# Abnormal Event Detection and Localization in a Video based on Similarity Structure

**Mahmood Fathy[1], Mohammad Sabokrou**
**Mojtaba Hosseini[3]**

Receive :2016/04/20 Accepted: 2016/07/26

**Abstract**

**This paper introduces a method for abnormal event detection in video. The video is divided into a set of cubic patches. A new descriptor for representing the video patches is proposed. This descriptor is created based on the structure similarity between a patch and nine neighboring patches of it. All training normal patches in respect to the proposed descriptor are represented and then modeled using a Gaussian distribution as the *reference model*. In test phase, those patches which are not fitted to the reference model are labeled as anomaly. We have evaluated the proposed method on two UCSD[1] and UMN[2] popular standard benchmarks. The performance of the presented method is similar to state-of-the-art methods and also is very fast.**

**Key-words: Anomaly, Video Processing, Similarity Structure, Gaussian distribution.**

## I. Introduction

Todays, the surveillance cameras are widely exploited for abnormal event detection. The abnormal events in video usually refers those ones which are rarely occurred, i. e. we look for an unknown event. Consequently, detecting abnormal event is completely a cumbersome task.

As there are not any abnormal samples in training video, the researcher have modeled the normal events which are occurred with high frequency as the *reference model*. In testing phase those samples which have a high diversion from this model are considered to be abnormal. The elementary proposed methods for abnormal event detection were based on modeling the trajectory of normal objects. These methods have a high complexity especially in crowded scenes, also they cannot handle occlusion problem efficiently. To overcome these challenges, the spatial-temporal features such as HoG and HoF are considered. These method was robust on occlusion problem, but the localization of abnormal events and speed of proposed method is two ongoing challenges. Appearing an object with unusual shape or motion can be considered as an abnormal object (event). Consequently, in this paper to have both shape and motion information the video is considered as a set of cubic patches. (i. e. k, w×h patches from a same location but different continues frames are extracted and then integrated as a 3D patch). For being fast and accurate, firstly the background of the frames are detected and removed, then the relation of those patches which are contained foreground pixels, with their adjacent are modeled as a *normal reference model*. In testing phase those patches which don't follow this model are detected as abnormal. An example of a normal and abnormal events is shown in Figure 1.

Figure 1. Example of abnormal and normal events. The panic escape is an abnormal event *(left):* Normal *(right)* Abnormal. Consider the relation of selected patch with its adjacent patches on normal and panic mode.

The rest of this paper is organized as follows. In section II the related works are surveyed, an explanation about the proposed method is provided in section III. In section IV the performance of our method is evaluated, section V concludes the paper.

## II. Related works

Early work in the subject area focused on modeling of object trajectories; see [1-9].

An object is labelled as being an anomaly if it does not follow learned normal trajectories. The main weaknesses of these methods are (1) that they are not robust with respect to occlusions, (2) and they are very complex for crowded scenes.

[1] Iran University of Science and Technology mahfathy@iust.ac.ir
[2] Malek Ashtar University of Technology sabokro@gmail.com
[3]Malek Asthar university of Technology mojtabahoseini@aut.ac.ir

For avoiding these weaknesses, researchers proposed some methods using spatial-temporal low-level features, such as optical flow or gradients. Zhang et al. [10] model the normal pattern of a video with a Markov Random Field (MRF) in respect to a number of features, such as rarity, unexpectedness, or relevance. Boiman and Irani [11] consider an event as being an anomaly if its reconstruction is impossible by using previous observations only. Adam et al. [12] use an exponential distribution for modeling the histograms of optical flow in local regions. Mahadevan et al. [13] use a Mixture of Dynamic Textures (MDT) for representing the video and fit a Gaussian mixture model to features. In [14], the MDT [13] is extended and explained with more details. Kim and Grauman [15] exploit a Mixture of Probabilistic PCA (MPPCA) model for representing local optical flow patterns. They also use an MRF for learning the normal patterns. A method based on the motion properties of pixels for behavior modeling is proposed by Benezeth et al.[16]. They described the video by learning a co-occurrence matrix for normal events across space-time. In [17], a Gaussian model is fitted to spatio-temporal gradient features, and a Hidden Markov Model (HMM) is exploited for detecting the abnormal events. The Social Force (SF) is introduced by Mehran et al [18] as an efficient technique for abnormal motion modeling for crowds. In [19] a method is proposed which is based on spatial-temporal oriented energy filtering. Cong et al [20] construct an over-complete normal basis set from normal data; if reconstructing a patch with this basis set is not possible then it is considered to be an anomaly.

In [21] a scene parsing approach is presented. All object hypotheses for the foreground of a frame are explained by normal training. Those hypotheses which cannot be explained by normal training are considered to be anomaly. The authors of [22] propose a method based on clustering the test data by using optic-flow features. [23] introduced an approach based on a cut/max-flow algorithm for segmenting the crowd motion. If a flow does not follow the regular motion model then it is considered as being an anomaly. Lu et al [24] propose a very fast (140-150 fps) anomaly detection method. Their method is based on sparse representation. In [25] an extension of the Bag of Video words (BOV) approach is used. In [26], a context-aware anomaly detection algorithm is proposed where the authors represent the video using motions and the context of videos. In [27], a method for modeling both motion and shape with respect to a descriptor (named "motion context") is proposed; they consider anomaly detection as a matching problem. Roshkhari et al [28] introduce a method for learning the events of a video by using the construction of a hierarchical codebook for dominant events in a video. Ullah et al [29] learn an MLP neural network using trained particles to extract the video behavior. A Gaussian Mixture Model (GMM) is exploited for learning the behavior of particles using extracted features. Also, in [30], a MLP neural network for extracting the corner features from normal training samples is proposed; authors also label the test samples using that MLP. Authors of [31] extract corner features and analyze those features based on their properties of motion by an enthalpy model; a random forest with corner features for detecting anomaly samples is exploited. Xu et al. [32] propose a unified anomaly energy function based on a hierarchical activity-pattern discovery for detecting anomalies.

Work reported in [33-34] model normal events based on a set of representative features which are learned on auto-encoders [35]; these authors use a one-class classifier for detecting anomalies as being outliers compared to the target (i.e. normal) class. In [36], the Histogram of Oriented Tracklets (HOT) is used for video representation and anomaly detection; these authors also introduce a new strategy for improving HOT. Yun et al [37] introduce an informative Structural Context Descriptor (SCD) to represent a crowd individually; in their work, a (spatial-temporal) SCD variation of a crowd is analyzed to localize the anomaly region.

A hierarchical framework for local and global anomaly detection is proposed in [38]. Normal interactions are extracted by finding frequent geometric relationships between sparse interest points; authors model the normal interaction template by Gaussian process regression. Xiao et al [29] exploit Sparse Semi-nonnegative Matrix Factorization (SSMF) for learning the local pattern of pixels. Their method learns a probability model

by using local patterns of pixels for considering both the spatial and temporal context. Their method is totally unsupervised. Anomalies are detected by the learned model. In [40] an efficient method for representing human activities in video data with respect to motion characteristics is introduced and named as motion influence map; authors label those blocks of a frame as being an anomaly which have a low occurrence. [41] Proposes an unsupervised framework for detecting the anomalies based on learning global activity patterns and local salient behavior patterns via clustering and sparse coding.

## III. Proposed Method

The video consider as a set of cubic patches size of 40×40×5. A high percent of extracted patches from a video are completely related to the background where these patches have not any role in abnormal event procedure. So, a simple background subtraction method is exploited on frames of video, and those patches which are extracted from background all labeled as normal and are ignored from the next processing. The patches which are extracted from foreground of every frame are represented by the proposed descriptor (see section IV-II). A Gaussian distribution are fitted on all normal represented patches. We called this Gaussian distribution as the *reference model*. In test component, first all background patches of the test frame are labeled as normal. The distance of remaining patches (i. e. foreground patches) form the *reference model* is computed. Those ones which are far from the *reference model* are considered to be anomaly. Figure 2 shows the overall scheme of our method. Also figure 3 shows the flowchart of our method. The details of *background subtraction*, creating the *reference model* and also *detecting abnormal* patches are provided in next subsection of this section.

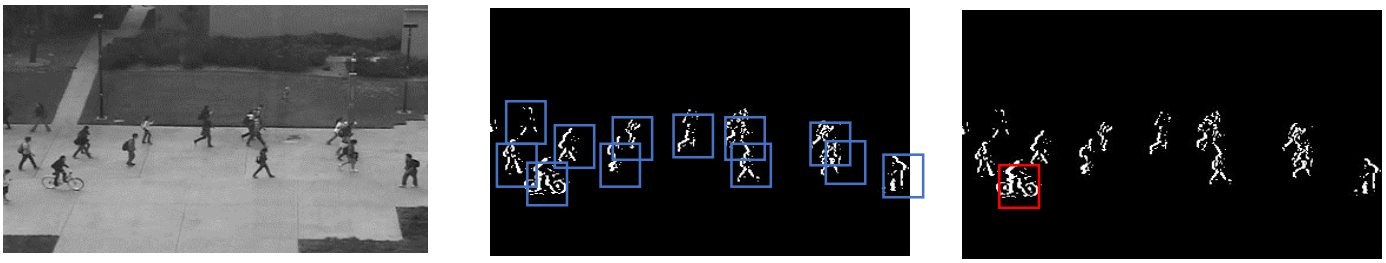

Figure 2. Overall schemes of our method. (*Left*) Original frame (*middle*) Applying a subtraction method for finding those patches which contain an object (here pedestrian) (*right*) that patch which are not fitted to reference model is indicated as anomaly.

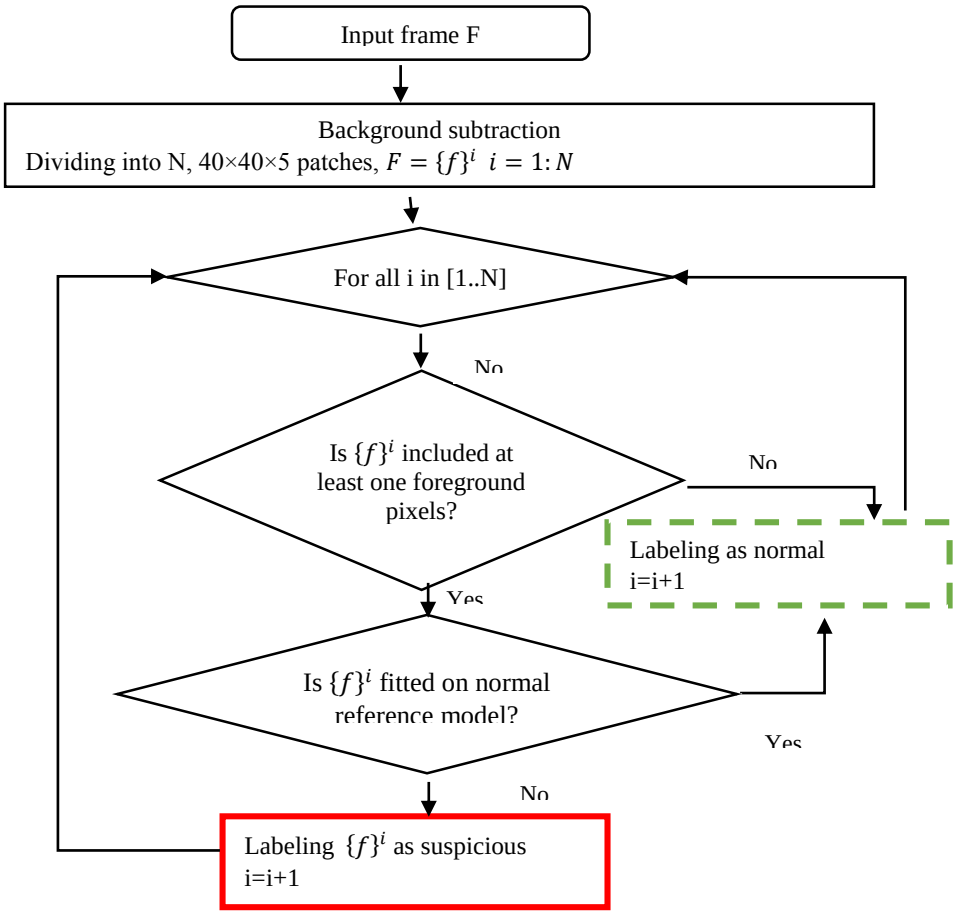


Figure 2. Flowchart of the proposed method

### IV-I. Background subtraction

The videos are captured by a fixed camera. So, for finding the background of $t^{th}$frame $I^t$ $(t > 10)$, the average of 10 previous frames is computed as B. Consequently the B is background model context. To compute the foreground, B is subtracted from $F^t$ i. e. F=$I^t - B$, where F is foreground. We find that also there are some noises in F. These noises are removed using an image-thresholing. The result of this thresholing is a binary image where the values of foreground pixel is 1, otherwise 0.

### IV-II. Descriptor: Feature extractor

The patches are described based on their relation with their adjacent patches in respect to both spatial and temporal context. 8 patches around and one patch behind of every patch is considered as it's adjacent. The Spatial-Temporal Structure Similarity (ST-SSIM) between a patch with 9 its neighboring patches are considered as the description of centroid patches. We have proposed the ST-SSIM based on SSIM measure [42-43] which is a tools for computing the structure similarity between two images. We have extended SSIM for calculating the structure similarity for two cubic patches. Suppose that the K$\in R^{W\times H\times T}$ and L$\in R^{W\times H\times T}$ be two cubic patches. Actually K (and L) is, several 2D patches with size of $W \times H$ which are extracted from same location of T continues frames and then is integrated as one cubic patch. Consequently K= $\{K_1, K_2, \dots, K_T\}$ and L= $\{L_1, L_2, \dots, L_T\}$ where $\{K_i\}^{i=1:T} \in R^{W\times H}$ and $\{L_i\}^{i=1:T}$. So, the ST-SSIM (K, L) can be computed using Equation 1.

$$\text{ST-SSIM}= \sum_{i=1}^{T} \beta^i \times SSIM(\{K_i, L_i\}) \quad (1)$$

Where SSIM is a popular measure and more details about it is provided in [42-43]. The $\beta \in [0.50, 0.99]$ is a parameter for adjusting the importance of occurring time of every 2D patches in a cubic patch. For example when we analyze the t[th] frame, the T previous patch also must be considered. (To be cubic). So, the $\beta$ allow us to decrease the effect of previous frames.

The patches based on ST-SSIM are represented to a feature vector. Suppose K be a patch, and $\{K\}^{i=1:9}$ be the neighboring patches of it, see Figure 4. The representation of K is $\{\alpha_1, \alpha_2, \dots, \alpha_9\}$ where $\alpha_i$ =ST-SSIM$\{K, K^i\}$.

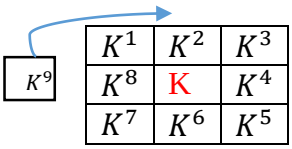


Figure 4. The adjacent patches of K. 8 patches are around of K and one is behind of K.

### IV-III. Creating the reference model

All training videos which contain normal events are divided into a set of cubic patches. Those patches which are related to background are ignored. The remaining patches are represented to 9 features as explained in previous sub-section. Then, a Gaussian distribution is fitted on all represented patches as G. We model the abnormal event detection as a one-class classification problem similar to [33].

### IV-IV. Abnormal event detection

In testing phase, also all background patches are considered to be normal as an early detection. But foreground patches must carefully be evaluated. Suppose K is a patch, it is represented to 9 features. The Mahalanobis distance between this feature vector and reference model G as D(K, G) be calculated. If D(K, G) be more than a pre-determined threshold such Ω, K is abnormal. Equation 2 indicates a summary of abnormal event detection.

$$f(K) = \begin{cases} Abnormal & if\ D(K,G) > \Omega \\ Normal & otherwise \end{cases} \quad (2)$$

### Experimental results

We empirically demonstrate that our method can detect suspicious event in video which is captured from surveillance system. All experiment are done using a PC with 3.5 GHz CPU and 8G RAM in MATLAB 2012a.

**Experimental settings:** In both training and testing the size of extracted patches are 40×40×5. The β in Equation (1) is selected to be 0.9. The Ω threshold also is experimentally determined.

**Datasets:** The proposed method is evaluated on UCSD Ped2 and UMN benchmarks. The UCSD

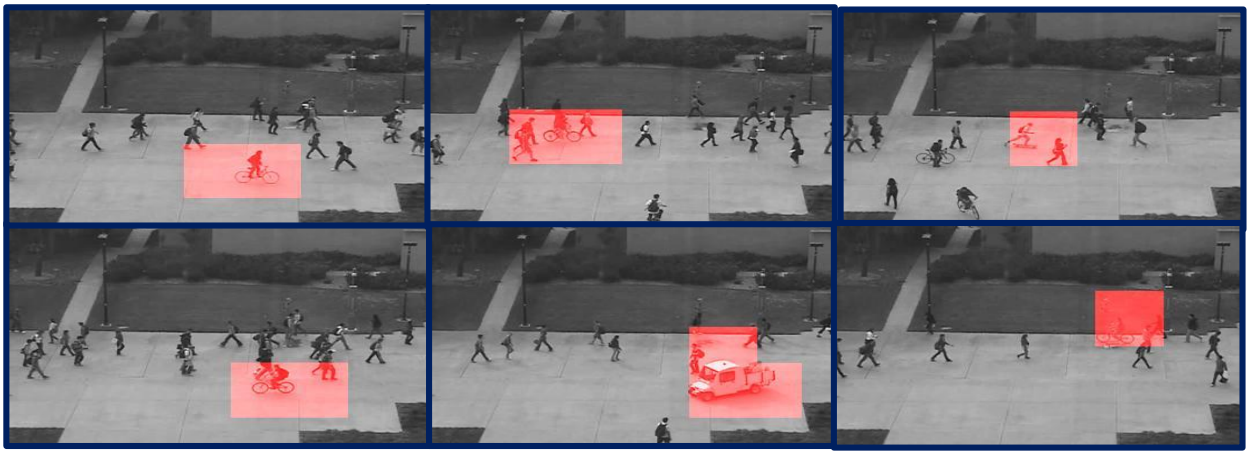

Figure 5. Qualitive results on USCD ped2

dataset includes two subsets, ped1 and ped2, that are from two different outdoor scenes. Both are recorded with a static camera at 10 fps, with the resolutions 158×234 and 240×360, respectively. The dominant mobile objects in these scenes are pedestrians. Therefore, any object (e.g., a car, skateboarder, wheelchair, or bicycle) is considered as being an anomaly. The UMN is related to a group of people are walking in an area, suddenly all people run away (escape); the escape is considered to be an unusual.

**Performance Evaluations:** Figure 5 shows the output of our method on 6 different UCSD Ped2 dataset. It confirms that our method has a good performance in both detecting and localizing unusual events. Also, a quantitative evaluation is provided is this section.

For analyzing the performance of the proposed method two measure is defined in [13] where defined as below:

**Frame-level:** if at least one pixels of a frame detected as unusual, that frame is labeled to be unusual.

**Pixel-level**: If 40% of abnormal ground-trust pixels are detected as an unusual with the algorithm, the frame is considered to be anomaly. This measure is related to localization accuracy.

We comprise the performance of our method in respect to Equal Error Rate (EER) of frame-level and Detection Rate (DR) of pixel-level.

The frame-level of our method in compassion of state-of-the-art method in Table I is provided. We outperforms most of other considered method. The proposed method by Li et. al [14] and saborkou et al. [33] has a bit higher performance than our method (by 3% and 1.5%, respectively)

Table I. Frame-level comparison

| Method | Frame-level EER |
|---|---|
| SF [18] | 42% |
| MPPCA [15] | 30% |
| SF+MPPCA [15] | 36% |
| Adam et.al [12] | 42% |
| MDT[13] | 25% |
| Li et al [14] | 18% |
| Saborkou et. al. [33] | 19.5% |
| Ours | 21% |

Also, Table II shows the detection rate of our method in comparison with state-of-the-art methods. We are better than the best state-of-the-art method by 2%.

Table II. Pixel-level comparison

| Method | Frame-level DR |
|---|---|
| SF [18] | 21% |
| MPPCA [15] | 18% |
| SF+MPPCA [15] | 28% |
| Adam et.al [12] | 24% |
| MDT [13] | 45% |
| Li et al [14] | 63.4% |
| Saborkou et. al.[33] | 76% |
| Ours | 78% |

Also, speed of the proposed method is 30 FPS, consequently, it can run as a real-time application.

The performance of our method on UMN dataset in comparison with state-of-the-art is provided in Table III. The results confirm that the proposed method is comparable to other considered methods. The Area Under Curve (AUC) of ROC and EER (both for frame-level) are used for evaluating the performance of the methods on this dataset.

Table III. Frame-level comparison on UMN dataset

| Method | EER/AUC (%) |
|---|---|
| Chaotic invariants [2] | 5.3/99.4 |
| SF [18] | 12.6/94.9 |
| Sparse [20] | 2.8/99.6 |
| Ours | 2.6/99.5 |

## IV. Conclusion

This paper proposed a method for modeling the normal patches using a new descriptor. The relation of each patch with their adjacent is modeled, in testing phase those patch which don't follow from this models are considered as abnormal event. The results confirm that the good performance of our method in comparison with several state-of-the art methods. Our method also enjoys low complexity.

## References

[1] Jianga, F., Yuan, J., Tsaftarisa, S. A.,katsaggelosa, A. K.: Anomalous video event detection using spatiotemporal context. Computer Vision Image Understanding, 115(3) 323-333 (2011)

[2] Wu, S., Moore, B., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: CVPR, (2010)

[3] Piciarelli, C., Foresti, G.L.: On-line trajectory clustering for anomalous events detection. Pattern Recognition Letters, 27(15) 1835-1842 (2006)
[4] Piciarelli, C., Micheloni, C., Foresti, G.L.: Trajectory-based anomalous event detection. IEEE Trans. Circuits Systems Video Technology, 18(11) 1544-1554 (2008).
[5] Antonakaki, P., Kosmopoulos, D., Perantonis, S.J.: Detecting abnormal human behaviour using multiple cameras. Signal Processing, 89(9) 1723-1738 (2009)
[6] Calderara, S., Heinemann, U., Prati, A., Cucchiara, R. Tishby, N.: Detecting anomalies in peoples trajectories using spectral graph analysis. Computer Vision Image Understanding. 115(8) 1099-1111 (2011)
[7] Morris, B.T., Trivedi, M.M.: Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. IEEE Trans. Pattern Analysis Machine Intelligence, 33(11), 2287-2301 (2011)
[8] Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE Trans. Pattern Analysis Machine Intelligence. 28(9) 1450-1464 (2006)
[9] Tung, F., Zelek, J.S., Clausi, D.A.: Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. Image and Vision Computing, 29(4) 230-140 (2011)
[10] Zhang, D., Gatica-Perez, D., Bengio, S., McCowan.: Semi-supervised adapted HMMS for unusual event detection. In: CVPR (2005)
[11] Boiman, O., Irani, Mi.: Detecting irregularities in images and in video. Int. J. Computer Vision. 74(1) 17-31 (2007)
[12] Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiplexed location monitors. IEEE Trans. Pattern Analysis Machine Intelligence, 30(3) 555-560 (2008)
[13] Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: CVPR, pages 1975-1981 (2010)
[14] Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Analysis Machine Intelligence, 36(1), 18-32 (2014)
[15] Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: CVPR, pp. 2921-2928 (2009)
[16] Benezeth, Y., Jodoin, P.M., Saligrama, V., Rosenberger, C.: Abnormal events detection based on spatio-temporal co-occurrences. In: CVPR, pp.1446-1453 (2009)
[17] Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: CVPR, pp.1446-1453 (2009)
[18] Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: CVPR, pp:935-942, 2009.
[19] Zaharescu, A., Wildes, R.: Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-drivenprocessing. In: ECCV 2010, 563-576 (2010)
[20] Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: CVPR, 3449-3456 (2011).
[21] Anti_c, B., Ommer, B.: Video parsing for abnormality detection In: ICCV. pp.2415-2422 (2011)
[22] Saligrama, V., Chen, Z.: Video anomaly detection based on local statistical aggregates. In: CVPR, (2012)
[23] Ullah, H., Conci, N.: Crowd motion segmentation and anomaly detection via multi-label optimization. In: ICPR workshop Pattern Recognition Crowd Analysis (2012)
[24] Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in MATLAB. In:ICCV. pp.2720-2727 (2013).
[25] Roshtkhari, M.J., Levine, M.D.: An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Computer Vision Image Understanding, 117(10) 1436-1452 (2013)
[26] Zhu, Y., Nayak, N., Roy-Chowdhury, A.: Context-aware modeling and recognition of activities in video. In: CVPR, 2491-2498 (2013)
[27] Cong, Y., Yuan, J., Tang, Y.: Video anomaly search in crowded scenes via spatio-temporal motion context. IEEE Trans. Information Forensics Security. 8(10) 1590-1599 (2013)
[28] Roshtkhari, M., Levine, M.: Online dominant and anomalous behavior detection in videos. In: CVPR, 2611-2618 (2013)
[29] Ullah, H., Tenuti, L., Conci, N.: Gaussian mixtures for anomaly detection in crowded scenes. IS&T/SPIE Electronic Imaging, 866303-866303 (2013)
[30] Ullah, H., Ullah, M., Conci, N.: Real-time anomaly detection in dense crowded scenes. IS&T/SPIE Electronic Imaging, 902608-902608 (2014)
[31] Ullah, H., Ullah, M., Conci, N.: Dominant Motion Analysis in Regular and Irregular Crowd Scenes. ECCV workshop HBU, 62-72 (2014)
[32] Xu, D., Song, R., Wu, X., Li, N., Feng, W., Qian, H.: Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing, 143, 144-152 (2014)

[33] Sabokrou, M., Fathy, M., Hoseini, M., Klette, R.: Real-Time Anomaly Detection and Localization in Crowded Scenes. In: CVPR workshop (2015)
[34] Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N., Kessler, F.B.: Learning deep representations of appearance and motion for anomalous event detection. In:BMVC (2015)
[35] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and com-
posing robust features with denoising autoencoders. In: Int. ACM Conf. Machine Learning, 1096-1103 (2008)
[36] Mousavi, H., Nabi, M., Galoogahi, H.K., Perina, A., Murino, V.: Abnormality detection with improved histogram of oriented tracklets. In: ICIAP (2015)
[37] Yuan, Y., Fang, J., Wang, Q.: Online Anomaly Detection in Crowd Scenes via
Structure Analysis. IEEE Trans. Cybernetics, 45(3) 562-575 (2015)
[38] Cheng, K.W., Chen, Y.T., Fang, W.H.: Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. In: CVPR. 2909-2917 (2015)
[39] Xiao, T., Zhang, C., Zha, H.: Learning to Detect Anomalies in Surveillance
Video. IEEE Signal Processing Letters, 22(9) 1477-1481 (2015)
[40] Lee, D.G., Suk, H.I., Park, S.K., Lee, S.W.: Motion Inuence Map for Unusual Human Activity Detection and Localization in Crowded Scenes. IEEE Trans. Circuits Systems Video Technology 25(10) 1612-162 (2015)
[41] Li, N., Wu, X., Xu, D., Guo, H., Feng, W.: Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. Neurocomputing, 155, 309-319 (2015)
[42] D. Brunet, E. R. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. IEEE Trans. Image Processing, 21(4):1488–1499, 2012.
[43] Wang, Zhou and Bovik, Alan Conrad and Sheikh, Hamid Rahim and Simoncelli, Eero P Image quality assessment: from error visibility to structural similarity In IEEE Transactions on Image Processing. 13(4), 600–612, 2004.