Novel Dataset for Fine-grained Abnormal Behavior Understanding in Crowd

Hamidreza Rabiee[1] Javad Haddadnia[2] Hossein Mousavi[3] Maziyar Kalantarzadeh[4]
Moin Nabi[5] Vittorio Murino[3]
[1]Department of Electrical Engineering, Hakim Sabzevari University, Iran
[2]Department of Biomedical Engineering, Hakim Sabzevari University, Iran
[3]PAVIS, Istituto Italiano di Tecnologia, Italy
[4]NODET, Iran [5]DISI, University of Trento, Italy

Abstract

Despite the huge research on crowd on behavior understanding in visual surveillance community, lack of publicly available realistic datasets for evaluating crowd behavioral interaction led not to have a fair common test bed for researchers to compare the strength of their methods in the real scenarios. This work presents a novel crowd dataset contains around 45,000 video clips which annotated by one of the five different fine-grained abnormal behavior categories. We also evaluated two state-of-the-art methods on our dataset, showing that our dataset can be effectively used as a benchmark for fine-grained abnormality detection. The details of the dataset and the results of the baseline methods are presented in the paper.

1. Introduction

Population growth and crowd behavior diversity have made crowd analysis target for different studies in a variety of areas over the last few years. It makes security, safety and managing of people more challenging issue in public and private places. It also meets a growing demand because almost everybody searches for a way to keep its belongings safe and secure. Analyzing crowd footage is one of the most efficient ways to evaluate following issues: it can help to i) better figure out crowd dynamics and relevant behaviors for public space design [17], ii) present public safety, group activity monitoring [18] and crowd control system [12], iii) visual surveillance [13, 23, 1], and also iv) establish of mathematical models which can present more precise simulation and applying them in computer games, movies, and television industries [2]. The way individuals are oriented in the scene is a very important parameter, which can affect the efficiency of crowd analysis algorithms. Crowded scenes can be placed in two categories based on the motion of the crowd [19]: structured one and unstructured one. In a former one, the motion of a crowd does not change repeatedly, and each spatial location of the scene includes just one major crowd behavior during the time. In contrast, the latter one shows disordered or random individuals motions and they might be randomly in different directions, and several crowd behaviors might occur in each spatial location [21]. Fig. 1 a-b shows an example of structured crowded environment and unstructured one. Apparently, they have different dynamics and visual specifications. An unstructured crowded scene seems to be a better choice for an algorithm to yield more realistic results. Individuals are capable of extracting helpful information from behavior models in the surveillance region, monitoring the scene for unusual events in real time and taking immediate action [7]. However, psychological research shows that the ability to monitor concurrent signals is really limited in humans [24]. In the extremely crowded scenes, multiple individuals and their behaviors have to be monitored which is a substantial issue even for a human observer. As a result, there is still a significant gap between efficiency of abnormal behavior detection in research labs and the real world because the majority of abnormal detection algorithms are tested on datasets having only a small number of abnormal behavior classes taken under controlled circumstances with similar scenarios. Although in the past few years many algorithms have been presented to track, recognize and understand the behaviors

Figure 1. Example of crowded scene. (a): Stage of a bicycle race (oriented scene). (b): people assembling in outdoor (disoriented scene).

978-1-5090-3811-4/16/$31.00 ©2016 IEEE

Dataset	UMN [12]	UCSD [11]	PETS2009 [5]	Violent Flows [6]	Rodriguez's [20]	UCF [22]
Number of videos	11	98	59	246	520	61
Annotation level	frame	frame/pixel	frame	video	video	video
Density	medium	high/medium	medium	dense	dense	dense
Type of scenarios	panic	abnormal object	panic	fight	pedestrian	crowd
Indoor/Outdoor	both	outdoor	outdoor	outdoor	outdoor	outdoor

Table 1. Datasets for crowd behavior analysis

of different objects in the videos [8], lack of publicly available benchmark datasets led not to have a common test bed for researchers to compare their algorithms. We select a set of criteria, by which we can compare proposed crowd datasets. These criteria include: *number of videos, annotation level, density, type of scenarios, Indoor/Outdoor*

The most important characteristics of previous crowd datasets based on aforementioned set of criteria are presented in table 1. As can be seen in table 1, every single proposed crowd dataset is useful for limited number of applications and can not be described as a comprehensive test bed for crowd analysis algorithms. As first contribution of this paper, a more comprehensive crowd dataset with many realistic scenarios is presented. Our dataset consists of a big set of video clips annotated with crowd behavior labels (e.g., "panic", "fight", "congestion", etc.). We use unstructured crowded scenes in our dataset, so individuals are free to choose random directions and change their ways as they want. We evaluated a set of state-of-the-art feature descriptors on our dataset, showing that it can be effectively used as a benchmark in crowd analysis communities. Unlike previous crowd datasets with limited number of crowd behavior scenarios, our dataset consists of different behavior types implemented by various scenarios make it more realistic. Using different objects, some abnormal conditions are also created in our dataset. Specifically, we address *fine-grained* abnormal behavior understanding - for example, not just detecting the abnormal events, but determining what is the "type" of abnormality - in the crowded environments. We separately evaluate the state-of-the-art feature descriptors by ground-truth behavior information extracted from our dataset and the average accuracy of each is presented in experimental results section. The remainder of the paper is decomposed as follows: After reviewing previous crowd datasets and introducing the novel properties of our dataset, in Section 2, we introduce our proposed dataset in details. In section 3, we present the methods we applied on our dataset. In section 4, the experimental results are presented and compared. We conclude the paper in section 5.

1.1. Related works

In this part, we list some existing state-of-the-art crowd datasets along with their important characteristics.

UMN [12] is a publicly available dataset including normal and abnormal crowd videos from the University of Minnesota. Each video consists of an initial part of a normal behavior and ends with sequences of an abnormal behavior. Despite the huge amount of abnormal behavior scenarios, only the panic one is included in this dataset which is not realistic in the real world.

UCSD [11] dataset was generated from a stationary camera mounted at an elevation, overlooking pedestrian walkways. The crowd density in the walkways ranges from sparse to very-crowded. Abnormal events are occurred due to either: i) the circulation of non-pedestrian objects in the walkways or ii) abnormal pedestrian motion patterns. As mentioned, the UCSD dataset regards only two definitions for abnormal events which cannot be fully responsible for abnormal behavior detection in crowded scene.

PETS2009 [5] consists of multi-sensor subsets with various crowd activities. The aim of this dataset is to use new or existing systems for i) crowd count and density estimation, ii) tracking of individual(s) within a crowd, and iii) detection of separate flows and specific crowd events, in a real-world environment. For instance, event recognition subset consists of scenarios such as "walking", "running", "evacuation" and "local dispersion". Lack of some other realistic scenarios including fight, fear, abnormal object, etc. is a deficiency for this dataset.

Violent-flows [6] is a dataset of real-world videos downloaded from the web consisting of crowd violence, along with standard benchmark protocols designed to test both violent/nonviolent classification and violence outbreak detection. The problem here is the average length of video clips which is 3.60 seconds and is a limiting parameter for analyzing the scene properly. Also the types of violent behaviors are related to just fighting most of the times in video clips.

Rodriguez's [20] was gathered by crawling and downloading videos from search engines and stock footage websites

Figure 2. (a):Normal and abnormal frames of UCSD dataset. (b):Normal and abnormal frames from the three scenarios of the UMN dataset. (c):Normal and Violent crowd from the Violence-in-crowds dataset. Videos are from different scenes. (d):Normal and abnormal frames of PETS2009 dataset dataset. (e): Frames of crowded scenes from Rodriguez dataset. (f): Normal and abnormal frames of UCF dataset.

behavior class	# frames
Panic	2002
Fight	4423
Congestion	2368
Obstacle	5120
Neutral	29713
Total:	43626

Table 2. Number of frames correspond to each behavior label along with total number of frames available in our dataset.

(e.g., Getty Images and YouTube). In addition to the large amount of crowd videos, the dataset consists of ground truth trajectories for 100 individuals, which were selected randomly from the set of all moving people. This dataset is not open to the public yet.

UCF [22] is acquired from the web (Getty Images, BBC Motion Gallery, YouTube, Thought Equity) and PETS2009, representing crowd and traffic scenes. It is publicly available in the form of image sequences. Unlike [12], it is mainly designed for crowd behaviors recognition, with ground truth labels. This dataset only focused on few crowd flow behaviors such as merging, splitting, circulating, blocking, which cannot fully cover all the crowd abnormal behaviors. Non of aforementioned datasets are not able to reflect human abnormal behavior in the real crowded conditions.

In Fig. 2 few sample frames for state-of-the-art crowd datasets are presented.

2. Proposed Dataset

The introduced dataset consists of 31 video sequences in total or as about 44,000 normal and abnormal video clips. The videos were recorded as 30 frames per second using a fixed video camera elevated at a height, overlooking individual walkways and the video resolution is 554×235. The crowd density in the scene was changeable, ranging from sparse to very crowded. In addition to normal and abnormal behavior scenarios, a few crowd scenes with abnormal objects regarded as threats to the crowd are also recorded make them more realistic. "Motorcycle crossing the crowded scene", "a suspicious backpack left by an individual in the crowd", "a motorcycle which is left between many people", etc. are some examples of such scenarios. In proposed dataset, we have represented five typical types of crowd behaviors. Each scenario topology was sketched in harmony with circumstances usually met in crowding issues. They accord with a transition on a flow of individuals in a free environment (neutral), a crowded scene containing abnormal objects (obstacles), evacuation of individuals from the scene (panic), physical altercation between individuals (Fight), group of people gathering together (congestion). For each behavior type, several videos from two field of views were recorded with different crowd densities changing from sparse to very crowded. All the videos in our dataset start with normal behavior frames and end with abnormal ones. In table 2, some useful details from recorded video clips including number of frames related to predefined behavior classes and total number of frames in our dataset are presented . From the table2, it is clear

Type of behavior	Scenarios
Panic	Suspicious backpack
	Hoodlum attack
	Earthquake
	Sniper attack
	Terrorist firework
Fight	Previous Personal issues between individuals that suddenly meet each other in the crowd
	Intentional or unintentional bad physical contact between two or more people in the crowd
Congestion	Demonstration
	Helping out an individual facing Health problem
	Break up a fight between two or more individuals
Obstacle or Abnormal object	Suspicious backpack
	Motorcycle crossing the crowd
	Motorcycle left in the crowd
	Bag theft with motorcycle
	An individual that fell to the ground for some reasons
Neutral	Moving individuals with almost fixed velocity in random direction
	Two or more people meeting one another

Table 3. Scenarios applied for each type of crowd behavior in our dataset

that similar number of frames are available for different behavior classes in our dataset except for "Neutral" behavior type which has more frames and is more likely in every real crowd. In Table 3 we annotate each behavior type with typical associated scenarios. Although there might be other scenarios for each type of behavior, we tried to use more probable examples in the crowd scene in our dataset. From the scenarios mentioned in Table 3, several videos have been recorded. For each scenario at least two video sequences that correspond to different velocity, field of view and number of individuals have been generated. In each instance, the pedestrian locations and direction of walking are randomly selected. Some sample frames of our dataset labeled with behavior types are presented in Fig. 3. We recorded different videos for each crowd behavior type wherein important parameters like number of individuals, type of scenarios, camera field of view, etc. were not fixed make the dataset more realistic and applicable.

3. Proposed Benchmark

In this part we apply two state-of-the-art methods on our dataset. Dense trajectories [25, 26] (See Fig. 4) which have shown to be efficient for action recognition are applied as first method on our dataset. As second benchmark, we use Histogram of Oriented Tracklet (HOT) [14, 15, 16] descriptor, which is suitable for the task of abnormality detection.

3.1. Low-level Motion Descriptors

A) Dense Trajectory: So far, all the works proposed for Crowd behavior recognition in dense crowded scenes are confronted with many difficulties because of complex motion patterns. To tackle existing challenges we used famous dense trajectory-based method. Dense trajectories are obtained by tracking closely packed feature points, extracted from each frame using multiple spatial scales, on a dense optical flow field using median filter. The length of a trajectory is limited to a fixed number of frames because trajectories tend to drift from their point of initialization. In order to dense coverage assurance and to guarantee the track availability on the dense grid in a frame, the trajectory is eliminated from the tracking process once it exceeds the length L.

Dense trajectories can cover most of the motion features of a video and therefore are able to be used as a tool to capture the apparent motion information and the local features of motions along with local image features. Fig.5(a) shows the dense trajectories computed for different crowded scenarios in our dataset. They are also more robust to irregular sudden motions in videos and capture complex motion patterns more precisely compare with state-of-the-art Kanade-Lucas-Tomasi (KLT) tracker [10] and as a result have taken up more attention in action recognition. We first resize the frames of all recorded videos to 554×235. The first T frames of each video sequence are then selected as the training set, and all the frames are considered as the testing set. Then, we extract dense trajectories using the code presented by [26]. In order to describe extracted dense trajectories we computed state-of-the-art feature descriptors, namely histogram of oriented gradients (HOG) [3], histogram of optical flow (HOF) [9] and motion boundary histogram (MBH) [4] within space-time patches to leverage the motion information in dense trajectories. The size of the patch is $N \times N$ pixels and L frames.

Figure 3. Example of different scenario frames. (a): four sample video frames of neutral scenario. (b): four sample video frames of panic scenario. (c): four sample video frames of fight scenario. (d): four sample video frames of obstacle (abnormal object) scenario. (e): four sample video frames of congestion scenario.

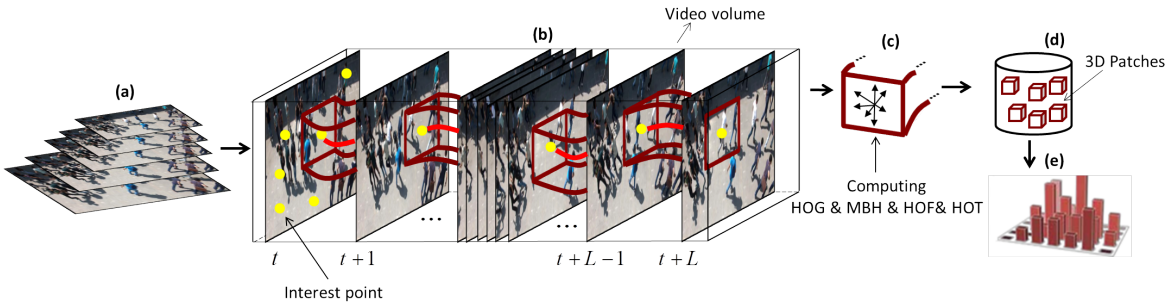

Figure 4. Illustration of our proposed benchmark applied on our dataset. (a): Interest points are sampled densely for multiple spatial scales. (b): For each spatial scale, tracking is performed over L frames. (c): Low-level visual feature descriptors (HOG, HOF, MBH and HOT) are extracted within space-time patches to leverage the motion information. The size of the patches are $N \times N \times L$. (d): A codebook for each descriptor is extracted separately. (e): Histograms of visual words are extracted to be used as a video descriptor.

B) Histogram of Oriented Tracklets: As another benchmark, we used HOT [14, 15, 16] descriptor on our dataset. This descriptor describes each spatio-temporal window in the video volume using motion trajectories represented by a set of tracklets. For this purpose, spatio-temporal cuboids are defined and statistics on the sparse trajectories are collected on the sparse trajectories that intersect them. More in detail, the magnitude and orientation of such intersecting tracklets are encoded in a histogram which is called histogram of oriented tracklets, in short HOT [14, 15, 16]. Fig.5 (b), (c) shows the HOT computed for different crowded scenarios in our dataset.

3.2. Video Representation and Classification

Providing a global feature representation from obtained feature descriptors is the next step. To do so, we employ a bag-of-words paradigm to build histograms for each video sequence. For this purpose, a codebook for each descriptor (HOG, HOF, MBH and HOT) is extracted separately. We fix the number of visual words per descriptor to 1000 which

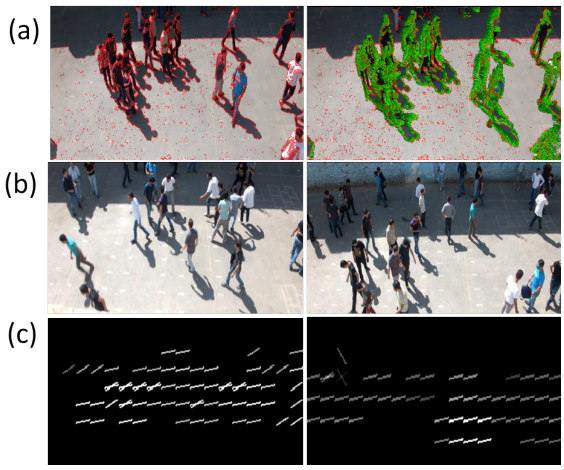

Figure 5. (a) Dense trajectories computed for different crowded scenarios in our dataset. Red marks are the end points of the trajectories.(b),(c) Histogram of Oriented Tracklets computed for two sample crowded scenarios in our dataset

has shown to yield fine results for a wide range of datasets. To restrict the complexity, a subset of 100,000 randomly selected training features is clustered using k-means. Descriptors are allocated to their closest vocabulary word using Euclidean distance. Histograms of visual words which are extracted doing so are used as a video descriptor. For classification of videos, a standard single class support vector (SVM) machines classifier is employed.

4. Experimental results

In this section, the aforementioned benchmarks are used to extract low-level visual features from our dataset. Note that the evaluation protocol is fixed during the experiments. The train and test data are divided in a leave-one-sequence-out fashion. More specifically, for 31 times (equal to number of video sequences) we leave one video clip of a sequence out for test and train data on all the remaining 30. In the evaluation process, the average accuracy both in tables and confusion matrices is used. We separately evaluate HOG, HOF, MBH,Trajectory, Dense trajectory and HOT low-level feature descriptors by ground-truth label information of the behavior and the average accuracy of each is presented in table 4. As can be seen, dense trajectory feature achieved 38.71 % accuracy in crowd behavior abnormality detection and has better performance comparing with other feature descriptors. In Fig.6 and Fig.7, the performance comparison between varied combinations of different types of behavior categories are shown by confusion matrices based on dense trajectory and HOT descriptor, respectively. As can be seen in Fig.6, the "*Panic*" category has the best result of 74.82 % compared to other behavior classes, probably due to solving a simpler task. The most confusion of this category was with "*fight*" which can be justified as the similarity of motion patterns in these two categories (very sharp movements). Also in Fig.7, the "*Panic*" category has the best result of 62.18 % compared to other behavior classes. The most confusion of this category was again with "*fight*" category.

	Our dataset Low-Level Visual Feature
Trajectory	35.30
HOG	38.80
HOF	37.69
MBH	38.53
HOT	38.17
Dense Trajectory	**38.71**

Table 4. Comparison of different feature descriptors (Trajectory, HOG, HOF, MBH, Dense Trajectory and HOT) on Low-Level Visual Feature. We report average accuracy for our dataset.

		Prediction				
		Panic	Fight	Congestion	Obstacle	Neutral
Truth	Panic	74.82%	15.18%	5.64%	3.39%	0.97%
	Fight	24.48%	30.47%	17.18%	18.24%	9.63%
	Congestion	32.17%	18.11%	23.43%	18.91%	7.38%
	Obstacle	9.25%	25.54%	19.02%	27.94%	18.25%
	Neutral	9.40%	16.80%	17.65%	19.27%	36.88%

Figure 6. Confusion matrix for DT [26]

		Prediction				
		Panic	Fight	Congestion	Obstacle	Neutral
Truth	Panic	62.18%	13.57%	12.43%	10.88%	0.94%
	Fight	14.10%	38.27%	17.77%	19.01%	10.85%
	Congestion	29.47%	21.77%	25.67%	15.32%	7.77%
	Obstacle	5.85%	26.59%	24.21%	28.20%	15.15%
	Neutral	8.69%	17.26%	17.78%	19.74%	36.53%

Figure 7. Confusion matrix for HOT descriptor [14, 15, 16]

5. Conclusions

In this paper we presents a novel multi-class crowd dataset, which has around 45,000 video clips all labeled via ground-truth behavior information, with interacting groups of individuals classified into one of five various behaviors. We evaluated the videos of our dataset employing the state-of-the-art feature descriptors and separately evaluate them via ground-truth behavior annotations.

References

[1] Y. Benabbas, N. Ihaddadene, and C. Djeraba. Motion pattern extraction and event detection for automatic visual surveillance. *Journal on Image and Video Processing*, 2011:7, 2011.

[2] R. Berggren. *Simulating crowd behaviour in computer games*. PhD thesis, BSc dissertation. Luleå University of Technology, 2005.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006*, pages 428–441. Springer, 2006.

[5] J. Ferryman, A. Shahrokni, et al. An overview of the pets 2009 challenge. 2009.

[6] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 1–6. IEEE, 2012.

[7] T. Hospedales, S. Gong, and T. Xiang. Video behaviour mining using a dynamic topic model. *International journal of computer vision*, 98(3):303–323, 2012.

[8] J. S. J. Junior, S. Musse, and C. Jung. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, 5(27):66–77, 2010.

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[10] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981.

[11] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE, 2010.

[12] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009.

[13] H. Mousavi, H. K. Galoogahi, A. Perina, and V. Murino. Detecting abnormal behavioral patterns in crowd scenarios. In *Toward Robotic Socially Believable Behaving Systems-Volume II*, pages 185–205. Springer International Publishing, 2016.

[14] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino. Analyzing tracklets for the detection of abnormal crowd behavior. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 148–155. IEEE, 2015.

[15] H. Mousavi, M. Nabi, H. K. Galoogahi, A. Perina, and V. Murino. Abnormality detection with improved histogram of oriented tracklets. In *Image Analysis and ProcessingI-CIAP 2015*, pages 722–732. Springer, 2015.

[16] H. Mousavi, M. Nabi, H. Kiani, A. Perina, and V. Murino. Crowd motion monitoring using tracklet-based commotion measure. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 2354–2358. IEEE, 2015.

[17] M. Moussaïd, D. Helbing, and G. Theraulaz. How simple rules determine pedestrian behavior and crowd disasters. *Proceedings of the National Academy of Sciences*, 108(17):6884–6888, 2011.

[18] M. Nabi, A. Del Bue, and V. Murino. Temporal poselets for collective activity detection and recognition. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 500–507, Dec 2013.

[19] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1389–1396. IEEE, 2009.

[20] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1235–1242. IEEE, 2011.

[21] N. N. A. Sjarif, S. M. Shamsuddin, S. Z. M. Hashim, and S. S. Yuhaniz. Crowd analysis and its applications. In *Software Engineering and Computer Systems*, pages 687–697. Springer, 2011.

[22] B. Solmaz, B. E. Moore, and M. Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):2064–2070, 2012.

[23] H. Su, H. Yang, S. Zheng, Y. Fan, and S. Wei. The large-scale crowd behavior perception based on spatio-temporal viscous fluid field. *Information Forensics and Security, IEEE Transactions on*, 8(10):1575–1589, 2013.

[24] N. Sulman, T. Sanocki, D. Goldgof, and R. Kasturi. How effective is human video surveillance performance. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–3. IEEE, 2008.

[25] B. Wang, M. Ye, X. Li, F. Zhao, and J. Ding. Abnormal crowd behavior detection using high-frequency and spatio-temporal features. *Machine Vision and Applications*, 23(3):501–511, 2012.

[26] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.