are shown in Table 5, from which we can see our system performs better compared with the other two methods on  $F_1$ -score and accuracy of segmentation areas. The qualitative segmentation results of datasets and real scenarios are shown in Fig. 9. We can see that our system segments crowd accurately with continuous and complete patches. Our system also provides moving directions of different groups.

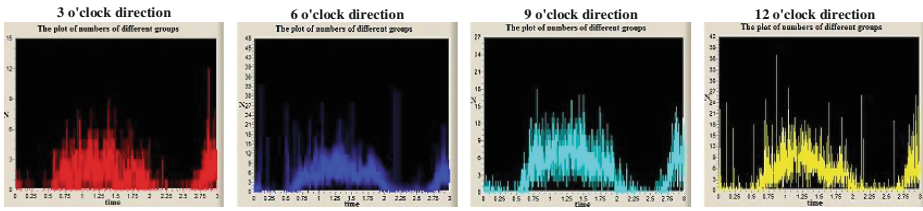


**Fig. 9.** The grouping distribution experimental results of the crowd segmentation module.



**Fig. 10.** The salient regions detection experimental results of the crowd saliency module.

The crowd saliency module of this system can detect salient regions, where two different direction group meet, an individual maneuvering through crowd and some other unstable events happen as shown in Fig. 10. Our saliency is different from traditional definition of saliency and we focus on whether there exists abnormal behaviors. The historical statistics analysis module can show query results of numbers of moving groups with different directions as shown in Fig. 11. As Tables 6 and 7 show, implementing the k-means algorithms and Gaussian blur on CUDA can speed up our system. Comparing to the implementation on CPU, the parallelization of k-means algorithms cuts the runtime in half.



**Fig. 11.** The plot of different directions

**Table 6.** Comparison of the k-Means runtime of CPU and CUDA

Image no	CPU Runtime(ms)	CUDA runtime(ms)
1	3495	1503
2	3000	1430
3	3622	1595
4	3515	1444
5	3593	1437
6	4477	1892
7	4586	1782
8	4460	2019

**Table 7.** Comparison of the Gaussian runtime of CPU and CUDA

Image no	CPU runtime(ms)	CUDA runtime(ms)
5	1529	1278
6	928	882
7	922	883
8	921	883
9	927	882
10	924	882
11	922	882
12	927	881
13	933	884

## 5 Conclusion

In this paper, we develop a unified framework that combines multiple crowd properties including stationary and dynamic features, local and global characteristics to analyze crowd scenes in a large-scale area. Our proposed system consists of crowd density module, crowd segmentation module, crowd saliency module, and historical statistics analysis module. Experimental results demonstrate our system's robustness and feasibility, which achieve favorable performance in various crowd scenes. In this case, our system can only give a warning level for salient regions. Hence, how to recognize the type of abnormal behaviors is an important problem, which will be explored in further.

**Acknowledgement.** This work was supported in Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos.15DZ1207403).

## References

1. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Conference on Computer Vision and Pattern Recognition, pp. 935–942 (2009)
2. Zhou, B., Wang, X., Tang, X.: Random field topic model for semantic region analysis in crowded scenes from tracklets. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441–3448 (2011)
3. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2007)
4. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. IEEE Conf. Comput. Vis. Pattern Recognit. **238**(6), 2054–2060 (2010)
5. Mehran, R., Moore, B.E., Shah, M.: A streakline representation of flow in crowded scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6313, pp. 439–452. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15558-1\\_32](https://doi.org/10.1007/978-3-642-15558-1_32)

6. Allain, P., Courty, N., Corpetti, T.: Crowd flow characterization with optimal control theory. *Appl. Biochem. Microbiol.* **51**(4), 432–441 (2010)
7. Raty, T.D.: Survey on contemporary remote surveillance systems for public safety. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **40**(5), 493–515 (2010)
8. Bodor, R., Morlok, R., Papanikolopoulos, N.: Dual-camera system for multi-level activity recognition. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1, pp. 643–648 (2004)
9. Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S.: Crowded scene analysis: a survey. *IEEE Trans. Circ. Syst. Video Technol.* **X**(1), 367–386 (2015)
10. Yang, H., Cao, Y., Su, H., Fan, Y., Zheng, S.: The large-scale crowd analysis based on sparse spatial-temporal local binary pattern. *Multimedia Tools Appl.* **73**(1), 41–60 (2014)
11. Tomasi, C.: Detection and tracking of point features. Technical report (1991). **9**(21), 9795–9802
12. Zhang, W., Wang, X., Zhao, D., Tang, X.: Graph degree linkage: agglomerative clustering on a directed graph. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*. LNCS, vol. 7572, pp. 428–441. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-33718-5\\_31](https://doi.org/10.1007/978-3-642-33718-5_31)
13. Mei, K.L., Kok, V.J., Chen, C.L., Chan, C.S.: Crowd saliency detection via global similarity structure. In: *ICPR*, pp. 3957–3962 (2014)
14. Su, H., Yang, H., Zheng, S.: The large-scale crowd density estimation based on effective region feature extraction method. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) *ACCV 2010*. LNCS, vol. 6494, pp. 302–313. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-19318-7\\_24](https://doi.org/10.1007/978-3-642-19318-7_24)
15. Zhou, B., Tang, X., Wang, X.: Coherent filtering: detecting coherent motions from crowd clutters. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*. LNCS, pp. 857–871. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-33709-3\\_61](https://doi.org/10.1007/978-3-642-33709-3_61)