Abnormality Detection with Improved Histogram of Oriented Tracklets

Hossein Mousavi[1](✉), Moin Nabi[1], Hamed Kiani Galoogahi[1], Alessandro Perina[1], and Vittorio Murino[1,2]

[1] Pattern Analysis and Computer Vision Department (PAVIS), Istituto Italiano di Tecnologia (IIT), Genova, Italy
{Hossein.Mousavi,Moin.Nabi,Kiani.Galoogahi, Alessandro.Perina,Vittorio.Murino}@iit.it
[2] Dipartimento di Informatica, University of Verona, Verona, Italy

Abstract. Recently the histogram of oriented tracklets (HOT) was shown to be an efficient video representation for abnormality detection and achieved state-of-the-arts on the available datasets. Unlike standard video descriptors that mainly employ low level motion features, e.g. optical flow, the HOT descriptor simultaneously encodes magnitude and orientation of tracklets as a mid-level representation over crowd motions. However, extracting tracklets in HOT suffers from poor salient point initialization and tracking drift in the presence of occlusion. Moreover, count-based HOT histogramming does not properly take into account the motion characteristics of abnormal motions. This paper extends the HOT by addressing these drawbacks introducing an enhanced version of HOT, named Improved HOT. First, we propose to initialize salient points in each frame instead of the first frame, as the HOT does. Second, we replace the naive count-based histogramming by the richer statistics of crowd movement (i.e., motion distribution). The evaluation of the Improved HOT on different datasets, namely UCSD, BEHAVE and UMN, yields compelling results in abnormality detection, by outperforming the original HOT and the state-of-the-art descriptors based on optical flow, dense trajectories and the social force models.

Keywords: Histogram of oriented tracklets · Abnormality detection · Tracklets · Crowd motion analysis

1 Introduction

The study of human behavior has become an active research topic in the areas of human-computer interaction, robot learning, user interface design, intelligent surveillance and crowd analysis. The task of crowd behavior detection refers to identifying the behavioral patterns of individuals involved in a crowd scenario. It is well noted in the sociological literature that a crowd goes beyond a set of individuals that independently display their personal behavioral patterns [1,2]. In other words, the behavior of each individual in a crowd may be influenced

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 722–732, 2015.
DOI: 10.1007/978-3-319-23234-8_66

by "crowd factors" (e.g., dynamics, goal, environment, event, etc.), thus, the individuals behave in a different way than if they were alone.

Based on the above explanation, existing computer vision techniques designed for the detection of individual behavioral patterns are not suitable for modeling and detecting events in crowd scenes. This has encouraged the vision community to design tailored techniques for modeling and understanding behavioral patterns in crowd scenarios. A large portion of recent works is dedicated to model and detect abnormal behaviors in video data. Existing works in the literature are basically different in terms of the type of abnormal behavior (e.g. panic [3], violence [4], escape [5]), types of features (histograms of low level features [6–8], optical flow [9,10], trajectories [11], spatio-temporal features [12,13], etc.), modeling frameworks and learning techniques such as markov model based [10], bayesian models [14], clustering based [15,16], commotion measure [17] and social force models [9].

Recently, a new video descriptor called Histogram of Oriented Tracklets, HOT, is proposed to detect abnormality in crowd scenarios [18]. The HOT descriptor encodes the motion patterns in the form of 2-dimensional histogram utilizing the magnitude and orientation of tracklets. The extensive experiments over abnormality datasets showed the superiority and simplicity of the HOT compared to the state-of-the-art descriptors. The promising performance achieved by HOT is mainly provided by: i) exploiting tracklets as mid-level motion representation, and ii) the capability of HOT descriptor to simultaneously encode the statistics of tracklet's orientation and magnitude in a unified descriptor.

This approach, however, suffers form two major drawbacks. **First**, the most of tracklets are extracted by tracking the salient points initialized in the first frame. Therefore, it is not able to extract new tracklets corresponded to salient points appearing in the next frames. Besides, since the tracklets are generated over long term salient points tracking, there is always the possible danger of drifting in the presence of occlusion. For instance, tracklets corresponded to an individual's hand can be wrongly drifted to another individual's hand due to the occlusion of hand shaking. **Second**, crowd motion statistics are naively encoded by counting the number of tracklets fall into each HOT bin. Such histogramming strategy has shown to be effective for feature description. In HOT, however, we empirically observed that it can drastically degrade the effect of magnitudes belong to abnormal motions. The dilemma here is to efficiently address these disadvantages by proposing Improved HOT.

Contribution. The major contributions of this work are listed as below:

1. We propose to extract tracklets by initializing salient points in each frame (i.e., frame level initialization), instead of the first frame (i.e., video level initialization) applied in HOT [18]. In this strategy, the potential salient points are detected at each frame and then tracked over the next L frames.
2. we propose to construct HOT histograms by exploiting richer statistics of crowd motions. In particular, the magnitude distribution (mean and vari-

ance) of tracklets are exploited in histogramming as opposed to the simple counting in the original HOT.

We extensively conduct a set of experiments over abnormal detection datasets including USCD [19], UMN [9] and Behave [20]. The results demonstrate the superiority of Improved HOT compared to HOT and the state-of-the-arts descriptors. The paper is organized as follows. Section 2 introduces the improved HOT and a short overview of the original HOT. Section 3 presents the experimental results followed by conclusions in Section 4.

2 Improved HOT

In order to have a comprehensive study, we recognize two main components in HOT including tracklet extraction and HOT histogram computation. For each component, we first briefly explain the original HOT and elaborate its drawbacks. Then, we introduce new strategies to improve the existing limitations.

2.1 Tracklet Extraction

The existing approaches for motion representation in crowds can be generally classified into two main categories: local motion based (e.g., optical flows) [9,10,14] and complete trajectories of objects [21,22] based. Both have some limitations. Without tracking objects, the information represented by local motions is limited, which weakens the models power. The crowd behavior recognized from local motions are less accurate, tend to be in short range and may fail in certain scenarios. The other type of approaches assumed that complete trajectories of objects were available and crowd video can be represented using the point trajectories. The accurate long-term observation is hard to be guaranteed due to scene clutter and tracking errors, but can effectively capture the global structure of crowd motions [23].

Tracklets exploited in the HOT [18], however, are mid-level representations between the two extremes discussed above. A tracklet is a fragment of a trajectory and is obtained by a tracker within a short period. Tracklets terminate when ambiguities caused by occlusions and scene clutters arise. They are more conservative and less likely to drift than long trajectories [23].

The tracklet extraction strategy employed in the original HOT is illustrated in Fig. 1(top). In this strategy, called video-level initialization [18], tracklets are initialized using the salient points detected in the first frame of the video, and then tracked until the tracker fails. The main drawback of this strategy is that tracklets are limited to the salient points which were extracted from the first frame or the salient points detected over re-initialization process (which only happens when tracker fails). This means that the new salient points appearing in the subsequent frames will not be fully detected, and thus, not considered for tracklet extraction. Moreover, such long term salient point tracking may lead to "drifting" problem. The drifting problem mainly occurs if a salient point miss-detected/tracked in presence of two particles occlusion. Such occlusions are not avoidable in real world crowd scenarios.

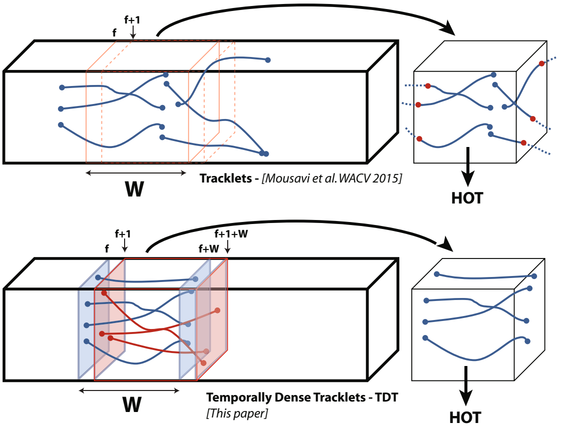

f
f+1
W
Tracklets - [Mousavi et al. WACV 2015]
HOT
f
f+1
f+W
f+1+W
W
Temporally Dense Tracklets - TDT
[This paper]
HOT

Fig. 1. Top: Video level tracklet initialization in HOT [18]. Bottom: Frame level tracklet extraction which is called Temporally Dense Tracklets.

We, on the other hand, proposed to re-initialize salient points in each single frame of the video and track the points over W frames, we called this Temporally Dense Tracklets (TDT) (Fig. 1(bottom)). This strategy is not limited to the points detected at the first frame and is capable of detecting all possible salient points over a given video. In other words, no matter how long is the captured video, this strategy is able to detect the salient points of all the appearing objects/individuals over the time. This results in producing a large pool of tracklets which can summarize the motion-patterns observed in the scene in each frame. we reset interest points in each frame and track it over W frames. In video-based tracklet extraction in HOT, on the other hand, an initial set of points are detected at the first frame of the video and tracked for the entire length of the tracklet.

2.2 Histogram of Tracklets Computation

The process of HOT computation explained in [18] starts by splitting the video in spatio-temporal cuboids. The magnitude ($M^{i,s}$) and orientation ($\theta^{i,s}$) of tracklet i passing from cuboid s are computed as:

$$M^{i,s} = \max_{t \in T} \left\{ m_t^{(i,s)} \right\} \qquad\qquad \theta^{i,s} = \arctan \frac{(y_{end}^{i,s} - y_{begin}^{i,s})}{(x_{end}^{i,s} - x_{begin}^{i,s})} \tag{1}$$

where (x_t, y_t) indicates the two-dimensional coordinates of the t^{th} point of the tracklet i and T indicates the length of each tracklet. m_t is the magnitude of the t^{th} point computed as $m_t = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2}$. $(x_{begin}^{i,s}, y_{begin}^{i,s})$ and $(x_{end}^{i,s}, y_{end}^{i,s})$ respectively show the entry and exit points of tracklet i in/from cuboid s. More details can be found in [18].

The magnitudes and orientations of all tracklets across cuboid s are independently quantized in O orientations and M magnitudes bins. The bins of the 2D-histogram $H_{o,m}^{s,f}$ are finally populated by counting the occurrence of the magnitude-orientation pairs in cuboid s. The frame that the HOT is computed for is indexed by f [18].

The major problem of such histogramming is, however, ignoring the magnitude characteristics of tracklets and only take into account the number of occurrences (by counting). Due to the fact that tracklets belong to abnormal motions exhibit strong magnitudes, simple histogramming degrades the weight of magnitudes belong to abnormal motions.

We differently encode the tracklets in each bin via motion magnitude distribution (mean and variance). In fact, the new histogramming technique, referred to as *weighted histogramming*, preserves the magnitude strength (mean) and the commotion (variance) of motion patterns in each HOT bin.

Given a set of J magnitude-orientation pairs $\{(\theta^{j,s}, M^{j,s})\}_{j=1}^{J}$ fall in the orientation bin o and magnitude bin m, the weighted histogramming returns two 2D histograms called mean-HOT and variance-HOT computed as:

$$mH_{o,m}^{s,f} = \frac{1}{J}\sum_{j=1}^{J} M^{j,s} \qquad vH_{o,m}^{s,f} = \frac{1}{J}\sum_{j=1}^{J}(M^{j,s} - mH_{o,m}^{s})^2 \tag{2}$$

where $mH_{o,m}^{s,f}$ and $vH_{o,m}^{s,f}$ respectively states the mean-HOT and variance-HOT corresponded to the cuboid s at frame f (following the original HOT we compute the Improved HOT per frame).

3 Abnormality Detection

Following the original HOT [18], we applied two approaches to compute the mean-HOT and variance-HOT per frame namely Fully bag of words (BW) and Per-frame, Per-sector (FS). Similarly, we employed the latent Dirichlet allocations (LDA) generative model for learning and classification.

Given a set of two-dimensional mean-HOT $mH_{om}^{s,f}$ and variance-HOT $vH_{o,m}^{s,f}$ for all cuboids s temporary centered at frame $f = 1, \dots, F$, we construct the LDA training corpus $\mathcal{D}$ based on two different detection strategies:

Fully bag of words (BW). In the first case, mean-HOTs and variance-HOTs are summed across spatial sectors:

$$(mD)^f = \sum_s (mH)_{o,m}^{s,f} \quad \text{and} \quad (vD)^f = \sum_s (vH)_{o,m}^{s,f} \tag{3}$$

The LDA training corpus $\mathcal{D}$ is then constructed by concatenating the vectorized $(mD)^f$ and $(vD)^f$ at each frame as $\mathcal{D} = \{[(mD)^f|(vD)^f]\}_{f=1}^{F}$, where the operator | concatenates two vectors.

Per-frame, Per-sector (FS). In the second case, mean-HOTs and variance-HOTs from all the different sectors of a frame are concatenated in a single descriptor to preserve the spatial information of each frame:

$$\begin{aligned} (mD)^f &= \left\{(mH)^{1,f}_{o,m}|(mH)^{2,f}_{o,m}|\dots|(mH)^{S,f}_{o,m}\right\} \\ (vD)^f &= \left\{(vH)^{1,f}_{o,m}|(vH)^{2,f}_{o,m}|\dots|(vH)^{S,f}_{o,m}\right\} \end{aligned} \tag{4}$$

Similarly, the LDA training corpus $\mathcal{D}$ is constructed by concatenating the vectorized $(mD)^f$ and $(vD)^f$ at each frame as $\mathcal{D} = \{[(mD)^f|(vD)^f]\}_{f=1}^{F}$. When training/testing data of normal and abnormal actions is available, the corpus of both positive(normal) and negative(abnormal) clips are constructed and fed into a linear SVM for learning and classification.

4 Experimental Evaluation

We compare the Improved HOT (iHOT) descriptor with the original HOT state-of-the-art descriptors in the literature, mainly the mixtures of dynamic textures framework [19] and leading optical flows based approaches [4,11,24,25]. To have a more comprehensive investigation, we conducted two different experiments. First, we evaluate the improvement provided by only TDT extraction strategy over the UMN and BEHAVE datasets. In the second experiment, we extensively evaluate the full framework of iHOT (namely, TDT tracklet extraction + weighted histogramming) over the UCSD dataset.

4.1 Crowd Datasets

Three publicly available datasets are employed for the evaluation, including USCD [19], UMN [9] and BEHAVE [20].

UCSD Dataset[1] The dataset contains two smaller subsets corresponded to two different scenes. The first, denoted by "`ped1`" contains clips of 158×238 pixels, which depict groups of people walking toward and away from the camera, with a certain degree of perspective distortion. The second, denoted by "`ped2`" has spatial resolution of 240×360 pixels and depicts a scene where most pedestrians move horizontally. The video footage of each scene is sliced into clips of 120-200 frames. We only considered anomaly at the frame level for this dataset.

BEHAVE Dataset[2] consists of a set of complex group activities including *meeting, splitting up, standing, walking together, ignoring each other, escaping, fighting* and *running.* Following [18], the *fighting* activity is selected as abnormalities (50 clips) and the rest as normal activities (271 clips).

UMN Dataset[3] includes 11 different scenarios of a panic and normal situations in three different indoor and outdoor scenes.

[1] Available at http://www.svcl.ucsd.edu/projects/anomaly/

[2] Available at http://groups.inf.ed.ac.uk/vision/behavedata/interactoins/

[3] http://mha.cs.umn.edu/movies/crowdactivity-all.avi

Table 1. AUC on the UMN dataset.

Dataset	iHOT-TDT	HOT [18]	SFM [9]	SR [26]	OF [9]	CI [27]
scene-1	**0.998**	0.993	0.990	0.995	0.964	n/a
scene-2	**0.991**	0.984	0.949	0.975	0.906	n/a
scene-3	**0.998**	0.991	0.989	0.964	0.967	n/a
all scenes	**0.994**	0.991	0.960	0.978	0.840	0.990

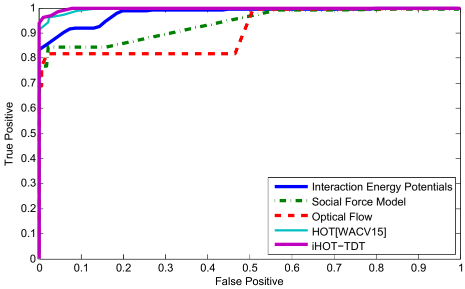


Fig. 2. ROC curve on BEHAVE dataset.

4.2 Evaluating TDT Tracklet Extraction

In this experiment, we evaluate the Improved HOT with TDT tracklet extraction (iHOT-TDT) comparing with the original HOT (video level initialization) and the state-of-the-arts on the UMN and BEHAVE datasets. The parameters are fixed to trackelt length $W = 11$, magnitude bins $M = 16$ and orientation bins $O = 8$. The classification strategy for the original HOT and the Improved HOT is limited to fully bag of words (BW).

Evaluation on UMN Dataset. In this experiment, we compared the iHOT-TDT with HOT [18], social force model (SFM) [9], sparse reconstruction (SP) [26], optical flow (OF) [9], Chaotic Invariants(CI) [27] following the standard evaluation of [9]. To have a finer evaluation, we deployed a protocol by consideration of UMN three scenes separately. We found this protocol so helpful to analyses the effect of proposed descriptor in each scene individually. The results on each scene (*scene-1*, *scene-2*, *scene-3*) and the whole dataset (*all scenes*) are reported in Table 1 in terms of AUC (Area Under the ROC Curve). The result demonstrates the superiority of our approach on this dataset for both scene-based and all scenes evaluations.

Evaluation on Behave Dataset. This experiment compares the iHOT-TDT with the optical flow based method, social force model [9] and interaction energy potential [11]. Following settings in [11], we used half of normal and abnormal videos for training and the rest for testing. We used BW classification strategy along with linear kernel SVM for frame level classification. The results are reported by the means of ROC as shown in Fig 2.

Table 2. Equal Error Rates on UCSD dataset using standard testing protocol. The results of the previous approaches are borrowed from [19]. The results of our approach and the original HOT are the best performance obtained at the parameter tuning experiment.

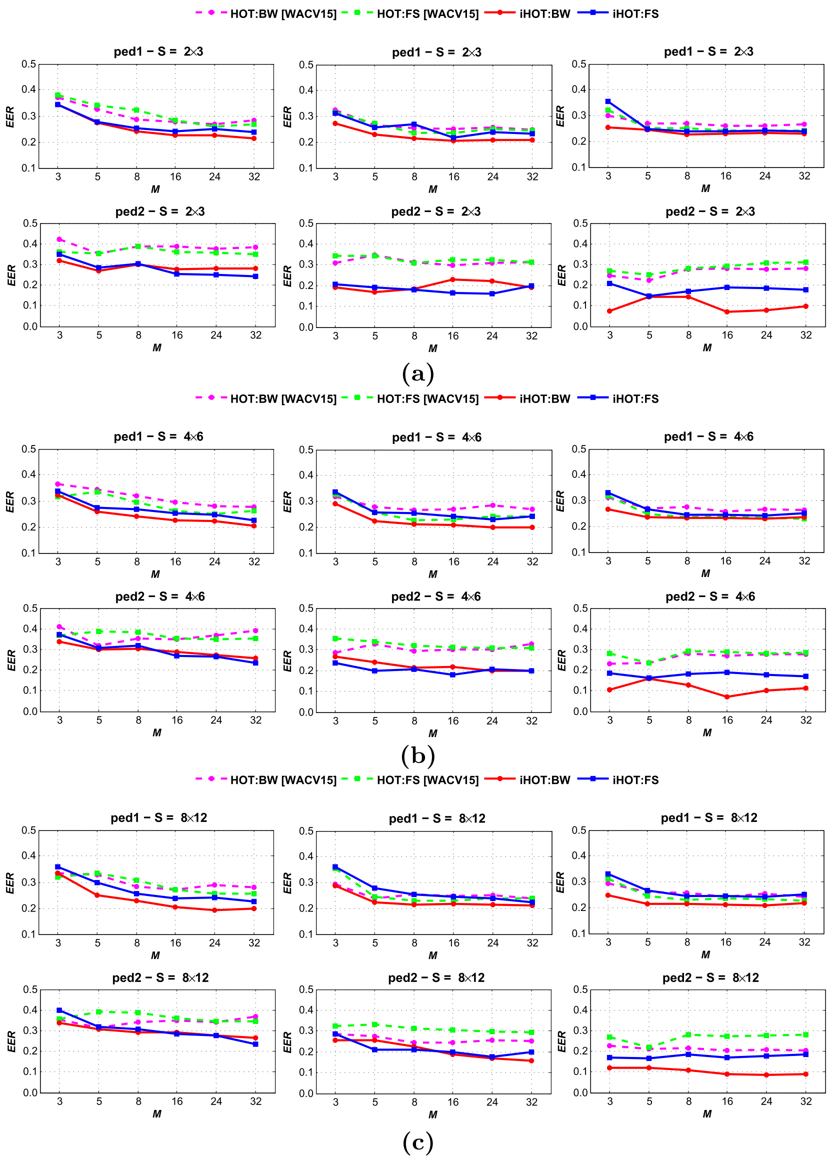
`ped1`		`ped2`	
Method	EER	Method	EER
MDT [19,29]	22.9%	MDT [19,29]	27.9%
SFM [9]	36.5%	SFM [9]	35.0%
LMH [24]	38.9%	LMH [24]	45.8%
HOT: BW [18]	23.84%	HOT: BW [18]	20.42%
HOT: FS [18]	22.53%	HOT: FS [18]	21.84%
iHOT: BW	**19.37**%	iHOT: BW	**8.59**%
iHOT: FS	22.27%	iHOT: FS	16.5%

4.3 Complete iHOT: TDT and Weighted Histogramming

We evaluate different parameter settings of the complete iHOT including spatial tessellation of the frame S, length of tracklets W and the quantization bins O and M. Following [18], we quantized tracklets orientation in $O = 8$ uniform bins [18]. We varied the temporal window of $W = \{5, 11, 21\}$ frames setting the tracklet length to W. Moreover, we varied the number of quantization levels for magnitude $M \in \{3, 5, 8, 16, 24, 32\}$. we considered three different spatial tessellations, called as *coarse* $S = 2 \times 3$, *medium* $S = 4 \times 6$ and *fine* $S = 8 \times 12$. The LDA topics number is fixed to $Z = 30$. This experiment was conducted on the UCSD dataset, `ped1` and `ped2`, comparing our approach with the original HOT in [18] using two different classification sensations: *Fully bag of words(BW)* and *Per-frame, Per-sector (FS)*. For our method, we considered complete iHOT (TDT + weighted histogramming).

The LDA likelihood [28] of the test frames was used to compute the EERs for our approach and the original HOT [18]. Results (EER, the smaller the better) for `ped1` and `ped2` are reported in Fig. 3 showing the robustness of both the original HOT and the Improved HOT. However, the EERs obtained by the Improved HOT for all the parameter combinations are obviously lower than those of the original HOT.

Table.2 compares the Improved HOT with the HOT and the state-of-the-arts descriptors. EERs of competitors are taken from [19] where the authors reported *best* results across all the model-method configurations. Despite such comparison cannot statistically highlight a clear winner, we limit ourselves to acknowledge how the new tracklet extraction strategy (TDT) and the weighted histogramming improves the performance of the original HOT [18] and, surely, outperforms the prior leading methods in the literature. Particularly, our improvement achieved superior performance than the original HOT for both classification strategies, BW and FS, on `ped1` and `ped2`. Please note that the EERs of the original HOT reported in Table 2 are slightly different than those obtained in the reference paper [18]. Since, here we fixed the LDA topics $Z = 30$, while, the results in [18] were the best achieved by varying $Z \in \{2, 4, 6, ..., 80\}$.



(a)

(b)

(c)

Fig. 3. Results for `ped1` and `ped2` varying the number of magnitude bins, tracklet length and spatial tessellation. (a) coarse tessellation, (b) medium tessellation, (c) fine tessellation. The first, second and third column respectively corresponded to the tracklet length of 5, 11 and 21.

5 Conclusion

In this paper, we introduced a modified version of histogram of oriented tracklets (HOT) descriptor for the task of abnormality detection in crowd scenes. We discussed and empirically showed that video level tracklet extraction employed by the original HOT which include poor initialization of salient points and tracking drift. To address this drawback we proposed Temporally Dense Tracklets to initialize the salient points in each frame. Moreover, we analyzed that the counting-based naive histogramming in the HOT is not capable of capturing statistics of abnormal motions. We proposed weighted histogramming to deal with this disadvantage by exploiting the distribution of crowd motions (mean and variance). The enhanced version of HOT is called Improved HOT (iHOT). The evaluations demonstrated the superiority of the Improved HOT compared to the original HOT and the prior video descriptors.

References

1. Wijermans, N., Jorna, R., Jager, W., Vliet, T.v.: Modelling crowd dynamics, influence factors related to the probability of a riot. In: Proceedings of the Fourth European Social Simualtion Association Conference (ESSA), Toulouse University of Social Sciences (2007)
2. Junior, S.J., et al.: Crowd analysis using computer vision techniques. IEEE Signal Processing Magazine **27**(5), 66–77 (2010)
3. Haque, M., Murshed, M.: Panic-driven event detection from surveillance video stream without track and motion features. In: 2010 IEEE International Conference on Multimedia and Expo (ICME), pp. 173–178. IEEE (2010)
4. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: CVPRW, pp. 1–6. IEEE (2012)
5. Wu, S., Wong, H.S., Yu, Z.: A bayesian model for crowd escape behavior detection. IEEE Transactions on Circuits and Systems for Video Technology **24**(1), 85–98 (2014)
6. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2004, vol. 2, p. II-819. IEEE (2004)
7. Xiang, T., Gong, S.: Video behavior profiling for anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(5), 893–908 (2008)
8. Reddy, V., Sanderson, C., Lovell, B.C.: Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 55–61. IEEE (2011)
9. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 935–942. IEEE (2009)
10. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 2921–2928. IEEE (2009)
11. Cui, X., Liu, Q., Gao, M., Metaxas, D.N.: Abnormal detection using interaction energy potentials. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3161–3167. IEEE (2011)

12. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 1446–1453. IEEE (2009)
13. Boiman, O., Irani, M.: Detecting irregularities in images and in video. International Journal of Computer Vision **74**(1), 17–31 (2007)
14. Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(3), 539–555 (2009)
15. Fu, Z., Hu, W., Tan, T.: Similarity based vehicle trajectory clustering and anomaly detection. In: IEEE International Conference on Image Processing. ICIP 2005, vol. 2, p. II-602. IEEE (2005)
16. Alvar, M., Torsello, A., Sanchez-Miralles, A., Armingol, J.M.: Abnormal behavior detection using dominant sets. Machine Vision and Applications, 1–18 (2014)
17. Mousavi, H., Nabi, M., Kiani, H., Perina, A., Murino, V.: Crowd motion monitoring using tracklet-based commotion measure. In: IEEE International Conference on Image Processing (ICIP). IEEE (2015)
18. Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., Murino, V.: Analyzing tracklets for the detection of abnormal crowd behavior. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 148–155. IEEE (2015)
19. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981. IEEE (2010)
20. Blunsden, S., Fisher, R.: The behave video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA **4**, 1–12 (2010)
21. Makris, D., Ellis, T.: Learning semantic scene models from observing activity in visual surveillance. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **35**(3), 397–408 (2005)
22. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(9), 1450–1464 (2006)
23. Zhou, B., Wang, X., Tang, X.: Random field topic model for semantic region analysis in crowded scenes from tracklets. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3441–3448. IEEE (2011)
24. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(3), 555–560 (2008)
25. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: CVPR, pp. 2921–2928. IEEE (2009)
26. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3449–3456. IEEE (2011)
27. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2054–2060. IEEE (2010)
28. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of machine Learning research **3**, 993–1022 (2003)
29. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(1), 18–32 (2014)