Convex Polytope Ensembles for Spatio-Temporal Anomaly Detection

Francesco Turchini, Lorenzo Seidenari, Alberto Del Bimbo

University of Florence
francesco.turchini@unifi.it, lorenzo.seidenari@unifi.it,
alberto.delbimbo@unifi.it

Abstract. Modern automated visual surveillance scenarios demand to process effectively a large set of visual stream with a limited amount of human resources. Actionable information is required in real-time, therefore abnormal pattern detection shall be performed in order to select the most useful streams for an operator to visually inspect. To tackle this challenging task we propose a novel method based on convex polytope ensembles to perform anomaly detection. Our method relies on local trajectory based features. We report State-of-the-Art results on pixel-level anomaly detection on the challenging publicly available UCSD Pedestrian dataset.

Keywords: computer vision, anomaly detection, surveillance

1 Introduction and Related Work

Nowadays a huge effort is put in securing cities and public spaces. Apart from human engagement in security policy with police forces and other security personnel, a lot of spending is dedicated to surveillance system deployment. Unfortunately while growing the amount of operators may enhance the security, growing the amount of sensors alone is not obtaining much benefits. While cameras are often installed as a deterrent for crimes, the usual approach is to use footage as evidence in investigations. More actionable information could be gathered if real-time video analysis provided to surveillance operators a subset of frames to inspect. Dadashi *et al.* [1] conducted a study to understand the role of automatic and semi-automatic video analysis in security context. They have shown that when reliable automatically computed information is provided workload is greatly reduced. This kind of support to human operators is key since, as reported in [2] the attention of operators, viewing multiple streams, greatly degrades just after 20 minute.

A very desirable feature in automatic visual surveillance system, is the ability to pick the right set of streams to watch. This can be casted as measuring the deviation of the most recent frames, from some nominal distribution of the imagery for the very same stream. More specifically an algorithm, selecting streams, should also provide localization of such anomalies. This is an important feature

since it allows to use high resolution PTZ cameras able to directly frame, at a higher quality, the abnormal pattern.

Modeling complex patterns requires to learn the distribution characterizing a set of video sequences, taken from a certain view. It is usually assumed that the camera is fixed, this allows to make models which are simpler and can learn patterns which are scene specific. Anomaly detection is usually casted as a one-class learning problem over features extracted from video sequences.

Most of the recent works are based on motion or spatio-temporal features. The seminal work from Adam *et al.* [3], learned local optical flow statistics and compared them to the one computed on forthcoming frames. Optical flow has been used extensively as low-level feature on which contextual models are then built [4], [5]. One of the main limitation of optical flow lies in the impossibility to model appearance abnormalities. Nonetheless, using just the appearance, is only suitable for low-frame rate scenarios [6], therefore many work resort to spatio-temporal representation, in order to jointly capture appearance and motion [7], [8], [9], [10], [11].

Several models have been applied to solve one-class learning. Non-parametric approaches [9][6], model feature distribution implicitly, by looking at distance between features. Parametric models, have the advantage of a lower memory footprint, they typically fit a mixture of density functions on the extracted features. Li *et al.* [10] learn a mixture of dynamic textures, computing likelihood over unseen patterns to perform inference. Similarly, Kim *et al.* [4] learn a mixture of Principal Components Analyzers, which jointly learns the distribution and perform dimensionality reduction. Feature learning has been rarely used except for Xu *et al.* [7], which use autoencoders to directly learn the representation, obtaining high accuracy. In this work, we only consider methods not using anomaly labels in learning, in such cases, the problem becomes a binary classification task with much less challenge.

In the past, trajectories were the feature of choice to model patterns in visual surveillance scenarios [12]. Trajectory based anomaly detection unfortunately requires high quality object tracking and can not find appearance abnormal patterns. In action recognition, the use of short local trajectories, namely dense trajectories, to extract features has led to a sensible increase in performance [13]. Several approaches build on this features, showing interesting further improvements and localization capabilities [14],[15]. Up to now we are not aware of such features being employed in unsupervised or semi-supervised tasks like anomaly detection.

Considering the relatively low computational requirement and high performance, we build on dense trajectories, which are known to be very well suited for a wide set of action recognition problems, since they are able to represent motion and appearance jointly. We propose to estimate the distribution of trajectory descriptors using convex polytopes [16]. Convex polytopes have been used in the past but never for computer vision problems. Our approach is inspired by [16], but is different since instead of modeling the distribution of data with a single polytope which is approximated using random projections, we consider

explicitly an ensemble of low-dimensional models. This approach is more suited to model multi-modal distributions and it allows to merge multiple features in a single decision.

We report state of the art results on the UCSD dataset both at pixel and frame level anomaly detection. Interestingly we found that local trajectory shape can get very good detection rates, potentially reducing the computational cost for feature extraction.

2 Anomaly Detection with Convex Polytopes

We tackle anomaly detection and localization as a single-class classification problem in a fully unsupervised way. As we can only train our system on a single class of input points (the non-abnormal class), we choose to employ the polytope ensemble technique as modeling method. In particular, we make use of Polytope Ensemble technique [16]. Polytope Ensemble considers a set of convex polytopes representing an approximation of the space containing the input feature points. We want a representation which is shaped according to the distribution of the points we can observe; among the convex class of polytopes, the convex hull has the geometric structure which is best tailored to model this kind of data distribution.

2.1 Model building

Given an input set of points $\boldsymbol{X} = \{x_1, \ldots, x_m\}$, its convex hull is defined as

$$\boldsymbol{C(X)} = \left\{ \sum_{i=1}^{|X|} \theta_i x_i | x_i \in \boldsymbol{X};\ \sum_i \theta_i = 1, \theta_i \geq 0 \forall i \right\} \tag{1}$$

By exploiting the convex hull properties, we can then identify an abnormal point simply checking whether it belongs to the convex hull or not.

Extended convex hull To ensure robustness of the model, we follow the procedure of [16] and modify the structure of the convex hull, performing a shift of its vertices closer or farther from its centroid. This allows to avoid overfitting and tune our system to cope with different practical conditions. Considering the set of vertices $\boldsymbol{V} \subset \boldsymbol{X}$ and the centroid of the polytope c_i, we can calculate the expanded polytope setting an α parameter such that

$$\boldsymbol{V_\alpha} = \{v + \alpha \frac{(v - c_i)}{||v - c_i||}, v \in \boldsymbol{V}\} \tag{2}$$

The new polytope defined by vertices in $\boldsymbol{V_\alpha}$ is a shrunken/enlarged version of the original convex hull. Negative values of α increase system sensivity, while positive values reduce it.

Ensemble building We rely on dense trajectory features [13]. We extract both motion and appearance descriptors using Improved Dense Trajectories algorithm. This allows us to jointly employ multiple features such as trajectory coordinates, HoG, HoF and MBH to achieve robust anomaly detection and localization. We set the ensemble size to T convex hulls. Then, for each feature and for each convex hull, we generate a random projection matrix P_i^f with norm 1 and size $d \times D_f$, where d is the size of the destination subspace, and D_f is the size of the feature f. We then apply this projections to the original data:

$$\boldsymbol{X}_{\boldsymbol{P_i^f}} = \{P_i^f x, \forall x \in \boldsymbol{X}\} \tag{3}$$

The i-th convex hull is calculated on $\boldsymbol{X}_{\boldsymbol{P_i^f}}$. Each convex hull will be characterized by a unique shape, as we generate a different random projection matrix at every iteration of model learning. A set of different sensitivity ensembles can be obtained by the aforementioned shrinking/expansion procedure, based on different values of the α parameter. It is not required to have an α set for each polytope since, as can be seen in Eq.2, shrinking factors are computed by scaling the distance of vertices from the centroid.

2.2 Anomaly Localization

At inference time, we test each extracted descriptor for inclusion in each convex hull of the ensemble, for each feature. We consider a local trajectory, with descriptors x_f as anomalous if the following condition is true:

$$x_f \notin C^f(\boldsymbol{X}_{\boldsymbol{P_i^f}}) \; \forall f, i \tag{4}$$

meaning that the descriptor is external to all the polytopes and that this happens for all the considered features (Trajectories, HoG, HoF, MBH).

These assumptions are rather strong, but they ensure that we reduce anomaly detection on unusual but yet ordinary patterns. When a descriptor is marked as abnormal, this detection lasts for the entire extent of the trajectory descriptor (15 frames by default). Detecting anomalies for individual trajectory descriptors allows to generate anomaly proposals in various areas of video frames, exploiting trajectory coordinates. We can then obtain an anomaly mask for each frame of each video by filtering these proposals. In Fig. 1 we represent the three main operations we perform to achieve anomaly detection and localization.

We take into consideration the set of trajectories $\boldsymbol{T_a} = \{t_1, t_2, \ldots, t_N\}$ which have been marked as anomalous after testing their inclusion into the convex hulls of the ensemble. Each trajectory t_i is a sequence of M points, $t_i = \{p_{i1}, \ldots, p_{iM}\}$ lasting M video frames. At frame f, we consider the points of the active anomalous trajectories, that is to say the set of points

$$\boldsymbol{P_a} = \{p_{in} \in t_i | n = f, \; t_i \in \boldsymbol{T_a}\} \tag{5}$$

Points identified by active anomalous trajectories at frame f are clustered with K-means algorithm to locate potentially abnormal areas of the frame. K-Means yields a partition $\boldsymbol{S_a}$ of the anomalous points set $\boldsymbol{P_a}$ in K Voronoi cells:

$$\boldsymbol{S_a} = \{S_1, \ldots, S_K | S_1 \cup \cdots \cup \; S_K = P_a, \; S_{k_1} \cap \; S_{k_2} = \emptyset \; \forall \; k_1, k_2\} \tag{6}$$

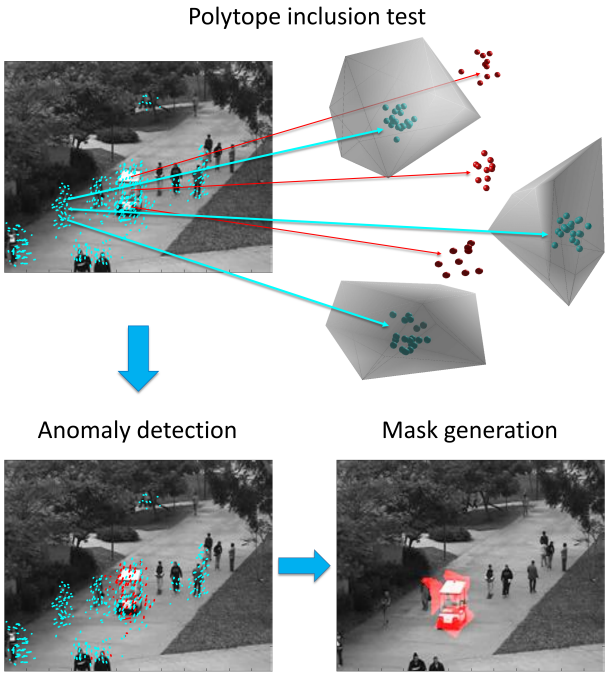


Fig. 1: Operating scheme of our anomaly detection and localization model

Each S_k represents an anomaly proposal for the considered frame. For each S_k, we verify if its cardinality is smaller than a fixed threshold, that is to say, if the anomaly proposal constitutes of a minimum number of points. We assume that small clusters are likely originated by spurious false positive detections, so we discard all the anomaly proposals S_k whose cardinality does not guarantee that the detection is reliable. Then, for each remaining S_k, we calculate the polygon described by its points. Each polygon represents an accepted anomaly proposal which contributes to the final anomaly mask creation for the frame.

3 Experimental Results

We conduct our experiments on the UCSD Pedestrian dataset. This dataset has been proposed by Mahadevan *et al.* [10], and it consist of two sets of videos, named Ped1 and Ped2, of pedestrian traffic. The dataset is not staged and features realistic scenarios. In the setting designed by the authors anomalous patterns are all the non-pedestrian entities appearing in the scene. We perform the evaluation on the Ped1 and Ped2 following the standard experimental protocol for this dataset which comprises two evaluation settings: frame-level and pixel-level [10].

In the frame-level criterion, detections are evaluated frame-wise, meaning that a frame is considered anomalous if at least an abnormal detection is pre-

dicted for that frame disregarding its location. In this setting it is possible to have "lucky guesses", predicting a frame correctly thanks to a detection which is spatially incorrect or with a too small overlap with the ground truth annotation.

Pixel-level evaluation is introduced to obtain a more detailed analysis of algorithm behavior. In this setting anomaly detections are compared with ground truth pixel masks. A frame is considered a true positive if there is at least 40% of pixel overlap between the ground truth and the predicted mask. A frame is considered a false positive in case anomalies are predicted in normal frames or if the overlap with ground truth masks is lower than 40%. We report the Receiver Operating Characteristic (ROC) curve of TPR and FPR varying system sensitivity, and the Rate of Detection (RD) of our system. We modify system sensitivity varying α in Eq.2.

First we perform an analysis of the contribution of different features. For simplicity, we divide features in three groups: trajectories, motion and appearance. We test each kind of feature alone and in combination with the others on UCSDPed1. We report the results of feature evaluation in Tab. 1.

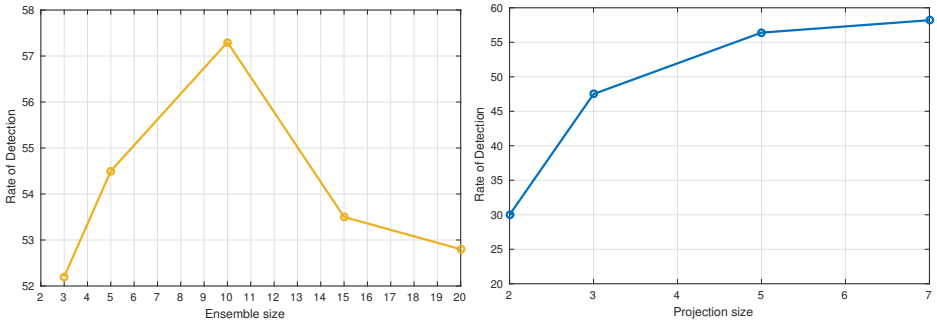
Trajectories	**Motion (HoF, MBH)**	**Appearance (HoG)**	**RD**
✓	-	-	57.9
-	✓	-	60.1
-	-	✓	48.9
✓	✓	✓	**62.2**

Table 1: Pixel Level Rate of Detection for different descriptors on UCSDPed1

Interestingly, local trajectories show very good performance. Anyhow, it appears clearly that motion descriptors give the main contribution to anomaly localization; however, as expected, best results are obtained fusing the contributions of all descriptors. In the following, we will then perform other tests using all the descriptors extracted from the dense trajectory pipeline.

Regarding our model, there are two parameters that can affect the performance. In the following experiments we want to understand how projection size and ensemble cardinality influence the correct detection of anomalies.

All projection size tests were obtained fixing ensemble size to 10 convex hulls, while all ensemble size tests were obtained fixing projection size to 5. We report detection rate variation charts in Fig. 2. As we expected, increasing projection size leads to consistent gain in rate of detection results. On the contrary, bigger ensembles do not always guarantee performance improvements. This outcome may be caused by the unpredictable behavior of the random projections when we raise the number of random generated projection matrices. The best trade-off from a computational point of view is obtained keeping an ensemble of 10 convex hulls and a projection size of 5 dimensions. Increasing projection size



(a): RD varying ensemble size (b): RD varying projection size

Fig. 2: Evaluation of ensemble size and projection size for our system on UCSDPed1.

over 7 causes convex hull generation and inclusion test to be nearly unfeasible due to very long computation time without bringing noticeable benefits.

With these settings fixed, we compare our results with the existing State-of-the-Art methods in fully unsupervised settings. First of all, it can be noted that with our method trajectory descriptors alone obtain very high Rate of Detection at the pixel level (57.9% as shown in Tab. 1), higher than most approaches on Ped1, excluding [11], and the deep learning based method by [7].

As we can see in Fig. 3 and 4, our method succeeds in limiting false positive detections, especially at low sensitivity, at the frame level. We detect and localize less than 20% of false positives facing more than 50% of true positives at lower sensitivity values on Ped1 setting. Our system behaves even better on Ped2 setting, where we correctly detect and localize more than 50% of true positive anomalies with less than 5% of mistakes. As we expect, false positive rate increases when our system becomes more sensitive to unseen patterns, however maintaining good robustness. Tab. 2 reports Rate of Detections for all considered methods for both datasets and both criteria, when reported by authors. Our method obtains a frame-level performance which is comparable to the State-of-the-Art and beat all existing methods on the more challenging pixel-level evaluation. Considering the evaluation protocol established in [10], frame level accuracy may not reflect the actual behavior of a method, because of lucky guesses, while the pixel-level criterion is stricter.

To show the high quality of our generated masks, we report a qualitative comparison on two frames. Notably our masks frame very tightly abnormal patterns, such as the bicycle rider and the truck in Fig. 5 and Fig. 6. With respect to [10] our masks are tighter. Methods such as MPPCA, Force Flow and LMH, are not able, especially in Ped2, to locate all anomalies. This is likely due to a lower quality of features employed.

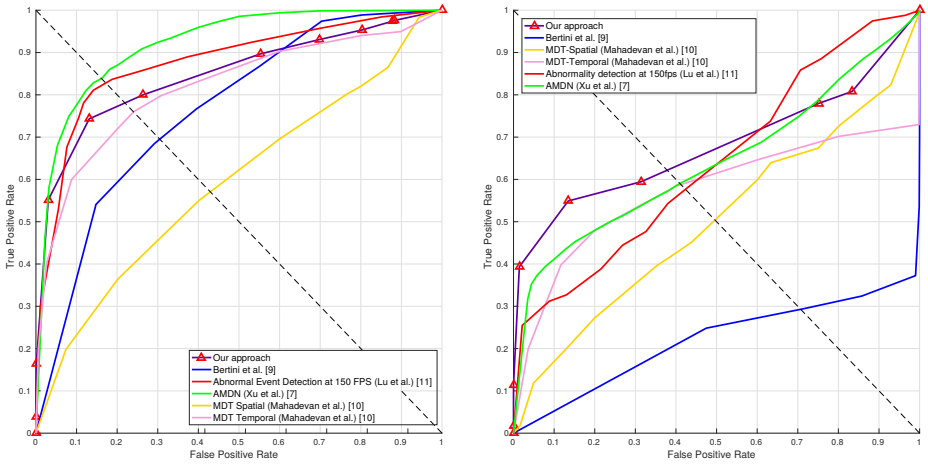


Fig. 3: TPR-FPR curves comparing our approach with various well-known methods on Ped1 setting. Left figure shows the Frame level criterion, right figure shows Pixel level criterion.

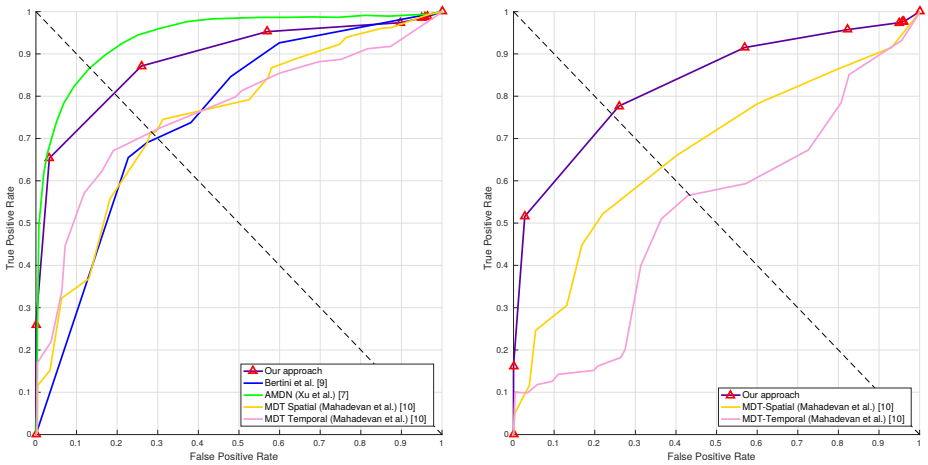


Fig. 4: TPR-FPR curves comparing our approach with various well-known methods on Ped2 setting. Left figure shows the Frame level criterion, right figure shows Pixel level criterion.

Method	Ped1		Ped2	
	Frame	**Pixel**	**Frame**	**Pixel**
Ours	78.1	**62.2**	80.7	**75.7**
Xu et al. [7]	78.0	59.9	**83.0**	-
MDT Spatial [10]	56.2	54.2	71.3	63.4
MDT Temporal [10]	77.1	48.2	72.1	56.8
150 fps [11]	**85.0**	59.1	-	-
Bertini et al. [9]	66.0	29.0	68.0	-
Mehran et al. [5]	63.5	40.9	65.0	27.6
Kim et al. [4]	64.4	23.2	64.2	22.4
Adam et al. [3]	61.1	32.6	54.2	22.4

Table 2: RD comparison of our method versus various well-known State-of-the-Art techniques on Ped1 and Ped2 (where available) settings, frame-level and pixel-level criteria.

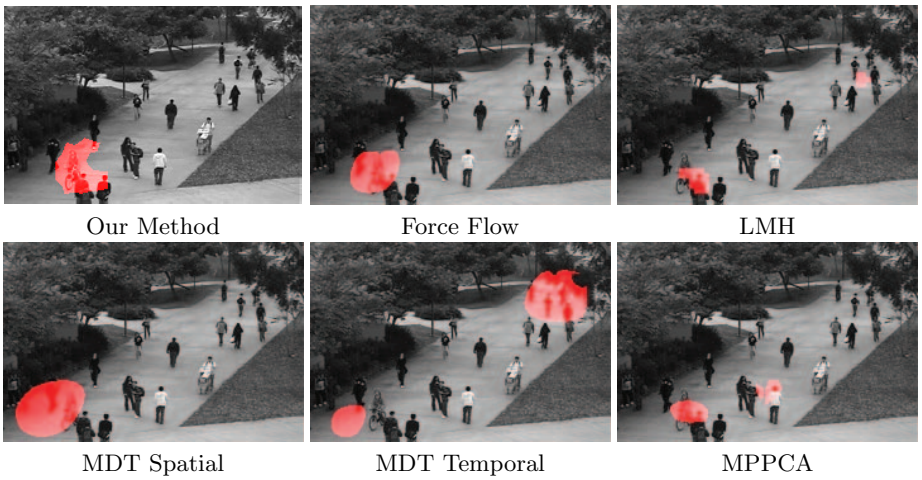


Fig. 5: Qualitative pixel level anomaly detection results on UCSD Ped1 comparing our method to previous approaches.

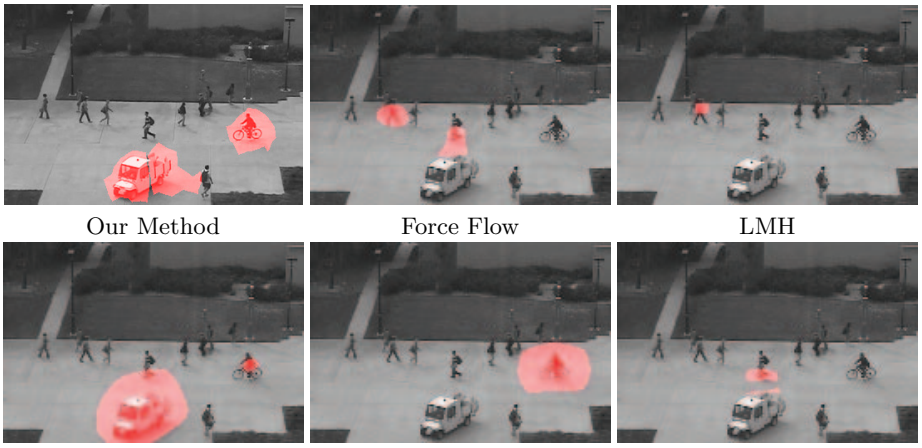


Fig. 6: Qualitative pixel level anomaly detection results on UCSD Ped2 comparing our method to previous approaches.

4 Conclusion

In this paper we show a novel, low memory footprint method to exploit dense trajectory features in anomaly detection. Our method is able to model a complex multimodal distribution yielded by spatio-temporal descriptors using a simple convex polytope ensemble. Moreover, when multiple views of the same datum are available our approach seamlessly performs feature fusion. Indeed, our method is very flexible, as it allows to combine multiple features maintaining the same operating mechanisms, and is tunable by a simple geometric transformation of polytope hulls. Our system can thus be adapted to cope with various practical conditions without losing its benefits both for anomaly detection and localization tasks. We also propose a technique to obtain precise masks by clustering abnormal trajectories; this mask generation technique allows us to achieve good robustness against false positive detections and is shown to obtain State-of-the-Art results in term of pixel-wise detection rate.

References

1. N Dadashi, AW Stedmon, and TP Pridmore. Semi-automated cctv surveillance: The effects of system confidence, system accuracy and task complexity on operator vigilance, reliance and workload. *Applied ergonomics*, 44(5):730–738, 2013.
2. Niels Haering, Péter L Venetianer, and Alan Lipton. The evolution of video surveillance: an overview. *Machine Vision and Applications*, 19(5):279–290, 2008.
3. Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560, 2008.
4. Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *Proc. of CVPR*, pages 2921–2928. IEEE, 2009.
5. Ramin Mehran, Alexis Oyama, and Mubarak 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.
6. Michael D Breitenstein, Helmut Grabner, and Luc Van Gool. Hunting nessie-real-time abnormality detection from webcams. In *Proc. of ICCV Workshops*, pages 1243–1250. IEEE, 2009.
7. Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *Computer Vision and Image Understanding*, 2015.
8. Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proc. of CVPR*, pages 1446–1453. IEEE, 2009.
9. Marco Bertini, Alberto Del Bimbo, and Lorenzo Seidenari. Multi-scale and real-time non-parametric approach for anomaly detection and localization. *Computer Vision and Image Understanding*, 116(3):320–329, 2012.
10. Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014.
11. Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proc. of ICCV*, pages 2720–2727, 2013.
12. Simone Calderara, Andrea Prati, and Rita Cucchiara. Mixtures of von mises distributions for people trajectory shape analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4):457–471, 2011.
13. Heng Wang, Dan Oneata, Jakob Verbeek, and Cordelia Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119(3):219–238, 2016.
14. Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Activity representation with motion hierarchies. *International journal of computer vision*, 107(3):219–238, 2014.
15. Francesco Turchini, Lorenzo Seidenari, and Alberto Del Bimbo. Understanding and localizing activities from correspondences of clustered trajectories. *Computer Vision and Image Understanding*, 2016.
16. Pierluigi Casale, Oriol Pujol, and Petia Radeva. Approximate polytope ensemble for one-class classification. *Pattern Recognition*, 47(2):854–864, 2014.