Abnormal Event Detection Using Convolutional Neural Networks and 1-Class SVM classifier

Samir Bouindour, Mohamad Mazen Hittawe, Sandy Mahfouz, Hichem Snoussi

University of Technology of Troyes, Institut Charles Delaunay/ LM2S 12 Rue Marie Curie, Troyes, France

Keywords: Anomaly detection . Deep learning . Features extraction . Classification .

Abstract

In this paper, we present a method based on deep learning for detection and localization of spatial and temporal abnormal events in surveillance videos using training samples containing only normal events. This work is divided into two stages, the first one is feature extraction for each patch of the input image using the first two convolution layers extracted from a pre-trained CNN. In second stage, one-class SVM is trained with resultant features. The SVM classifier allows a fast and robust abnormal detection with respect to the presence of outliers in the training dataset. Experimental tests have conducted on UCSD Ped2 dataset, this dataset is considered as complex due to low resolution and presence of many occlusions. Our results showed high performance and were compared with state-of-the art methods.

1 Introduction

Anomaly detection and localization is a challenging task in video surveillance due to the fact that the definition of anomaly is subjective and context-dependent. The use of surveillance cameras requires that computer vision technologies need to be involved in the analysis of very large volumes of video data. Detection of anomalies in crowded scenes is one of the challenging and exciting applications in this area. In the context of crowded scene videos, anomalies are formed by rare shapes or rare motions or both of them. Many works are based on the analysis of the trajectories to detect the abnormal events. These methods often require accurate tracking solutions and are also very sensitive to occlusion. They can be used in detecting deviant trajectories in scenes containing few objects but become not efficient and time consuming in overcrowded scenes. There are also other methods which do not require the analysis of trajectories, these methods extract spatio-temporal features in order to represent the events in the video. Among these methods, low-level local visual features such as motion or texture (Histogram of Oriented Gradients, Histograms of optical flow) to model the background and to construct the template behavior have been used. Since the activity pattern of a pixel can not be used for behavioral understanding, these methods are not optimal for complex events.

Other works were carried out in the same perspective of extracting spatio-temporal features, inspired by the concept of Bag of Video words (BOV). This concept is based on using local video volumes obtained either by dense sampling or by selection of points of interest for the detection of abnormal events. However, the relationship between video volumes is often not taken into account. Other methods that considered the contextual information have been recently proposed, but usually, these methods require high computational capacity, which often makes them unusable in real time. Other methods used supervised learning, where training examples are composed of both normal and abnormal events, which makes them not usable in the field of video surveillance as it is very difficult to identify and reproduce all the abnormal events.

In this article, we propose a method based on deep learning for the detection and localization of abnormal events that may reflect suspicious situations. The main challenge is extracting robust descriptors and defining classification algorithms adapted to detect suspicious behaviors with the minimum values of false alarms, while ensuring a good rate of detection.

The rest of the paper is organized as follows: In Section 2, we present some of the state-of the-art methods. In Section 3, we detail our proposed detection method. We present experimental results to evaluate our method in Section 4. Finally, Section 5 concludes the paper.

2 Related work

The previous works on the detection of abnormal events have focused on the analysis of the trajectories. A moving object is considered as abnormal if it does not respect the normal learned trajectories [1, 2, 3, 4, 5, 6, 7]. The main drawback of these methods is the fact that the algorithms are very sensitive to occlusions and generally require precise tracking techniques in order to locate the objects and differentiate them, which is very difficult in the crowded scenes.

To avoid these weaknesses, many works have been devoted to the extraction and analysis of low-level local features such as motion or texture, either by constructing the model of pixel-level background and the model of normal events. In [8], the authors used temporal video filtering. In [9], the authors constructed the normal model of a video with Markov random field (MRF), with respect to some features such as rarity or relevance. Boiman and Irani [10] considered that if the event reconstruction using only the previous observations is impossible then this event is classified as an anomaly. The authors in [11]

used an exponential distribution for the modeling of the histograms of optical flow (HOF) in local regions. In [12], the authors proposed to use a mixture of dynamic texture (MDT) to represent the video. Another method based on the optical flow for data clustering was proposed by [13]. In [14], the authors used learning co-occurrence matrix to build the model of the events in video. In [15], a Gaussian model is constructed with spatio-temporal gradient features, and a hidden Markov model (HMM) is then used in order to detect anomalies. The authors in [16] included the spatio-temporal compositions of the motion to build background and behavior models, the new events are compared to the model to decide whether the event is abnormal or not. In [17], the events are described using spatio-temporal oriented energy filters to construct the activity model for each pixel. The authors in [18] used Histograms of Optical Flow Orientation (HOFs) as motion descriptor and one class SVM is then trained to detect the abnormal events, the authors showed that the solution based on optical flow is efficient for abnormal moving object. However, the main drawback of this method is not to detect abnormal shapes in case of less movements.

The methods based on low-level local features are not optimal for complex abnormal events detection, as the activity pattern of a pixel can not be used for behavioral understanding. Their applicability in video surveillance is limited to the detection of local temporal phenomena.

Many other works are based on the use of spatio-temporal video volumes, obtained either by dense sampling [10, 19, 20] or by selection of interesting points [21, 22]. The social force (SF) has been introduced by [23] for abnormal motion modeling in crowded scenes. The real time performance of the dense sampling methods is directly related to the number of video volumes and the features associated with them, that is why methods which use interest point selection have an advantage over those using dense sampling. However, the problem of such methods is the deficiency of the interest point detection in some complex environments.

In most of previous methods, contextual information has not been enough considered for event understanding. In order to avoid this weakness, authors in [24] used three-dimensional spatio-temporal pyramid matching. In [10], the authors presented a method based on dense sampling in order to obtain the spatio-temporal video volumes, they defined a large region around each volume to consider contextual information. However, high computational cost is still the main drawback of such methods.

Recently, some works based on deep learning have been proposed for anomaly detection. In [25], authors used the optical flow to select video volumes of interest and a convolutional neural network (CNN) for features extraction. In this work, authors used learning examples for both normal and abnormal events. This is the major drawback of this method, as the number of abnormal events and their diversity make impossible to provide relevant learning examples. Many unseen and not well represented abnormal events may not be detected, during the test phase. In [26], the authors proposed fully convolutional neural networks (FCNs). It is a combination of a pre-trained CNN and a new convolutional layer which is trained with sparse auto-encoder. Our work is mainly based on [26] for feature extraction. Our contribution is the extension of this work to integrate a robust classifier based on 1-Class SVM methodology taking into account the possible presence of some outliers in training data and reduce the influence of size of the training dataset on the computational cost.

```
Initialization:
N=nbr-frames
For i =1: N
    #F: frame
    X=features-extraction (Fi)
    #size(X) = 529  vectors of 256 dimensions
    Features_Train=[ Features_Train; X]
End
#size(Features_Train)=(529 * N) vectors of 256 dimensions
Model = Train_SVM(Features_Train)
```

```
Abnormal events detection:
For each new frame Fi
    Features_Test =features-extraction (Fi)
    #size(Features_Test) = 529 vectors of 256 dimensions
    [ Labels, Score ] = Predict (Model, Features_Test)
    # size (Score) = 529*1
    # Each score represents if the small patch is normal or abnormal in the input frame
    For j=1:529
        If score (j) < threshold
            Patch_j is Abnormal
        End
    End
End
```

Figure 1. Algorithm of abnormal events detection and localization .

3 Proposed abnormal event detection method

In this section, we present a method for detecting and localization of abnormal events in scenes, based only on training dataset of normal situations. Our work is divided into two main stages, the first one consists of extracting robust and discriminative features using the first two convolutional layers of a pre-trained CNN, and in the second stage, we use the resultant features to train a one-class nonlinear SVM algorithm.

3.1 Features extraction

The network Alexnet [27] has been trained on a very large number of images for the places and objects recognition which

gives it the ability of extracting robust features from images. An abnormal event is characterized by the presence of abnormal shapes, abnormal movements or the presence of both of them. In terms of features extraction, using one frame at each time as input for CNN does not take in account the motion information, so we use for each frame F_t, as proposed in [26], the volume D consisting of the frames $D = [F_t, F_t - 1, F_t - 2]$ as inputs to the network.

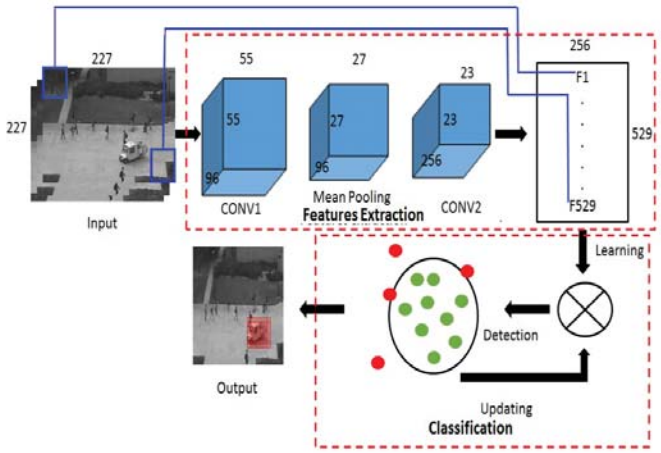

Figure 2. Flowchart of the proposed method

Convolutional layers are considered as an important stage in CNNs because they allow to generate the feature maps. However, there are also another important concept which is represented by pooling layers. The pooling layers allow to reduce the spatial size of the future maps progressively between the convolutional layers, which reduce the number of parameters in network and control the overfitting. Thus, if the complete CNN (from end to end) for each input frame is used, then, the resultant features after the fully connected layers will be feature vector instead of feature map. Representing each frame by a feature vector allows to decide if the frame is normal or not. However, this representation does not allow to localize the anomalies inside the frame.

Our used feature map is extracted from the first two convolutional layers and one pooling layer between them in order to achieve the objective of detecting and localizing anomalies in each frame. The resultant feature map is a matrix of dimension 529×256, each raw representing the vector feature of one small patch in the input frame as shown in Fig. (2).

3.2 Classification

Support Vector Machine (SVM) proposed by [28] is considered as a statistical learning method for classification and regression. Afterward, SVM has been adapted to non-linear problems with using kernel methods[3, 18]. The kernel function is defined as the following :

$$k(X, X^{'}) = \Phi(X).\Phi(X') \tag{1}$$

$\Phi(X)$ is defined for solving non-linear classification problems and project the original input data χ to new feature space $\mathcal{H}$ where the classification problem has a linear solution. In our case, we use the polynomial kernel :

$$k(X, X^{'}) = (X, X^{'})^d \tag{2}$$

where d is polynomials degree.

The objective of this learning method is to find a hyperplane separator to classify the input data which can be described mathematically as the following :

$$f(X) = W.X + b \tag{3}$$

Corresponding to the decision function:

$$y(X) = sgn(f(X)) \tag{4}$$

Statistical learning theory states that the optimal classifier can be found by maximizing the margin [3]. This can be expressed as a minimization problem:

$$Min\frac{1}{2} \parallel W \parallel^2 \tag{5}$$

subject to :

$$Y_i * (W.X_i + b) \geq 1, i = \{1, ..., n\} \tag{6}$$

where n is the size of input training data and Y_i is data label (-1 or +1). In 1-class SVM framework, the data from only one class are available which matches our problem framework as we use only the normal event examples from the observed scene. The objective of one-class SVM, is to define a region in the space X which contains most of the data. This could be achieved by looking for a hyperplane in the feature space, and then try to maximize its distance from the the origin, while only a small set of data is located between the hyperplane and the origin [3].

The global scheme of the proposed abnormal events detection and localization is presented in Fig. 2. The algorithm is composed of two stages, the first one is the features extraction, and the second step is the classification. Fig. 1 summarizes the proposed algorithm:

Step 1 – Initialization : The feature are extracted from all the N training images (The resultant matrix is $Features\_train$ of dimension $529 \times N \times 256$). Then, these features have been used in order to train nonlinear one-class SVM.

Step 2 – Abnormal events detection : For each new test frame, we extract features (the resultant matrix is $Features\_test$ of dimension 529×256, each raw in $Features\_test$ represents one patch in the test frame) which will be given to the trained SVM model in order to decide if is normal or abnormal.

4 Experimental results

The proposed method is evaluated on UCSD Ped2 dataset (`http://www.svcl.ucsd.edu/projects/anomaly`) representing complex abnormal behaviors containing different outdoor scenes. This dataset contains 16 folders for training and 12 folders for testing. The Ped2 dataset

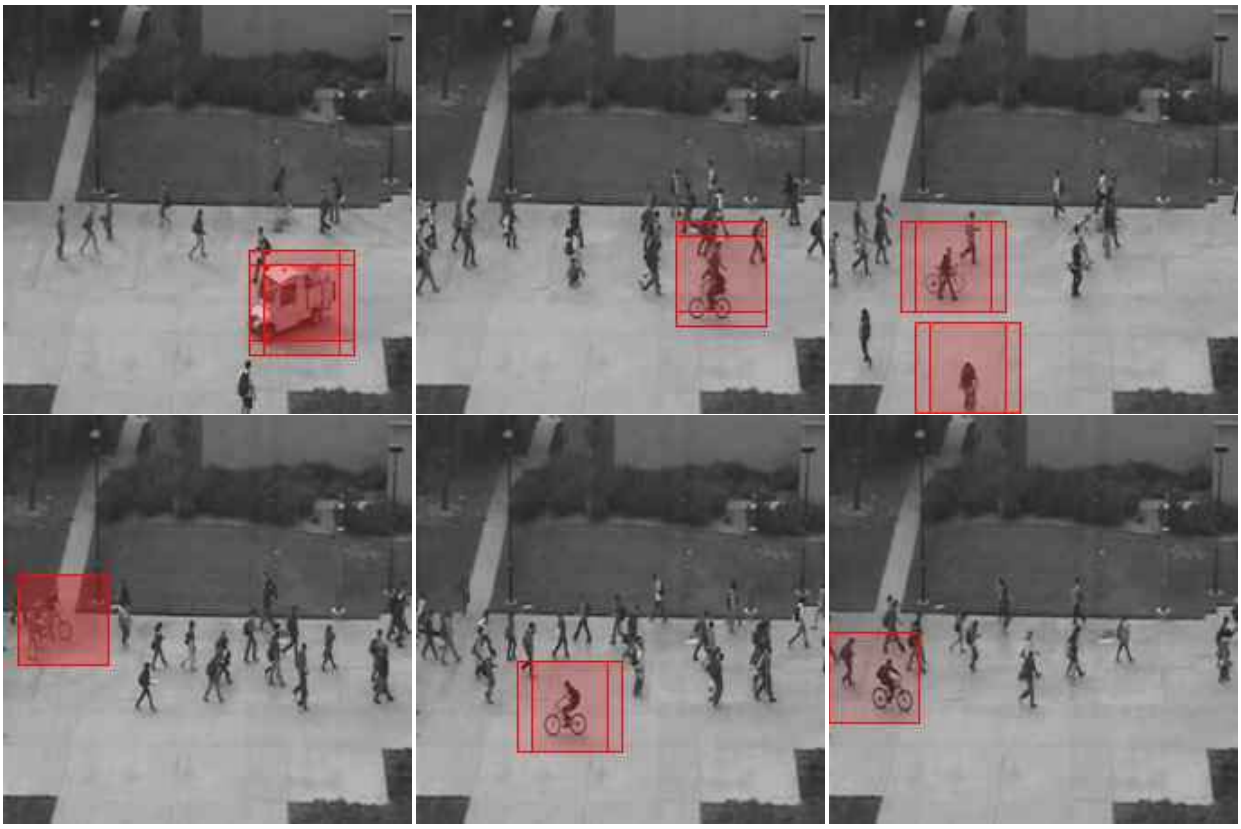

Figure 3. Qualitative results of the proposed method on UCSD PED2 dataset datset.

contains walking pedestrian images with resolutions 240×360 pixels. The main difficulties in UCSD dataset are the low resolution and the presence of many occlusions. The videos are taken for different crowd densities, the abnormal events in these videos are defined as the presence of non pedestrian actions like (bicycle, skateboarder, car ... etc) inside normal patterns of pedestrian.

4.1 Performance evaluation

As explained previously, our algorithm is based on using non-linear one-class SVM, in order to construct the model of normal events. Many experiments have been performed in order to select the best kernel in the SVM framework. We finally selected a polynomial kernel, as it yields the best results. Our algorithm has been implemented and tested under Matlab, for the deep learning part we used caffe [29] and we used the statistics and Machine Learning Toolbox for the SVM classifier. Note that, we use only frames containing normal events in training stage.

Fig. 3 shows visual results on the USCD Ped2 dataset. The red areas represent the abnormal events.

In order to evaluate the performance of the proposed method, we used the measure of EER (Equal Error Rate). EER is calculated as the following:

$$EER = \frac{FP + FN}{NF} \tag{7}$$

where :

- FP represents the number of frames detecting abnormal events by our method where there are no abnormal events.
- FN represents the number of frames with actual abnormal events not detected by our method.
- NF is the number of frames in each folder.

It should also be noted that for each test folder only the images of the corresponding train folder are used. In the table 1, we can clearly note that despite the reduced number of training images, our algorithm still achieves good results with very small training datasets (between 120 and 180 frames). For the other folders in Ped2 dataset, the results were not representative despite that we got good EER. In fact, the accuracy of the location of the anomaly in these folders is insufficient to be relevant.

Table 2 summarizes the comparison results of the proposed method with state-of-the-art methods on PED2 dataset. We can clearly observe that our method performs at frame level (if one pixel detects an anomaly then it is considered as being an anomaly) better that most of the state-of-the-art methods, except the FCN method [26], which achieved the best performances. However, the authors of [26] used mahalanobis distance with respect to the training data in order to detect the anomalies, which is computational consuming as the size of the training dataset increases. The one-class SVM classifier is

Folder	Nbr frames train	Nbr frames test	TP	TN	FP	FN	EER
Folder1	120	180	47	60	0	73	40%
Folder3	150	150	118	4	0	28	18%
Folder4	180	180	111	30	0	39	21%
Folder5	180	150	127	0	21	2	15%
Folder7	150	180	111	31	14	24	21%
Folder9	180	120	86	0	0	34	28%

Table 1. Quantitative results (EER frame level)of the proposed method on the UCSD Ped2 dataset

Method	EER
Adam. [11]	42%
Bertini. [30]	30%
Zhou. [25]	24.4%
Kim. [31]	30%
SF [23]	42%
FCN [26]	11%
Ours	***24.2%***

Table 2. Quantitative comparison of the proposed method and the state-of-the-art for anomaly detection using the Ped2 dataset and EERFL

invariant to the training dataset size, which allows its applicability in even very large video datasets.

5 Conclusion

A novel approach based on coupling deep learning and non-linear one-class SVM algorithm for detection and localization of spatial and temporal abnormal events has been proposed in this paper. It is based on cascaded stages: features extraction from normal events using pre-trained CNN and one-class SVM learned on these features. Results show high performance in detection and localization anomaly compared with state-of-the-art methods.

References

[1] Shandong Wu, Brian E Moore, and Mubarak Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2054–2060. IEEE, 2010.

[2] Fan Jiang, Junsong Yuan, Sotirios A Tsaftaris, and Aggelos K Katsaggelos. Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, 115(3):323–333, 2011.

[3] Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology*, 18(11):1544–1554, 2008.

[4] Panagiota Antonakaki, Dimitrios Kosmopoulos, and Stavros J Perantonis. Detecting abnormal human behaviour using multiple cameras. *Signal Processing*, 89(9):1723–1738, 2009.

[5] Simone Calderara, Uri Heinemann, Andrea Prati, Rita Cucchiara, and Naftali Tishby. Detecting anomalies in peoples trajectories using spectral graph analysis. *Computer Vision and Image Understanding*, 115(8):1099–1111, 2011.

[6] Brendan Tran Morris and Mohan M Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE transactions on pattern analysis and machine intelligence*, 33(11):2287–2301, 2011.

[7] Frederick Tung, John S Zelek, and David A Clausi. Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. *Image and Vision Computing*, 29(4):230–240, 2011.

[8] Erhan Baki Ermis, Venkatesh Saligrama, Pierre-Marc Jodoin, and Janusz Konrad. Motion segmentation and abnormal behavior detection via behavior clustering. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 769–772. IEEE, 2008.

[9] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, and Iain McCowan. Semi-supervised adapted hmms for unusual event detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 611–618. IEEE, 2005.

[10] Oren Boiman and Michal Irani. Detecting irregularities in images and in video. *International journal of computer vision*, 74(1):17–31, 2007.

[11] 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.

[12] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE, 2010.

[13] Venkatesh Saligrama and Zhu Chen. Video anomaly detection based on local statistical aggregates. In *Computer*

Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2112–2119. IEEE, 2012.

[14] Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2458–2465. IEEE, 2009.

[15] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1446–1453. IEEE, 2009.

[16] Anurag Mittal, Antoine Monnet, and Nikos Paragios. Scene modeling and change detection in dynamic scenes: A subspace approach. *Computer vision and image understanding*, 113(1):63–79, 2009.

[17] Andrei Zaharescu and Richard Wildes. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In *European Conference on Computer Vision*, pages 563–576. Springer, 2010.

[18] Tian Wang and Hichem Snoussi. Histograms of optical flow orientation for visual abnormal events detection. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 13–18. IEEE, 2012.

[19] Bhaskar Chakraborty, Michael B Holte, Thomas B Moeslund, and Jordi Gonzàlez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396–410, 2012.

[20] Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic. Spatiotemporal localization and categorization of human actions in unsegmented image sequences. *IEEE transactions on Image Processing*, 20(4):1126–1140, 2011.

[21] O. Beya, M. Hittawe, D. Sidibé, and F. Mériaudeau. Automatic detection and tracking of animal sperm cells in microscopy images. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2015 11th International Conference on*, pages 155–159. IEEE, 2015.

[22] Xudong Zhu and Zhijing Liu. Human behavior clustering for anomaly detection. *Frontiers of Computer Science in china*, 5(3):279–289, 2011.

[23] 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.

[24] Jingen Liu and Mubarak Shah. Learning human actions via information maximization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[25] Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, 47:358–368, 2016.

[26] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, et al. Fully convolutional neural network for fast anomaly detection in crowded scenes. *arXiv preprint arXiv:1609.00866*, 2016.

[27] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. 2014.

[28] Vladimir Vapnik. Pattern recognition using generalized portrait method. *Automation and remote control*, 24:774–780, 1963.

[29] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[30] 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.

[31] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2921–2928. IEEE, 2009.